EFFECT_CI: A SAS® Macro for Constructing Confidence Intervals Around Standardized Mean Differences

Melinda R. Hess, Jeffrey D. Kromrey, University of South Florida
Melinda R. Hess, University of South Florida, EDU162, Tampa, FL 33620

ABSTRACT

In recent years, researchers in a variety of fields have increasingly recognized the importance of (a) effect size statistics to communicate the magnitudes of treatment effects, and (b) confidence intervals to communicate the extent of uncertainty in parameter estimates due to sampling error. Although confidence intervals are frequently used for simple statistics such as means and correlation coefficients, the sampling distributions of effect sizes present greater challenges for constructing accurate interval estimates. Steiger and Fouladi (1992, 1997) described an interval inversion approach that provides a method for constructing confidence intervals for a variety of complex statistics, including sample effect sizes. By transforming the sample standardized mean difference effect size (*d*) into a noncentrality parameter, the noncentral *t* distribution is used to identify values of noncentrality for which the sample effect size is expected to occur (for example) 2.5% of the time and 97.5% of the time. These values of noncentrality are then transformed to provide the endpoints of a 95% confidence band around the sample value of *d*. This paper presents a SAS macro that calculates confidence intervals for standardized mean differences using the interval inversion approach. Inputs to the macro include the observed sample effect size and the sample sizes for the two groups. The macro computes and reports 80%, 90% and 95% confidence intervals around the sample effect size. The paper provides a demonstration of the SAS/IML code, sample output, and examples of applications in simulation studies.

INTRODUCTION

As with many aspects of research, there is increasing attention being levied on not only the impact of research results, but on proper and adequate conduct of that research as well as the need to provide comprehensive and sufficient reporting of appropriate statistics. One of these statistics in particular, effect size, has been receiving increasing recognition as a critical element in research applications that should be reported in the literature (Nix & Barnette, 1998). The new edition of the American Psychological Association (APA)'s style manual for publication (APA, 2001) cites the failure to report effect sizes, as well as other research issues, as defects in reporting research. However, Thompson (1998) has noted that the 'encouragement' of the APA has not seemed to induce sufficient leverage for researchers to consistently report this informative statistic. As a result, some professional journals have made it a requirement of authors to supply this information before consideration for publication. Unfortunately, recent research has indicated that while many journals technically require this information, few are enforcing their own requirements (McMillan, Snyder & Lewis, 2002).

The report by Wilkinson and the APA Task Force on Statistical Inference (1999) not only addresses the need for effect size reporting but also stresses the obligation of researchers to provide estimates for confidence intervals for all principal outcomes, including, but not limited to, effect size information. Analytical experts are increasingly investigating the use of confidence intervals for various parameter estimations in lieu of traditional point estimates (Nix & Barnette, 1998; Grissom & Kim, 2001). While the use of intervals around effect sizes has been the topic of various theoretical discussions, empirical investigation has just begun under limited conditions.

INTERVAL ESTIMATES FOR EFFECT SIZES

Research into robust and reliable effect size computation is ongoing and currently there are a variety of effect size indices available to researchers such as Cohen's *d*, Hedge's *g* and the trimmed *d* (Hogarty & Kromrey, 1999). For purposes of simplicity at this stage of the research, only Cohen's *d* is addressed in this paper. Cohen's *d* is defined as the difference in means between groups divided by the pooled standard deviation and is given by:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}}$$

where $\overline{X}_i, S_i^2$ and n_i are the sample mean, variance and size of group *i*.

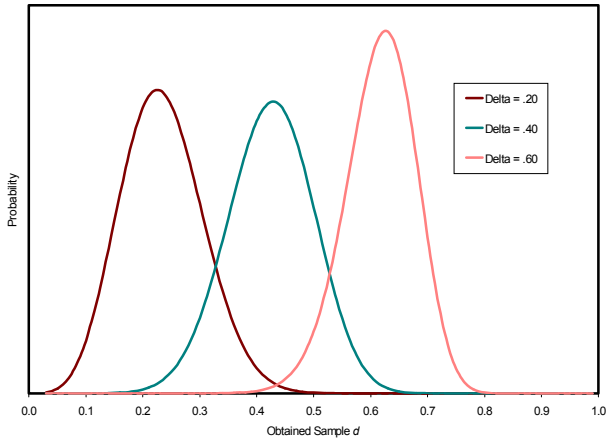
A variety of methods for constructing confidence bands around the sample value of Cohen's *d* have been suggested, including use of the normal distribution (relying on the asymptotic normality of the sampling distribution of *d*) and the application of the hyperbolic sine transformation to accelerate the approach to normality (Hedges & Olkin, 1985). Additionally, confidence intervals for effect sizes may be constructed using bootstrap approaches, techniques commonly recognized as efficient methods for providing estimates for, among other things, confidence intervals and standard errors (Efron & Gong, 1983; Efron & Tibshirani, 1986; Stine, 1990). However, comparisons of approaches to interval estimation for standardized mean differences suggest that one of the most promising methods is that of interval inversion.

THE INTERVAL INVERSION APPROACH

The interval inversion approach to confidence interval estimation was proposed by Steiger and Fouladi (1992, 1997). This method has shown promise in similar applications of confidence interval estimation for relatively complicated parameters (Kromrey & Hess, 2001; Hess & Kromrey, 2002). This method uses the sampling distribution of *d* to estimate the values of the population effect size δ for which the sample effect size, obtained from sample sizes of n_1 and n_2, would be expected (for example) 2.5% of the time and 97.5% of the time. Because analytical formulae for obtaining these values are not available, numerical methods are used (see, for example, Press, Teukolsky, Vetterling & Flannery, 1992).

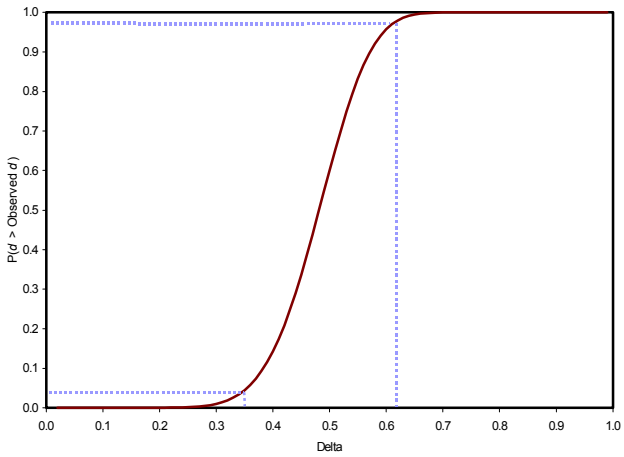
An illustration of this method of interval estimation is provided in Figures 1 and 2. Assume that a research analysis yields a sample effect size, *d*, of 0.48, from a study of two independent groups of observations with $n_1 = n_2 = 20$. The sampling distribution of *d* is graphed in Figure 1 for three potential values of the population effect size δ. If the population effect size is .60, then nearly all of the sampling distribution is greater than the observed effect size of .48; if the population value is .40, then approximately 40% of the sampling distribution is greater than the observed value of *d*; and if the population effect size is .20, then only a small portion of the sampling distribution is greater than the observed value.

Figure 1
Probability Densities of *d* Under Three Values of δ



Given the obtained value of *d* from a sample, one can compute the proportion of each sampling distribution that is greater than the observed value. Extending this thinking to an infinite number of potential values of δ (rather than the three illustrated in Figure 1), one can plot the proportion of the sampling distribution of *d* that is greater than the observed value of *d*, $\Pr(d > d_{obs})$, as a function of the population parameter. Such a graph is provided in Figure 2. Using the interval inversion approach, the endpoints of the 95% confidence interval are the values of δ for which $\Pr(d > d_{obs}) = .025$, and $\Pr(d > d_{obs}) = .975$ for the lower and upper limits, respectively.

Figure 2
Probability of *d* > Observed *d* as a Function of δ



In practice, programming the interval inversion method is simplified when the sampling distribution of a function *d* is used, rather than that of *d* itself. For example,

$$d\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

is distributed as Student's *t*, with $df = n_1 + n_2 - 2$, and noncentrality parameter, λ, where

$$\lambda = \delta\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Using this function of the standardized mean difference, the interval inversion method evaluates the noncentral *t* distribution and identifies values of noncentrality for which the observed sample noncentrality is expected to occur (for example) 2.5% of the time and 97.5% of the time. These values of noncentrality are then transformed to provide the endpoints of a 95% confidence band around the sample value of *d*. Such an approach may be easily implemented with SAS.

MACRO EFFECT_CI

A SAS/IML macro was designed to compute confidence intervals for Cohen's *d*, using the interval inversion method. The macro was developed to provide researchers with an easily accessible tool for constructing confidence intervals around sample effect sizes. Inputs to the macro include the name of the data set containing the observations, the obtained effect size and the sample sizes for the two groups being compared.

The effect sizes and associated samples sizes, n_1 and n_2, are read into column vectors for use within the macro. With these summary statistics organized into the three vectors, confidence intervals may be obtained for multiple sample effect sizes with a single call to the macro.

Within PROC IML, the subroutine FIND_DELTA uses the interval inversion method to compute a specific endpoint of a confidence band for a given Type I error rate. For example, the subroutine is called twice to compute the endpoints for a 95% confidence band, once for the .025 and again for the .975 percentile point. In this macro, the subroutine is called six times to compute the two endpoints for each of the Type I error rates under consideration: 80%, 90% and 95% for each observation input.

Using the sample sizes for a given observation, the degrees of freedom (*df*) are computed. The values of the obtained effect size and sample sizes are then used to compute the noncentrality parameter (nc). Once the degrees of freedom and noncentrality parameter are computed, the endpoint of the confidence interval is found using a 3-step process. First, a value for the endpoint is found that is slightly too large through the use of a do loop that increments the value of the hypothetical population effect size by .001 each time until the probability from student's *t* distribution exceeds the target percentile. A similar process is used to obtain an endpoint value that is slightly too small. Once these two boundary values are found, the macro uses the method of bisection to successively halve the distance between the two values until that difference is sufficiently small (1×10^{-11}). Once this convergence is reached, the subroutine returns the target value of δ that is the endpoint of the confidence interval.

The provided version of the macro calls this subroutine six times to provide the six endpoints of interest. Once all six iterations have been accomplished, the obtained sample effect size and the endpoints of the three confidence intervals at the various alpha levels are printed using the FILE PRINT statement.

```
* +----------------------------------------+
   Input to the macro:
      data = name of data set
      effect_size = obtained sample
      value of Cohen d
      n1 = sample size of group one
      n2 = sample size of group two
   Output is printed table of confidence
      intervals
* +----------------------------------------+;

%macro EFFECT_CI(data, effect_size, n1, n2);

proc iml;

start find_delta(obs_stat, n1, n2, pctl,
       delta_t);

  df = n1 + n2 - 2;

  * Step 1: Find value of delta that is a
            little too high;
  OK = 0;
  delta_t = 0;
  * start loop with pop effect size = 0;
  do until (OK = 1);
     nc = delta_t # sqrt(n1#n2/(n1+n2));
     cumprob = PROBT(obs_stat,df,nc);
     if cumprob<pctl then OK = 1;
     if cumprob>pctl then
      delta_t = delta_t + .001;
  end;
  high = delta_t;

  * Step 2: Find value of delta that is a
            little too low;
  OK = 0;
  delta_t = 0;
  * start loop with pop effect size = 0;
  do until (OK = 1);
     nc = delta_t # sqrt(n1#n2/(n1+n2));
     cumprob = PROBT(obs_stat,df,nc);
     if cumprob>pctl then OK = 1;
     if cumprob<pctl then
      delta_t = delta_t - .001;
  end;
  low = delta_t;

  * Step 3: Successively halve the interval
            between low and high to obtain
            final value of percentile;
  change = 1;
  small = .00000000001;
  do until (change<small);
    half = (high + low)/2;
    nc = half # sqrt(n1#n2/(n1+n2));
    cum_h = PROBT(obs_stat,df,nc);
    if cum_h < pctl then high =  half;
          * still too high;
    if cum_h > pctl then low  = half;
          * still too low;
    change = abs(high - low);
    Delta_t = (high + low)/2;
  end;
finish;

use &data;
 read all var{&effect_size} into effect_vec;
 read all var{&n1} into n1;
 read all var{&n2} into n2;
 k = nrow(effect_vec);

 file print;
 put @1 'Confidence Intervals Around Sample
```

```
        Effect Sizes' //
      @16 '95% CI' @36 '90% CI' @56 '80% CI' /
      @2 'Effect' @10 '------------------'
      @30 '------------------'
      @50 '------------------' /
      @3 'Size' @12 'Lower     Upper'
      @32 'Lower     Upper'
      @52 'Lower     Upper' /
      @1 '--------' @10 '--------- ---------'
      @30 '--------- ---------'
      @50 '--------- ---------';

do i = 1 to k;

  obs_stat = effect_vec[i,1] #
         sqrt(n1[i,1]#n2[i,1]/(n1[i,1] +
         n2[i,1]));

  run find_delta(obs_stat, n1[i,1], n2[i,1],
                .025, delta025);
  run find_delta(obs_stat, n1[i,1], n2[i,1],
                .975, delta975);
  run find_delta(obs_stat, n1[i,1], n2[i,1],
                .05, delta05);
  run find_delta(obs_stat, n1[i,1], n2[i,1],
                .95, delta95);
  run find_delta(obs_stat, n1[i,1], n2[i,1],
                .10, delta10);
  run find_delta(obs_stat, n1[i,1], n2[i,1],
                .90, delta90);

  print_effect = effect_vec[i,1];
  file print;
  put   @1 print_effect 8.3 @10 delta975 8.3
       @20 delta025 8.3 @30 delta95 8.3
       @40 delta05 8.3 @50 delta90 8.3
       @60 delta10 8.3;
end;
quit;

%mend EFFECT_CI;
```

The output for the macro is formatted in such a way as to provide a single line for each observation/study input into the macro, with lower and upper endpoints given at three Type I error rates.

With relatively minor changes to the macro, essentially the percentiles specified as arguments to the subroutine FIND_DELTA, one can compute intervals for any desired level of Type I error rate. Additionally, the macro could be further modified to compute or print other statistics of interest, including sample sizes and widths of confidence intervals.

INVOKING THE MACRO

The easiest way in which the macro EFFECT_CI may be used is to simply create a SAS dataset that inputs the sample effect size(s) and sample sizes. The macro is then called, using as arguments the name of the dataset, the name of the variable that contains the effect sizes and the names of the two variables that contain the sample sizes for each effect size. For example, the following code reads three sample effect sizes and their corresponding sample sizes. The data are read into a SAS dataset called ONE and are referenced by the variable names sample_d, treatment_n and control_n. The call to the macro EFFECT_CI requests the estimation of confidence intervals for each of the effect sizes.

```
data one;
   input sample_d treatment_n control_n;
cards;

 0.246 25 30
 0.572 80 80
-0.885 25 15
;
%EFFECT_CI (one,sample_d,treatment_n,control_n)
run;
```

OUTPUT FROM MACRO EFFECT_CI

Table 1 provides an example of the output provided by Macro EFFECT_CI. The output includes each sample effect size provided as input to the macro as well as confidence bounds for three levels of Type error: alpha = 0.05, 0.10, and 0.20.

Table 1. Example of output from Macro EFFECT_CI

Confidence Intervals Around Sample Effect Sizes

	95% CI		90% CI		80% CI	
Effect size	Lower	Upper	Lower	Upper	Lower	Upper
0.246	-0.288	0.778	-0.202	0.692	0.103	0.593
0.572	0.255	0.887	0.306	0.836	0.364	0.778
-0.885	-1.550	-0.210	-1.442	-0.317	-1.317	-0.441

In this example, the lower and upper endpoints of the 95%, 90% and 80% confidence intervals are provided for the three different observations. The 95% confidence band for the study with the rather large negative effect size of -0.8850 extends from -1.5498 to -0.2096.

EMPIRICAL STUDIES OF INTERVALS

Simulation studies on confidence interval construction (Hess & Kromrey, 2003; Hess & Kromrey, 2002) under conditions where both populations had equal variances suggested that the interval inversion approach provided accurate confidence intervals around the sample effect size across a broad range of sample sizes, population distribution shapes, and values of δ. However, for conditions with heterogeneous variances (especially when paired with unequal sample sizes), no approach to interval estimation provided accurate confidence bands. For

such conditions, *d* is a biased estimate of δ (see, for example, Kraemer & Andrews, 1982; Wilcox & Muska, 1999). Such bias in a statistical point estimate means that the confidence interval may be the appropriate width to obtain the nominal level of confidence, but the interval is being constructed in the wrong location. A recent simulation study (Hogarty & Kromrey, 2001) confirmed the substantial bias in Cohen's *d* as a point estimator when populations are heterogeneous in variance.

For such conditions, alternative non-parametric indices appear promising. For example, Hogarty and Kromrey (2001) found that ordinal indices of effect size (Cliff's *d*, Cliff, 1993, 1996; or the closely related $\hat{A}$ proposed by Vargha & Delaney, 2000) provided relatively unbiased point estimates of the corresponding parameters under a variety of distributional conditions.

CONCLUSIONS

The macro EFFECT_CI provides accurate confidence intervals for standardized mean differences obtained under a variety of data conditions. However, under extremely large sample sizes (total *N* of several thousand) or large sample effect sizes (*d* of 5.00 or larger), the SAS function PROBT is unable to evaluate the cumulative distribution function. The limits of the PROBT function depend upon the combination of sample effect size, degrees of freedom for the *t* distribution and noncentrality parameter. If such conditions are reached during the computation of the confidence intervals, the macro will stop executing and return an error message that an invalid argument has been sent to the PROBT function. For data structures with very large samples or very large effect sizes, alternative methods of confidence interval construction are recommended. For example, Hedges and Olkin (1985) suggested confidence intervals constructed using the normal distribution and the standard error of the sample effect size

$$\hat{\sigma}_d = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}}$$

With larger sample sizes, such a normal approximation to the sampling distribution of *d* is quite accurate.

Although the macro has been provided to yield 80%, 90%, and 95% confidence intervals, the code may be modified easily to accommodate other levels of confidence. Further, one-sided confidence intervals may be obtained with this macro. Such intervals are used to provide confidence statements such as 'we are 95% sure that the population effect size is at least 0.678.' Such modifications to the code simply require changing the fourth argument to the FIND_DELTA subroutine. For example, using arguments of .005 and .995 will provide the endpoints of a 99% confidence interval.

In addition, other effect sizes (besides the standardized mean difference) may be used, but this alteration will require more modification to the program code. Most effect sizes can be converted to noncentrality parameters, and such parameters can be used within the structure provided by the macro EFFECT_CI. For example, effect sizes that are often used in the context of multiple correlation analysis or multiple regression analysis are given by

$$f^2 = \frac{R^2}{1 - R^2} \quad \text{and} \quad f^2 = \frac{\Delta R^2}{1 - R_L^2}$$

These effect sizes are converted into noncentrality parameters for the F distribution, using

$$\lambda = f^2 \left(df_{num} + df_{den} + 1 \right)$$

Using the program structure provided in EFFECT_CI, this noncentrality parameter may be used with the PROBF function to obtain confidence bands for this effect size.

In summary, the use of effect sizes has grown in popularity in recent years (although such application remains far from universal). Because effect sizes, in many instances, provide useful information to supplement more traditional inferential statistics, advocacy for their use is appropriate. Similarly, the use of interval estimates to complement point estimates and hypothesis tests is a worthy endeavor. This macro is provided to facilitate researchers' calculation and use of confidence intervals for standardized mean differences.

REFERENCES

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144-151.

Carpenter, J. & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine, 19*, 1141-1164.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114, 494-509.

Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research, 31*, 331-350.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cooper H. & Hedges, L. (1994). *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Efron, B. & Gong, G. (1983). *A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 37(1),* pg 36-49.

Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1(1),* p. 54-77.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43(4),* p.521-532.

Grissom R.J. & Kim J.J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, 6(2), p. 135-146.

Hedges L.V. & Olkin I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Hess, M.R. & Kromrey, J.D. (2002, April). *Confidence intervals for the standardized mean difference: An empirical comparison of methods for interval estimation of effect sizes.* Paper presented at the American Educational Research Association, New Orleans, LA.

Hess, M.R. & Kromrey, J.D. (2003, April). *Confidence Intervals for Standardized Mean Differences: An Empirical Comparison of Bootstrap Methods Under Non-normality and Heterogeneous Variances.* Paper presented at the American Educational Research Association, Chicago, IL..

Hogarty K. Y. & Kromrey, J.D. (1999, August). *Traditional and robust effect size estimates: Power and Type I error control in meta-analytic tests of homogeneity*. Paper presented at the Joint Statistical Meetings, Baltimore.

Hogarty, K. Y. & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Kromrey, K. Y. & Hess, M. H. (2001, April). *Interval Estimates of R^2: An empirical comparison of accuracy and precision under violations of the normality assumption*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Kromrey, J. D. & Hogarty, K. Y. (1999, April). *Traditional and robust effect size estimates: an empirical comparison in meta-analytic tests of homogeneity*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

McMillan, J.H., Snyder, A. & Lewis, K.L., (2002, April). *Reporting Effect Size: The Road Less Traveled.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Nix, T.W. & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, *5*(2), p. 3-14.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd Ed.). New York: Campbridge.

Robey, R.R. & Barcikowski, R.S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45,* 283-288.

Steiger, J. H. & Fouladi, R. T. (1992). R2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research, Methods, Instruments, and Computers, 4,* 581-582.

Steiger, J. H. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum, p. 221-257.

Stine, R. (1990). An introduction to bootstrap methods. *Sociological Methods and Research, 18 (2&3),* p. 243-291.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools*, *5*(2), p. 33-38.

Vargha, A. & Delaney, H.D. (2000). A critique and improvement of the CL Common Language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*, 101-132.

Wilkinson & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

ACKNOWLEDGEMENTS

This work was supported, in part, by the University of South Florida and the National Science Foundation, under Grant No. REC-9988080. The opinions expressed are those of the authors and do not reflect the views of the National Science Foundation or the University of South Florida.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact Melinda Hess at:

University of South Florida
4202 East Fowler Ave. EDU 162
Tampa, FL 33620
Work Phone: 813-974-5739
Email: mhess@helios.acomp.usf.edu