Paper ST- 05

FDR_TEST: A SAS Macro for Calculating New Methods of Error Control in Multiple Hypothesis Testing

Jeffrey D. Kromrey, Kristine Y. Hogarty, University of South Florida

Jeffrey D. Kromrey, University of South Florida, EDU 162, Tampa, FL 33620

ABSTRACT

The testing of multiple null hypotheses in a single study is a common occurrence in applied research. The problem of Type I error inflation or probability pyramiding in such contexts has been well-known for many years. General procedures for the control of Type I error rates in multiple testing are the Bonferroni procedure and its' more recent modifications. These procedures partition a desired level of Familywise error across the set of hypotheses being tested. Recent work on multiple testing by Benjamini and Hochberg (1993, 1995, 2000) has focused on controlling the False Discovery Rate (FDR) rather than rates of Type I error. The adaptive (BH-A) and non-adaptive (BH) procedures for controlling the FDR in a set of tests promise increases in statistical power relative to other procedures. This paper presents a SAS macro that calculates probabilities under five decision rules that may be used in multiple testing (per hypothesis rule, Bonferroni, Hochberg, BH, BH-A) The macro evaluates a set of probabilities that are supplied as an input. Macro outputs include the results of the five decision rules applied to the set of probabilities. The paper provides a demonstration of the SAS/IML code and examples of the application of the code in simulation studies.

INTRODUCTION

It is not unusual for applied researchers to test a host of null hypotheses during the conduct of a single research study. Relevant contexts include testing all elements of a correlation matrix, testing individual regression weights in a multiple regression analysis, conducting subgroup analyses in survey research, structural equation modeling, and conducting follow-up tests in analysis of variance. In educational research, it is especially common for researchers to investigate an array of factors, employing a myriad of tests related to a common underlying phenomenon. For example, recent research in the area of educational reform related to the National Science Foundation's Systemic Initiatives (Kromrey et al., 2002) has investigated the influence of participation in reform efforts and changes in student outcomes. The simultaneous examination of student achievement across grade levels, subject area and years of participation calls for a method of adjustment designed to control erroneous rejections without the often associated deleterious loss in statistical power.

The problem of Type I error inflation or probability pyramiding in such contexts has been well-known for many years (Ryan, 1959; Toothacker, 1993), although techniques for the control of Type I error are most commonly applied in ANOVA contexts. General procedures for the control of Type I error rates in multiple testing are the Bonferroni procedure (Dunn, 1961) and its more recent modifications (Holm, 1979; Hochberg, 1988). These procedures partition a desired level of Familywise error (α_{FW}) across the set of hypotheses being tested, such that the probability of a Type I error in the set of tests is no larger than α_{FW}.

Recent work on multiple testing by Benjamini and Hochberg (1993, 1995, 2000) has focused on controlling the False Discovery Rate (FDR) rather than rates of Type I error. The FDR is the proportion of rejected null hypotheses that represent Type I errors. In the Benjamini and Hochberg procedures, control is not sought for the number of hypotheses tested (*m*), but for the number of tested null hypotheses that we think are really true (m_0). The Benjamini and Hochberg procedures (an adaptive procedure, referred to henceforth as *BH-A*, and a non-adaptive procedure referred to as *BH*) for controlling the False Discovery Rate in a set of tests promise increases in statistical power by forgoing control of Familywise Type I error rates (α_{FW}) in lieu of controlling the False Discovery Rate. Keselman, Cribbie and Holland (1999) verified the power advantages of the BH procedure in pairwise follow-up tests in ANOVA, although the power advantages were limited to conditions in which the number of groups was at least 5.

An Example of the Differences

Presented in Table 1 are results of testing *m* null hypothoses. Each hypothesis test (H_i) yields a p-value (P_i) and the hypotheses have been ordered in terms of their p-values such that $P_1 \leq P_2 \leq \ldots P_m$. A researcher employing the unprotected t-test procedure would evaluate each P_i at the hypothesiswise alpha level (e.g., .05) to make a reject/fail-to-reject decision. Similarly, a researcher employing the Bonferroni procedure would set Familywise alpha at a reasonable level, then conduct each test at a reduced level of hypothesiswise alpha $\alpha_{HW} = m^{-1}\alpha_{FW}$. For the use of a modified Bonferroni procedure such as Hochberg's procedure, the Familywise alpha is adjusted sequentially such that the smallest observed probability in the set is compared to $m^{-1}\alpha$ and the next smallest is compared to $(m-1)^{-1}\alpha$, and so forth. For the Hochberg procedure, testing begins at the bottom of the set of ordered p-values and proceeds up the list until a null

hypothesis is rejected. At that point, all hypotheses further up the list (i.e., smaller values of *i*) are also rejected.

Table 1
Decision Criteria in Multiple Hypothesis Testing.

Hypothesis (i)	P	Bonf	Mod Bonf	BH and BH-A
H_1	P_1	α/m	α/m	q/m_0
H_2	P_2	α/m	$\alpha/(m-1)$	$2q/m_0$
H_3	P_3	α/m	$\alpha/(m-2)$	$3q/m_0$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
H_m	P_m	α/m	$\alpha/1$	mq/m_0

The BH and BH-A procedures are similar to the Modified Bonferroni procedure except that the desired False Discovery Rate (*q*) rather than Familywise Type I error rate is employed to establish the rejection criteria. Sequential rejection criteria are established for the ordered probabilities resulting from the hypothesis tests, such that the smallest probability will be compared to $m_0^{-1}q$, the second smallest to $m_0^{-1}2q$, and so forth. As with the Hochberg procedure, the testing begins at the bottom of the ordered list of tests and once a null hypothesis is rejected, all hypotheses higher in the list are also rejected (smaller values of *i*).

As a concrete example of the differences in the rejection criteria realized by these approaches, consider a researcher conducting a set of nine hypothesis tests. The probabilities of the test statistics for the nine tests range from .0046 to .9600, and are presented below (Table 2) arranged from smallest to largest probability.

Table 2
Example of Decision Criteria for Testing Nine Hypotheses.

Hypothesis (i)	p	Bonf	Modified Bonf	BH and BH-A
1	0.0046	.0055	.0055	.0055
2	0.0074	.0055	.0062	.0111
3	0.0133	.0055	.0071	.0166
4	0.4241	.0055	.0083	.0222
5	0.4989	.0055	.0100	.0277
6	0.5870	.0055	.0125	.0333
7	0.7240	.0055	.0166	.0388
8	0.8094	.0055	.0250	.0444
9	0.9600	.0055	.0500	.0500

In the Bonferroni approach, each probability is compared to the value .0055 (that is, .05/9) to reach a reject/fail to reject decision. The sequential criteria for the modified Bonferroni and the BH/BH-A procedures illustrate that the criteria are identical for the hypothesis with the smallest probability (.0055, the same value as the Bonferroni criterion) and for the hypothesis with the largest probability (.05). For all steps in between these extremes, however, the BH and BH-A procedures provide larger criteria than those of the modified Bonferroni.

The BH procedure and the BH-A procedure differ in the definition of m_0. The BH procedure sets $m_0 = m$, so that the number of anticipated true nulls is set equal to the number of hypotheses tested. With the BH-A, the value of m_0 is estimated from the sequential pattern of observed probabilities associated with the *m* hypothesis tests conducted. This estimation proceeds by considering a graph of the ordered sequence of p-values, as illustrated in Figure 1 for a hypothetical set of nine tests. A series of *m* lines is fit to these data, such that each line passes through the two points ($m + 1$, 1) and (i, P_i), and the slope of each line is calculated, beginning with $i = 1$. The sequence of slopes is evaluated and the process is discontinued when the slope for line *i* is less than the slope for line $i - 1$. The estimated value of m_0 is then obtained as the smaller of ($1/S_j + 1$) and *m* (where S_j is the slope of the line at the point when the process of slope calculation was discontinued). For these data, the slope decreases at $i = 4$, resulting in ($1/S_j + 1$) = 11 and $m_0 = m$ = 9.

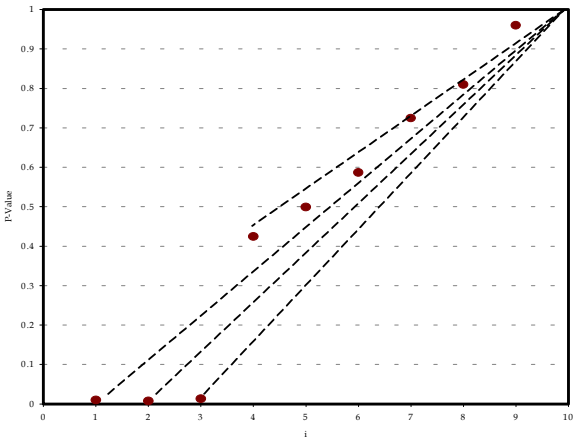

Figure 1. Plot of Ordered P Values.

AN EMPIRICAL STUDY

In a recent study conducted by Hogarty and Kromrey (2002), the relative performance of the procedures was evaluated in terms of Familywise Type I error rate, False Discovery rate and statistical power. The Familywise Type I error rate estimate is simply the proportion of samples in which at least one Type I error occurred:

$$\alpha_{FW} = \frac{k_I}{k}$$

where k_I = the number of samples in which at least one Type I error was committed and *k* = the number of samples generated.

In contrast, the False Discovery Rate is the ratio of the number of Type I errors to the number of null hypotheses rejected. This rate was calculated for each set of tests and averaged over the simulations conducted:

$$\text{FDR} = \frac{1}{k}\sum_{k}\left(\frac{\textit{Number of true } H_0 \textit{ rejected}}{\textit{Number of } H_0 \textit{ rejected}}\right)$$

Statistical power in the context of multiple hypothesis tests was also evaluated, in terms of any-hypothesis power, per-hypothesis power and all-hypotheses power. Any-hypothesis power is the probability of rejecting *at least one* false null hypothesis in a set of tests. In contrast, all-hypotheses power is the probability of rejecting *all* false null hypotheses in the set of tests and per-hypothesis power is the probability of rejecting each false null hypothesis in a set of tests (Toothaker, 1991). Generally, all-hypothesis power is less than or equal to per-hypothesis power, which is less than or equal to any-hypothesis power. The overall Type I error rates and False Discovery rates are presented below. The results in their entirety, delineated by the central design factors of the study (sample size, number of hypotheses tested, proportion of true null hypotheses, effect size of non-null hypotheses), are available from the authors.

Familywise Type I Error Rates

The distributions of estimated Familywise Type I error rates across the conditions examined are presented in Figure 2. This figure illustrates an inflation of Type I error rate for the unprotected t-test in the majority of conditions examined. In contrast, the Bonferroni and Hochberg procedures maintained Type I error control at near nominal levels across all conditions examined. The BH procedure evidenced inflated Familywise error rates (because the procedure was not designed to control this rate), but to a much lower extent than that of the unprotected t-test. Finally, the BH-A procedure showed notably more liberal Type I error control than the non-adaptive version.

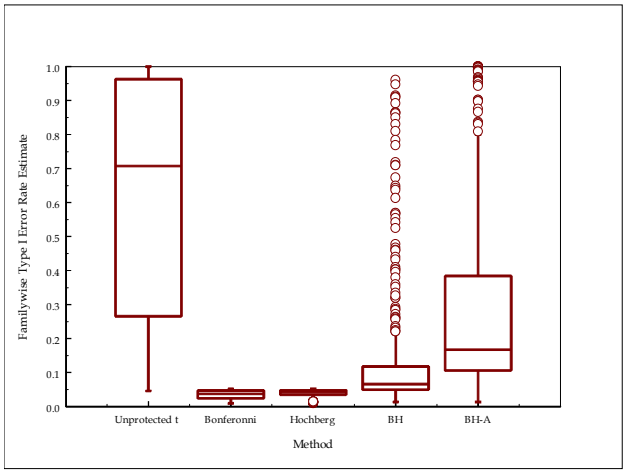

Figure 2. Distribution of Familywise Type I Error Rate Estimates.

False Discovery Rates

The distributions of estimated False Discovery rates across the conditions examined are presented in Figure 3. Consistent with the Familywise error results, this figure illustrates the inflation of the False Discovery rates for the unprotected t-test in the majority of conditions examined. Both the Bonferroni and Hochberg procedures evidenced conservative control of the False Discovery rate while the original BH procedure maintained the rate very near the nominal level across the conditions examined. Finally, the adaptive version of the BH procedure showed notably more liberal False Discovery rate control than the non-adaptive version. These error rates were further analyzed for each of the research design factors included in the study. Details of these additional analyses (as well as details on statistical power comparisons) are available in Hogarty and Kromrey (2002).

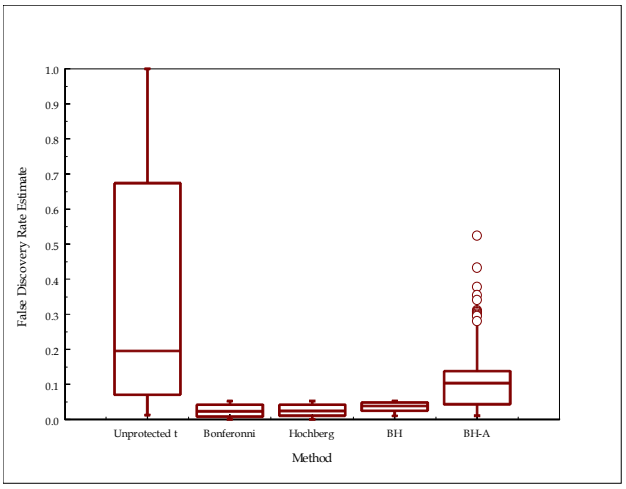

Figure 3. Distribution of False Discovery Rate Estimates.

MACRO FDR_TEST

The macro FDR_TEST uses PROC IML to evaluate a set of probabilities using (a) an unprotected (Hypothesiswise error rate) approach, (b) the traditional Bonferroni and the Hochberg modified Bonferroni approaches for Familywise error rate control, and (c) the BH and BH-A approaches for FDR control. The macro FDR_TEST was developed to provide researchers with an easily accessible tool for using these approaches to error control in the context of multiple hypothesis testing. Any set of probabilities representing a family of tests can be read into SAS and evaluated using the macro. The output from the macro is an easily interpretable table providing reject and fail-to-reject decisions regarding each of the probabilities in the set.

Inputs to the macro include the name of the SAS dataset that contains the set of probabilities to be evaluated, the variable name for the probabilities in the SAS dataset, and the nominal level of alpha and FDR for testing the set of probabilities. The nominal level specified as the third argument in the call to the macro is used as the Hypothesiswise alpha level in the unprotected tests, as the Familywise alpha level for the Bonferroni and Hochberg procedures, and as the FDR level for the BH and BH-A approaches.

```
%macro FDR_TEST (dataset,p_values,q);
proc iml;
 use &dataset;
 read all var{&p_values} into pvalues;
 q_value = &q;
* +-----------------------------------+
```

```
   Obtain ranks and anti-ranks of the
   p-values. Create vector of DV numbers
   to keep track of the tests
* +-------------------------------------+;
  m = nrow(pvalues);
  rank_p = rank(pvalues);
  antirank_p = m + 1 - rank_p;
  ordered_ps = J(m,1,0);
  DV_number = j(m,1,0);
  do i = 1 to m;
    position = antirank_p[i,1];
    ordered_ps[position,1] = pvalues[i,1];
    dv_number[position,1] = i;
  end;

* +-------------------------------------+
   Create vectors to hold decisions:
   1 = reject, 0 = fail to reject
* +-------------------------------------+;

  unprot = J(m,1,0);
  bonf = J(m,1,0);
  hoch = J(m,1,0);
  FDR = J(m,1,0);

  do i = 1 to m;
    ranki = m + 1 - i;
    if ordered_ps[i,1] < q_value then
    unprot[i,1] = 1;
    if ordered_ps[i,1] < q_value/m then
    bonf[i,1] = 1;
    if ordered_ps[i,1] < q_value/i then
    hoch[i,1] = 1;
    if ordered_ps[i,1] < (q_value#ranki)/m
    then FDR[i,1] = 1;
  end;

* +-------------------------------------+
   Second loop through Hochberg and FDR
   for the 'step-up' criterion.
* +-------------------------------------+;

  gotone_H = 0;
  gotone_F = 0;
  do i = 1 to m;
      if hoch[i,1]=1 then gotone_H = 1;
      if gotone_H=1 then hoch[i,1] = 1;
      if FDR[i,1]=1 then gotone_F = 1;
      if gotone_F=1 then FDR[i,1] = 1;
  end;

* +-------------------------------------+
    Adaptive FDR
* +-------------------------------------+;
  FDR_A = FDR;
  m0_hat = m;
  S = J(m,1,0);
* +-------------------------------------+
    When estimated slope decreases, save
    the value of 'i' as 'Less_than'
* +-------------------------------------+;
  Less_than = 0;
  do i = 1 to m;
    ranki = m + 1 - i;
    S[i,1] = ((1-ordered_ps[i,1])/(m + 1 -
    ranki));
    if (i > 1 & Less_than = 0) then do;
      if S[i,1] < S[i - 1,1] then Less_than
      = i;
    end;
  end;
* +-------------------------------------+
    If a decrease was found, estimate m0
* +-------------------------------------+;
  if Less_than > 0 then do;
    m0_hat = round((1/S[Less_than,1]) + 1);
    if m0_hat >= m then m0_hat = m;
  end;
* +-------------------------------------+
    If estimated m0 is less than m,
    apply the adaptive FDR procedure
* +-------------------------------------+;

  if m0_hat < m then do;
    do i = 1 to m;
      if ordered_ps[i,1] < q_value#i)/m0_hat
      then FDR_A[i,1] = 1;
    end;
  end;
* +-------------------------------------+
     Second loop through FDR_A for the
     'step-up' criterion.
* +-------------------------------------+;
  gotone_F = 0;

  do i = 1 to m;
    if FDR_A[i,1] = 1 then gotone_F = 1;
    if gotone_F = 1 then FDR_A[i,1] = 1;
  end;

* +-------------------------------------+
    Sort decisions into correct sequence
    in the vectors using DV_number
* +-------------------------------------+;

  decisions=unprot||bonf||hoch||FDR||FDR_A;
```

```
  do i = 1 to m;
    position = DV_number[i,1];
    unprot[position,1] = decisions[i,1];
    bonf[position,1] = decisions[i,2];
    hoch[position,1] = decisions[i,3];
    FDR[position,1] = decisions[i,4];
    FDR_A[position,1] = decisions[i,5];
  end;

file print;
  put @1 'Multiple Testing Results'/
  @1 '------------------------'/
  @1 'Number of Tests: ' @48 m 3. /
  @1 'False Discovery Rate /
      Familywise Error Rate:'
  @47 q_value 4.2 //
  @1 'FWE Control   FDR Control' /
  @1 'Test  Unprot --- -------------' /
  @1 'Prob   test   Bonf   Hoch    BH
  BH-A' /
  @1 '------ --- ------ ------ ------' ;

  do i = 1 to m;
    pvalue_p = pvalues[i,1];
    if unprot[i,1]=1 then unprot_p = 'Rej';
    if unprot[i,1]=0 then unprot_p = 'FTR';
    if bonf[i,1] = 1 then print_b = 'Rej';
    if bonf[i,1] = 0 then print_b = 'FTR';
    if hoch[i,1] = 1 then print_h = 'Rej';
    if hoch[i,1] = 0 then print_h = 'FTR';
    if fdr[i,1] = 1 then print_f1 = 'Rej';
    if fdr[i,1] = 0 then print_f1 = 'FTR';
    if fdr_a[i,1]=1 then print_f2 = 'Rej';
    if fdr_a[i,1]=0 then print_f2 = 'FTR';
    file print;
      put @1 pvalue_p 6.4 @10 unprot_p 3.
        @17 print_b 3. @24 print_h 3.
        @31 print_f1 3. @38 print_f2 3.;
  end;

  file print;
  put
    @1 '------------------------------' ;
quit;
%mend FDR_TEST;
```

INVOKING THE MACRO

The easiest way in which the macro FDR_TEST may be used is to simply create a SAS dataset that inputs the probabilities for the set of tests being evaluated. The macro is then called, using as arguments the name of the dataset, the name of the variable that contains the probabilities and the desired nominal level at which tests will be conducted. For example, the following code reads eight probabilities (that may have been obtained from eight independent t-tests or analyses of variance, or that may represent the tests of eight correlation coefficients or regression parameter estimates). These probabilities are read into a dataset called ONE and are referenced by the variable name pvalue. Arranging the probabilities in ascending order facilitates interpretation of the results.The call to the macro FDR_TEST requests an evaluation of these eight probabilities using the .05 level of statistical significance.

```
data one;
 input pvalue;
cards;
.0001
.0002
.0011
.0022
.0123
.0211
.0304
.0664
;
%FDR_TEST (one,pvalue,.05)
run;
```

OUTPUT FROM MACRO FDR_TEST

Table 3 provides an example of the output produced by the macro FDR_TEST. The individual test probabilities are listed in the order in which they were delivered to the macro and the dichotomous decisions (reject vs. fail to reject) that were obtained under each decision rule are provided in columns of the table.

Table 3

```
Multiple Testing Results
------------------------
Number of Tests:                          8
FDR / FWER :                           0.05
```

		FWE Control		FDR Control	
Test Prob	Unprot test	Bonf	Hoch	BH	BH-A
0.0001	Rej	Rej	Rej	Rej	Rej
0.0002	Rej	Rej	Rej	Rej	Rej
0.0011	Rej	Rej	Rej	Rej	Rej
0.0022	Rej	Rej	Rej	Rej	Rej
0.0123	Rej	FTR	Rej	Rej	Rej
0.0211	Rej	FTR	FTR	Rej	Rej
0.0304	Rej	FTR	FTR	Rej	Rej
0.0664	FTR	FTR	FTR	FTR	FTR

In this example, both the BH and BH-A procedures provided rejection of all null hypotheses tested except for the last one (p = .0664), which are the same decisions that

were reached using the unprotected testing procedure. In contrast, the Hochberg modified Bonferroni procedure rejected five of the eight tests and the original Bonferroni procedure rejected only four of the eight tests.

Tthe macro can easily be modified to include labels for the individual tests that have been evaluated. For example, an alphanumeric variable could be included whose values provide a label for each test. This variable could be included as an additional argument to the macro and the values could be included in the printed output. In addition, the probabilities that are being evaluated can be obtained directly from output of other SAS procedures, utilizing the flexible ODS component of recent SAS versions.

Researchers conducting multiple hypothesis tests must remain cognizant of the statistical and conceptual issues surrounding the control of Type I error rates. Specifically, the reader is referred to the seminal work on Type I error control by Ryan (1959), where three major error rates are defined: the per comparison rate, the per experiment rate, and the experiment-wise rate. Additionally, there exist a variety of possible definitions of "families of hypotheses" resulting in fundamental differences among researchers regarding the error rates to be kept under control in a given research situation.

Finally, whereas agreement regarding what defines a family may appear somewhat elusive, few would argue that many issues exist that either separately or in combination need to be considered when making multiplicity adjustments. Accordingly, the related nature of the questions being considered, the number of comparisons examined, the degree of controversy surround the research, the notion of who stands to benefit, and the nature of the study are all important considerations when deciding on the proper adjustment and procedure (Proschan & Waclawiw, 2000).

REFERENCES

Benjamini, Y. & Hochberg, Y. (1993). The adaptive control of the false discovery rate in multiple independent testing problems. *Series in Statistics 93.1, Technical Report of the Department of Statistics and OR*, Tel Aviv University, Tel Aviv, Israel.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate – a new and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*, 289 – 300.

Benjamini, Y. & Hochberg, Y. (2000). On the adaptive control of the False Discovery Rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, *25*, 60 – 83.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 52-64.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*, 800 – 803.

Hogarty, K.Y. & Kromrey, J.D. (2002). F*amily Wise Errors, False Discovery Rates and Statistical Power: An Empirical Comparison of Multiple Hypothesis Testing Procedures*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65—70.

Keselman, H. J., Cribbie, R. & Holland. B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparisonwise Type I error control. *Psychological Methods*, *4*, 58-69.

Kromrey, J. D., Ferron, J. M, Parshall, C. G., Hogarty, K. Y., Grinnell, L., Hess, M. R., Lee, R., Romano, J., Sentovich, C., Watson, F., Dawkins, G. & Niles, J. (2002, April). *Evidence of attainment: A comparison of methods for representing and communicating student outcomes in systemic reform*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Proschan, M. A. & Waclawiw, M.A. (2000). Practical guidelines for multiplicity adjustment in clinical trials. *Controlled Clinical Trials, 21*, 527-539.

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

Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, *56*, 26-47.

Toothacker, Larry E. (1993). *Multiple comparison procedures*. Newbury Park: Sage

ACKNOWLEDGMENTS

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 Jeff Kromrey at:

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