

Pairwise Multiple Comparison Test Procedures

H. J. Keselman
University of Manitoba

Robert A. Cribbie
University of Manitoba

and

Burt Holland
Temple University

INTRODUCTION

Researchers in the social sciences are often interested in comparing the means of several treatment conditions ($\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_J$; $j = 1, \dots, J$) on a specific dependent measure. When each treatment group mean is compared with every other group mean, the tests are designated pairwise comparisons. When computing all pairwise comparisons the researcher must consider various issues (the issues pertain to other classes of multiple tests as well): (a) the multiplicity effect of examining many tests of significance, (b) the selection of an appropriate level of significance (α), and (c) the selection of an 'appropriate' multiple comparison procedure (MCP). The goals of the research should guide these decisions. Researchers faced with these decisions have often settled on 'traditional' choices [e.g., familywise error (FWE) control, $\alpha = .05$, and Tukey's (1953) method, respectively]. Indeed, a recent survey of the statistical practices of educational and psychological researchers indicates that of the many MCPs that are available, the Tukey and Scheffe (1959) methods are most preferred (Keselman et al., 1998).

With respect to the selection of a MCP, the researcher must be aware that his/her choice can often significantly affect the results of the experiment. For example, many MCPs (e.g., those that are based on traditional test statistics) are not appropriate (and may lead to incorrect decisions) when assumptions of the test statistics are not met (e.g., normality, variance homogeneity). Furthermore, several MCPs have recently been proposed that according to published results and/or statistical theory significantly improve on the properties (e.g., power) of existing procedures, while still maintaining the specified error rate at or below α .

Therefore, the goal of this chapter is to describe some of the newer MCPs within the context of one-way completely randomized designs when validity

assumptions are satisfied, as well as when the assumptions are not satisfied. That is, our goal is to help popularize newer procedures, procedures which should provide researchers with more robust and/or more powerful tests of their pairwise comparison null hypotheses.

It is also important to note that the MCPs that are presented in our paper were also selected for discussion, by-in-large, because researchers can, in most cases, obtain numerical results with a statistical package, and in particular, through the SAS (1999) system of computer programs. The SAS system (see Westfall et al., 1999) presents a comprehensive up-to-date array of MCPs. Accordingly, we acknowledge at the beginning of our presentation that some of the material we present follows closely Westfall et al.'s presentation, however, our paper focuses on MCPs for examining all possible pairwise comparisons between treatment group means. We also present procedures that are not available through the SAS system. In particular, we discuss a number of procedures that we believe are either new and interesting ways of examining pairwise comparisons (e.g., the model comparison approach of Dayton, 1998) or have been shown to be insensitive to the usual assumptions associated with some of the procedures discussed by Westfall et al. (e.g., MCPs based on robust estimators).

Type I Error Control

Researchers who test a hypothesis concerning mean differences between two treatment groups are often faced with the task of specifying a significance level, or decision criterion, for determining whether or not the difference is significant. The level of significance specifies the maximum probability of rejecting the null hypothesis when it is true (i.e., committing a Type I error). As α

decreases researchers can be more confident that rejection of the null hypothesis signifies a true difference between population means, although the probability of not detecting a false null hypothesis (i.e., a Type II error) increases. Researchers faced with the difficult, yet important, task of quantifying the relative importance of Type I and Type II errors have traditionally selected some accepted level of significance, for example $\alpha = .05$.

However, determining how to control Type I errors is much less simple when multiple tests of significance (e.g., all possible pairwise comparisons between group means) will be computed. This is because when multiple tests of significance are computed, how one chooses to control Type I errors can affect whether one can conclude that effects are statistically significant or not. Choosing among the various strategies that one can adopt to control Type I errors could be based on how one wishes to deal with the multiplicity of testing issue.

The multiplicity problem in statistical inference refers to selecting the statistically significant findings from a large set of findings (tests) to support one's research hypotheses. Selecting the statistically significant findings from a larger pool of results that also contain nonsignificant findings is problematic since when multiple tests of significance are computed, the probability that at least one will be significant by chance alone increases with the number of tests examined.

Discussions on how to deal with multiplicity of testing have permeated many literatures for decades and continue to this day. In one camp are those who believe that the occurrence of any false positive must be guarded at all costs (see Games, 1971; Miller, 1981; Ryan, 1959, 1960, 1962; Westfall & Young, 1993). That is, as promulgated by Thomas Ryan, pursuing a false lead can result in the waste of much time and expense, and is an error of inference that

accordingly should be stringently controlled. Those in this camp deal with the multiplicity issue by setting α for the entire set of tests computed.

For example, in the pairwise multiple comparison problem, Tukey's (1953) MCP uses a critical value wherein the probability of making at least one Type I error in the set of pairwise comparisons tests is equal to α . This type of control has been referred to in the literature as experimentwise or FWE control. These respective terms come from setting a level of significance over all tests computed in an experiment, hence experimentwise control, or setting the level of significance over a set (family) of conceptually related tests, hence FWE control. Multiple comparisonists seem to have settled on the familywise label. Thus, in the remainder of the paper, when we speak about overall error control, we are referring to FWE. As indicated, for the set of pairwise tests, Tukey's procedure sets a FWE for the family consisting of all pairwise comparisons.

Those in the opposing camp maintain that stringent Type I error control results in a loss of statistical power and consequently important treatment effects go undetected (see Rothman, 1990; Saville, 1990; Wilson, 1962). Members of this camp typically believe the error rate should be set per comparison [the probability of rejecting a given comparison] (hereafter referred to as the comparisonwise error-CWE rate) and usually recommend a five percent level of significance, allowing the overall error rate (i.e., FWE) to inflate with the number of tests computed. In effect, those who adopt comparisonwise control ignore the multiplicity issue.

For example, a researcher comparing four groups ($J = 4$) may be interested in determining if there are significant pairwise mean differences between any of the groups. If the probability of committing a Type I error is set at α for each comparison, then the probability that at least one Type I error is committed over all $C = J(J - 1)/2$ pairwise comparisons can be much higher than

α . On the other hand, if the probability of committing a Type I error is set at α for the entire family of pairwise comparisons, then the probability of committing a Type I error for each of the C comparisons can be much lower than α . Clearly then, the conclusions of an experiment can be greatly affected by the level of significance and unit of analysis over which Type I error control is imposed.

As indicated, several different error rates have been proposed in the multiple comparison literature. The majority of discussion in the literature has focused on the FWE and CWE rates (e.g., Kirk, 1995; Ryan, 1959; Miller, 1981; Toothaker, 1991; Tukey, 1953), although other error rates, such as the false discovery rate (FDR) also have been proposed (e.g., Benjamini & Hochberg, 1995).

The FWE rate relates to a family (containing, in general, say k elements) of comparisons. A family of comparisons, as we indicated, refers to a set of conceptually related comparisons, e.g., all possible pairwise comparisons, all possible complex comparisons, trend comparisons, etc. As Miller (1981) points out, specification of a family of comparisons, being self defined by the researcher, can vary depending on the research paradigm. For example, in the context of a one-way design, numerous families can be defined: A family of all comparisons performed on the data, a family of all pairwise comparisons, a family of all complex comparisons. (Readers should keep in mind that if multiple families of comparisons are defined [e.g., one for pairwise comparisons and one for complex comparisons], then given that erroneous conclusions can be reached within each family, the overall Type I FWE rate will be a function of the multiple subfamilywise rates.) Researchers may find helpful the guidelines offered by Westfall and Young (1993, p. 220) for specification of a family; they include:

- The questions asked form a natural and coherent unit. For example,

they all result from a single experiment.

- All tests are considered simultaneously. (For example, when the results of a large study are summarized for publication, all tests are considered simultaneously. Usually, only a subset of the collection is selected for display, but the entire collection should constitute the “family” to avoid selection effects.)
- It is considered *a priori* probable that many or all members of the “family” of null hypotheses are in fact true.

Specifying family size is a very important component of multiple testing (In this chapter family size is all possible pairwise comparisons.). As Westfall et al. (1999, p. 10) note, differences in conclusions reached from statistical analyses that control for multiplicity of testing (FWE) and those that do not (CWE) are directly related to family size. That is, the larger the family size, the less likely individual tests will be found to be statistically significant with familywise control. Accordingly, to achieve as much sensitivity as possible to detect true differences and yet maintain control over multiplicity effects, Westfall et al. recommend that researchers “choose smaller, more focused families rather than broad ones, and (to avoid cheating) that such determination must be made *a priori*...” (p. 10).

Definitions of the comparisonwise and familywise error rates appear in many sources (e.g., Kirk, 1995; Ryan, 1959; Miller, 1981; Toothaker, 1991; Tukey, 1953). Nonetheless, for completeness, we provide the reader with definitions of these rates of error. The comparisonwise error rate is defined as

$$\text{CWE} = P(\text{Reject } H_0 \mid H_0 \text{ is true}).$$

That is, the comparisonwise error rate is the usual probability of rejecting a null hypothesis (H_0), given that (I) the null hypothesis is true. On the other hand, the

familywise error rate for multiple tests of significance in which some hypotheses ($H_{0j_1}, H_{0j_2}, \dots, H_{0j_m}$) are true and the remaining ($k - m$) are false is given by

$$\text{FWE} = P(\text{Reject at least one of } H_{0j_1}, H_{0j_2}, \dots, H_{0j_m} \mid H_{0j_1}, H_{0j_2}, \dots, H_{0j_m} \text{ all are true}).$$

This rate of error obviously will depend on which hypotheses are true and which are false within any particular application. To deal with this ambiguity, a number of authors (see Westfall et al., 1999) define FWE control to be the maximum FWE rate, which occurs, when all null hypotheses are true.

Not only does the FWE rate depend on the number of null hypotheses that are true but as well on the distributional characteristics of the data and the correlations among the test statistics. Because of this, an assortment of MCPs have been developed, each intended to provide FWE control.

Controlling the FWE rate has been recommended by many researchers (e.g., Hancock & Klockars, 1996; Petrinovich & Hardyck, 1969; Ryan, 1959, 1962; Tukey, 1953) and is "the most commonly endorsed approach to accomplishing Type I error control" (Seaman, Levin & Serlin, 1991, p. 577). Keselman et al. (1998) report that approximately 85 percent of researchers conducting pairwise comparisons adopt some form of FWE control.

Although many MCPs purport to control FWE, some provide 'strong' FWE control while others only provide 'weak' FWE control. Procedures are said to provide strong control if FWE is maintained across all null hypotheses; that is, under the complete null configuration ($\mu_1 = \mu_2 = \dots = \mu_J$) and all possible partial null configurations (An example of a partial null hypothesis is $\mu_1 = \mu_2 = \dots = \mu_{J-1} \neq \mu_J$). Weak control, on the other hand, only provides protection for the complete null hypothesis, that is, not for all partial null hypotheses as well.

The distinction between strong and weak FWE control is important because as Westfall et al. (1999) note, the two types of FWE control, in fact, control different error rates. Weak control only controls the Type I error rate for falsely rejecting the complete null hypothesis and accordingly allows the rate to exceed, say 5%, for the composite null hypotheses. On the other hand, strong control sets the error rate at, say 5%, for all (component) hypotheses. For example, if $CWE = 1 - (1 - 0.05)^{1/k}$, the familywise rate is controlled in a strong sense for testing k independent tests. Examples of MCPs that only weakly control FWE are the Newman (1939)-Keuls (1952) and Duncan (1955) procedures.

False Discovery Rate Control. Work in the area of multiple hypothesis testing is far from static, and one of the newer interesting contributions to this area is an alternative conceptualization for defining errors in the multiple testing problem; that is the FDR, presented by Benjamini and Hochberg (1995). FDR is defined by these authors as the expected proportion of the number of erroneous rejections to the total number of rejections.

We elaborate on the FDR within the context of pairwise comparisons. Suppose we have J means, $\mu_1, \mu_2, \dots, \mu_J$, and our interest is in testing the family of C pairwise hypotheses, $H_0: \mu_j - \mu_{j'} = 0$, of which m_0 are true. Let S equal the number of correctly rejected hypotheses from the set of R rejections; the number of falsely rejected pairs will be V . In terms of the random variable V , the comparisonwise error rate is $E(V/C)$, while the familywise rate is given by $P(V \geq 1)$. Thus, testing each and every comparison at α guarantees that $E(V/C) \leq \alpha$, while testing each and every comparison at α/C (Bonferroni) guarantees $P(V \geq 1) \leq \alpha$.

According to Benjamini and Hochberg (1995) the proportion of errors committed by falsely rejecting null hypotheses can be expressed through the random variable $Q = V/(V + S)$, that is, the proportion of rejected hypotheses which are erroneously rejected. (It is important to note that Q is defined to be zero when $R = 0$; that is, the error rate is zero when there are no rejections.) FDR was defined by Benjamini and Hochberg as the mean of Q , that is

$$E(Q) = E\left(\frac{V}{V+S}\right) = E\left(\frac{V}{R}\right), \text{ or}$$

$$E(Q) = E\left(\frac{\text{Number of false rejections}}{\text{Number of rejections}}\right).$$

That is, FDR is the mean of the proportion of the falsely declared pairwise tests among all pairwise tests declared significant.

As Benjamini and Hochberg (1995) indicate, this error rate has a number of important properties:

(a) If $\mu_1 = \mu_2 = \dots = \mu_J$, then all C pairwise comparisons truly equal zero, and therefore the FDR is equivalent to the familywise rate; that is, in the case of the complete null being true, FDR control implies familywise control. Specifically, in the case of the complete null hypothesis being true, $S = 0$ and therefore $V = R$. So, if $V = 0$, then $Q = 0$, and if $V > 0$ then $Q = 1$ and accordingly $P(V \geq 1) = E(Q)$.

(b) When $m_0 < C$, the FDR is smaller than or equal to the familywise rate of error. For example, when $m_0 < C$, the FDR is smaller than or equal to the familywise rate of error because in this case, $FWE = P(R \geq 1) \geq E(V/R) = E(Q)$. This indicates that if the familywise rate is controlled for a procedure, then FDR is

as well. Moreover, if one adopts a procedure which provides strong (i.e., over all possible mean configurations) FDR control, rather than strong familywise control, then based on the preceding relationship, a gain in power can be expected.

(c) V/R tends to be smaller when there are fewer pairs of equal means and when the nonequal pairs are more divergent, resulting in a greater differences in FDR and the familywise value and thus a greater likelihood of increased power by adopting FDR control.

In addition to these characteristics, Benjamini and Hochberg (1995) provide a number of illustrations where FDR control seems more reasonable than familywise or comparisonwise control. Exploratory research, for example, would be one area of application for FDR control. That is, in new areas of inquiry where we are merely trying to see what parameters might be important for the phenomenon under investigation, a few errors of inference should be tolerable; thus, one can reasonably adopt the less stringent FDR method of control which does not completely ignore the multiple testing problem, as does comparisonwise control, and yet, provides greater sensitivity than familywise control. Only at later stages in the development of our conceptual formulations does one need more stringent familywise control. Another area where FDR control might be preferred over familywise control, suggested by Benjamini and Hochberg (1995), would be when two treatments (say, treatments for dyslexia) are being compared in multiple subgroups (say, kids of different ages). In studies of this sort, where an overall decision regarding the efficacy of the treatment is not of interest but, rather where separate recommendations would be made within each subgroup, researchers likely should be willing to tolerate a few errors of inference and accordingly would profit from adopting FDR rather than familywise control.

Adjusted p-values

As indicated, FWE control is the rate of error control that is currently favored by social science researchers. In its typical application, researchers compare a test statistic to a FWE critical value. Another approach for assessing statistical significance is with adjusted p-values, \tilde{p}_c , $c = 1, \dots, C$ (Westfall et al., 1999; Westfall & Wolfinger, 1997; Westfall & Young, 1993). As Westfall and Young note " \tilde{p}_c is the smallest significance level for which one still rejects a given hypothesis in a family, given a particular (familywise) controlling procedure (p. 11)." Thus, authors do not need to look up (or determine) FWE critical values and moreover consumers of these findings can apply their own assessment of statistical significance from the adjusted p-value rather than from the standard (i.e., FWE) significance level of the experimenter. The latter point is consistent with the current practice of reporting a p-value for a single test statistic rather than stating that the 'result was significant' at the say .05 value; that is, current practice allows the consumer to take a p-value and apply his/her own personal standard of significance in judging the importance of the finding. For example, if $\tilde{p}_c = 0.09$, the researcher/reader can conclude that the test is statistically significant at the FWE = 0.10 level, but not at the FWE = 0.05 level.

To illustrate the calculation of an adjusted p-value consider the usual Bonferroni procedure. In its usual application, H_{0c} is rejected if the p-value is less than or equal to α/C . Note that this is equivalent to rejecting any H_{0c} for which $C \cdot p_c$ is less than or equal to α . Therefore, Bonferroni adjusted p-values are:

$$\tilde{p}_c = \begin{cases} C \cdot p_c & \text{if } C \cdot p_c \leq 1 \\ 1 & \text{if } C \cdot p_c > 1. \end{cases}$$

Adjusted p-values are provided by the SAS (1999) system for many popular MCPs (see Westfall et al., 1999).

Power

Just as the rate of Type I error control can be viewed from varied perspectives when there are multiple tests of significance, the power to detect nonnull hypotheses also can be conceptualized in many ways. Over the years many different conceptualizations of power for (pairwise) comparisons have appeared in the literature (e.g., all-pairs, any pair, per-pair); our presentation, however, will be based on the work of Westfall et al. (1999, pp. 137-144).

According to Westfall et al. (1999), when multiple tests of significance are (to be) examined, power can be defined from four different perspectives: (1) complete power; (2) minimal power; (3) individual power; and (4) proportional power. The definitions they provide are:

- Complete Power-- $P(\text{reject all } H_c\text{s that are false})$
- Minimal Power-- $P(\text{reject at least one } H_c \text{ that is false})$
- Individual Power-- $P(\text{reject a particular } H_c \text{ that is false})$
- Proportional Power (average proportion of false H_c s that are rejected).

Complete power is the probability of detecting all nonnull hypotheses, a very desirable outcome, though very difficult to achieve, even in very well controlled and executed research designs. For example, as Westfall et al. note, if ten independent tests of significance each have individually a power of 0.8 to detect a nonnull effect, the power to detect them all equals $(.8)^{10} = 0.107!$ Minimal power, on the other hand, is the probability of detecting at least one nonnull

hypothesis and corresponds conceptually to the Type I FWE rate. Individual power is the probability of detecting a particular nonnull hypothesis, with a MCP critical value. However, it is questionable why a researcher would use a MCP critical value if he/she were just interested in detecting one particular nonnull difference. That is, in such a situation why control for multiplicity? Lastly, proportional power indicates what proportion of false null hypotheses one is likely to detect.

Types of MCPs

MCPs can examine pairwise hypotheses either simultaneously, or sequentially. A simultaneous MCP conducts all comparisons regardless of whether the omnibus test, or any other comparison, is significant (or not significant) using a constant critical value. Such procedures are frequently referred to as simultaneous test procedures (STPs) (see Einot & Gabriel, 1975). A sequential (stepwise) MCP considers either the significance of the omnibus test or the significance of other comparisons (or both) in evaluating the significance of a particular comparison; multiple critical values are used to assess statistical significance. MCPs that require a significant omnibus test in order to conduct pairwise comparisons have been referred to as protected tests.

MCPs that consider the significance of other comparisons when evaluating the significance of a particular comparison can be either step-down or step-up procedures. Step-down procedures begin by testing the most extreme test statistic and nonsignificance of the most extreme test statistics implies nonsignificance for less extreme test statistics. Step-up procedures begin by testing the least extreme test statistic and significance of least extreme test statistics can imply significance for larger test statistics. In the equal sample sizes

case, if a smaller pairwise difference is statistically significant, so is a larger pairwise difference, and conversely. However, in the unequal sample size cases, one can have a smaller pairwise difference be significant and a larger pairwise difference non-significant if the sample sizes for the means comprising the smaller difference are much larger than the sample sizes for the means comprising the larger difference.

One additional point regarding STP and stepwise procedures is important to note. STPs allow researchers to examine simultaneous intervals around the statistics of interest whereas stepwise procedures do not (see however, Bofinger, Hayter & Liu, 1993).

DESIGN, NOTATION AND TEST STATISTICS

A mathematical model that can be adopted when examining pairwise mean differences in a one-way completely randomized design is:

$$Y_{ij} = \mu_j + \epsilon_{ij},$$

where Y_{ij} is the score of the i th subject ($i = 1, \dots, n$) in the j th group ($\sum_j n = N$), μ_j is the j th group mean, and ϵ_{ij} is the random error for the i th subject in the j th group. In the typical application of the model, it is assumed that the ϵ_{ij} s are normally and independently distributed and that the treatment group variances (σ_j^2 s) are equal. Relevant sample estimates include

$$\hat{\mu}_j = \bar{Y}_j = \sum_{i=1}^n Y_{ij}/n \quad \text{and} \quad \hat{\sigma}^2 = \text{MSE} = \sum_{j=1}^J \sum_{i=1}^n (Y_{ij} - \bar{Y}_j)^2 / J(n - 1).$$

A confidence interval for a pairwise difference $\mu_j - \mu_{j'}$ has the form

$$\bar{Y}_j - \bar{Y}_{j'} \pm c_\alpha \hat{\sigma} \sqrt{2/n} ,$$

where c_α is selected such that $FWE = \alpha$. In the case of all possible pairwise comparisons, one needs a c_α for the set such that they simultaneously surround the true differences with a specified level of significance. That is, for all $j \neq j'$, c_α must satisfy

$$P(\bar{Y}_j - \bar{Y}_{j'} - c_\alpha \hat{\sigma} \sqrt{2/n} \leq \mu_j - \mu_{j'} \leq \bar{Y}_j - \bar{Y}_{j'} + c_\alpha \hat{\sigma} \sqrt{2/n}) = 1 - \alpha .$$

The interval is equivalent to

$$P(\max_{j,j'} \frac{|(\bar{Y}_j - \mu_j) - (\bar{Y}_{j'} - \mu_{j'})|}{\hat{\sigma} \sqrt{2/n}} \leq c_\alpha) = 1 - \alpha ,$$

where max stands for maximum. Evident from this last expression is that c_α is related to the Studentized range distribution (see Scheffe, 1959, p. 28). Specifically, if z_1, z_2, \dots, z_n are standard normal independent random variates and V is a random variable, independent of the z s, and is chi-square distributed with df degrees of freedom, then

$$q_{(J, df)} = \max_{j,j'} \frac{|z_j - z_{j'}|}{\sqrt{V/df}}$$

has a Studentized range distribution with parameters J and df . Another relation that should be noted, is that it can be shown that c_α satisfies

$$P(q_{J, J(n-1)} / \sqrt{2} \leq c_\alpha) = 1 - \alpha .$$

A hypothesis for the comparison ($H_c: \mu_j - \mu_{j'} = 0$) can be examined with the test statistic:

$$t_c = \bar{Y}_j - \bar{Y}_{j'} / (2 \text{MSE}/n)^{1/2}.$$

The preceding can also be specified from a general linear model perspective (see Westfall et al., 1999, Chapter 5). That is, the data can be conceived as coming from the model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where \mathbf{Y} is an $N \times 1$ observational vector, \mathbf{X} is the $N \times p$ design matrix, β is the $p \times 1$ vector of unknown parameters and ϵ is the $N \times 1$ vector of random errors.

The usual assumptions to the model relate to the characteristics of the random errors. Specifically, it is assumed that the $\epsilon_1, \epsilon_2, \dots, \epsilon_N$ all (a) have a mean of zero, (b) have common variance, σ^2 , (c) are independent random variables, and (d) are normally distributed. Important estimates of the model are obtained in the following manner:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}. \\ \hat{\sigma}^2 &= (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})/df,\end{aligned}$$

where $(\bullet)^{-}$ denotes a generalized inverse and $df = (N - \text{rank } \mathbf{X})$ (see Westfall et al., 1999, p. 87).

One can specify estimable (see Scheffe, 1959, p. 13) functions of the parameters, $\mathbf{c}'\beta$, where for this chapter, the functions would be the pairwise

comparisons, such as say $\mathbf{c}'\beta = \mu_1 - \mu_2$, where $\mathbf{c}' = (0 \ 1 \ -1 \ 0 \ \dots \ 0)$, which would be estimated by $\mathbf{c}'\hat{\beta}$.

To form simultaneous intervals or obtain simultaneous tests of the estimable functions (pairwise comparisons), one needs to know the dependence structures of the estimable functions. As Westfall et al. (1999) point out, simultaneous inferences rely on the joint distribution of the quantities

$$T_i = \frac{\mathbf{c}'_i\hat{\beta} - \mathbf{c}'_i\beta}{\hat{\sigma}\sqrt{\mathbf{c}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_i}},$$

where $\hat{\sigma}\sqrt{\mathbf{c}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_i}$ is the standard error (SE) of $\mathbf{c}'_i\hat{\beta}$. The joint distribution of the T_i is a multivariate t distribution, with $df = (N - \text{rank}\mathbf{X})$ and dispersion matrix $\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{C} = (\mathbf{c}_1, \dots, \mathbf{c}_k)$ and \mathbf{D} is a diagonal matrix where the i th element equals $\mathbf{c}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_i$.

Confidence intervals of the estimable functions have the form

$$\mathbf{c}'_i\hat{\beta} \pm c_\alpha \text{SE}(\mathbf{c}'_i\hat{\beta}),$$

where c_α is chosen such that the FWE = α . Bonferroni-type methods can be used to set the simultaneous intervals such the confidence coefficient will not exceed $1 - \alpha$. However, because the Bonferroni procedure is overly conservative, we know that these intervals will simultaneously contain the true values more than $100(1 - \alpha)$ percent of the time. This approach however can be improved by taking the correlational structure among the estimable functions into account, that is, by setting a simultaneous critical value via the multivariate t distribution. That is,

$$P\left(\left|\frac{\mathbf{c}_i' \hat{\beta} - \mathbf{c}_i' \beta}{\hat{\sigma} \sqrt{\mathbf{c}_i' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{c}_i}}\right| \leq c_\alpha, \text{ for all } i\right) = 1 - \alpha.$$

As Westfall et al. (1999, p. 89) note, “The value of c_α is the $1 - \alpha$ quantile of the distribution of $\max_i |T_i|$, where the vector $\mathbf{T}' = (T_1, \dots, T_k)$ has the multivariate t distribution.”

MCPs FOR NORMALLY DISTRIBUTED DATA/ HOMOGENEOUS POPULATION VARIANCES

Tukey. Tukey (1953) proposed a STP for all pairwise comparisons in what Toothaker described as possibly “the most frequently cited unpublished paper in the history of statistics” (1991, p. 41). Tukey’s MCP uses a critical value obtained from the Studentized range distribution. The procedure accounts for dependencies (correlations) among the pairwise comparisons in deriving a simultaneous critical value. In particular, statistical significance, with FWE control, is assessed by comparing

$$|t_{j,j'}| > q_{(J, J(n-1))} / \sqrt{2} .$$

Tukey’s procedure can be implemented in SASs (1999) general linear model (GLM) program.

At this juncture we can illustrate the use of adjusted p-values and refer to various SAS programs that can be used to obtain relevant information. As Westfall et al. (1999, p. 52) note, to find the exact significance level for Tukey’s (1953) test one must determine the value of $c_\alpha = q_{(J, J(n-1))} / \sqrt{2}$ for which $(\bar{Y}_j - \bar{Y}_{j'}) + c_\alpha \hat{\sigma} \sqrt{2/n} = 0$. It can be seen that the solution is $c_\alpha = |t_{j,j'}|$. Accordingly, the adjusted p-value is

$$\tilde{p}_{j,j'} = P(q_{(J, J(n-1))}/\sqrt{2} |t_{j,j'}|).$$

Westfall et al. (1999, p. 52) enumerate the GLM syntax, with an accompanying numerical example, to obtain adjusted p-values for Tukey's test.

Recall that we defined various power rates in the multiple comparison problem: complete power, minimal power, individual power, and proportional power. SAS software allows users to compute these values.

To illustrate, consider the power to detect a particular pairwise difference, that is, individual power. To detect a difference (δ) between μ_j and $\mu_{j'}$ either

$$t_{j,j'} = \frac{\bar{Y}_j - \bar{Y}_{j'}}{\hat{\sigma}\sqrt{2/n}} > c_\alpha$$

or

$$t_{j,j'} = \frac{\bar{Y}_j - \bar{Y}_{j'}}{\hat{\sigma}\sqrt{2/n}} < -c_\alpha,$$

where, as indicated, $c_\alpha = q_{(J, J(n-1))}/\sqrt{2}$. The individual power, therefore, is the sum of the probabilities of these two events. These two probabilities can be obtained from SAS's PROBT function; that is, PROBT calculates probabilities for the noncentral Student t distribution with $J(n-1)$ df and noncentrality parameter $(\delta/\sigma)\sqrt{n/2}$. Westfall et al. (1999, p. 140) provide a macro (%Individual Power) for obtaining numerical results. These authors also provide a macro (%SimPower) that computes complete, minimal and proportional power.

Fisher-Hayter. Fisher (1935) proposed conducting multiple t-tests on the C pairwise comparisons following rejection of the omnibus ANOVA null hypothesis

(See Keselman, Games & Rogan, 1979; Kirk, 1995). The pairwise null hypotheses are assessed for statistical significance by referring t_c to $t_{(\alpha/2, \nu)}$, where $t_{(\alpha/2, \nu)}$ is the upper $100(1-\alpha)$ percentile from Student's distribution with parameter ν . If the ANOVA F is nonsignificant, comparisons among means are not conducted; that is, the pairwise hypotheses are retained as null.

It should be noted that Fisher's (1935) Least Significant Difference (LSD) procedure only provides Type I error protection via the level of significance associated with the ANOVA null hypothesis, that is, the complete null hypothesis. For other configurations of means not specified under the ANOVA null hypothesis (e.g., $\mu_1 = \mu_2 = \dots = \mu_{J-1} < \mu_J$), the rate of familywise Type I error can be much in excess of the level of significance (Hayter, 1986; Hochberg & Tamhane, 1987; Keselman et al., 1991; Ryan, 1980).

Hayter (1986) proved that the maximum FWE for all partitions of the means, which occurs when $J-1$ of the means are equal and the remaining one is very disparate from this group, is equal to $P(q_{(J-1, \nu)} > 2 t_{(\alpha/2, \nu)})$. One can see that for $J > 3$ the maximum FWE will exceed the level of significance. In fact, Hayter (1986) showed that for $\nu = \infty$ and $\alpha = .05$, FWE attains values of .1222 and .9044 for $J = 4$ and $J = 8$, respectively. Thus, this usual form of the LSD does not provide a satisfactory two stage procedure for researchers when $J > 3$.

Accordingly, Hayter (1986) proposed a modification to Fisher's LSD that would provide strong control over FWE. Like the LSD procedure, no comparisons are tested unless the omnibus test is significant. If the omnibus test is significant, then H_c is rejected if:

$$|t_c| > q_{(J-1, df)} / \sqrt{2}.$$

Studentized range critical values can be obtained through SASs PROBMC (see Westfall et al., 1999, p. 46).

It should be noted that many authors recommend Fisher's two-stage test for pairwise comparisons when $J = 3$ (see Keselman, Cribbie & Holland, 1999; Levin, Serlin, & Seaman, 1994). These recommendations are based on Type I error control, power and ease of computation issues.

PROCEDURES THAT CONTROL THE FALSE DISCOVERY RATE

BH. As previously indicated, Benjamini and Hochberg (1995) proposed controlling the FDR, instead of the often conservative FWE or the often liberal CWE. For FDR control, the p_c -values are ordered (smallest to largest) p_1, \dots, p_C , and for any $c = C, C-1, \dots, 1$, if $p_c \leq \alpha/(C-c+1)$, reject all $H_{c'}$ ($c' \leq c$).

BH has been shown to control the FWE for several situations of dependent tests, that is, for a wide variety of multivariate distributions that make BH applicable to most testing situations social scientists might encounter (see Sarkar, 1988; Sarkar & Chang, 1997). In addition, simulation studies comparing the power of the BH procedure to several FWE controlling procedures have shown that as the number of treatment groups increases (beyond $J = 4$), the power advantage of the BH procedure over the FWE controlling procedures becomes increasingly large (Benjamini et al., 1994; Keselman et al., 1999). The power of FWE controlling procedures is highly dependent on the family size (i.e., number of comparisons), decreasing rapidly with larger families (Holland & Cheung, 2000; Miller, 1981). Therefore, control of the FDR results in more power than FWE controlling procedures in experiments with many treatment groups, but yet provides more control over Type I errors than CWE controlling procedures.

Statistical significance can be assessed once again with adjusted p-values. For the FDR method of control, the adjusted p-values would equal

$$\begin{aligned}\tilde{p}_{(C)} &= p_{(C)} \\ \tilde{p}_{(C-1)} &= \min(\tilde{p}_{(C)}, [C/(C-1)] \cdot p_{(C-1)}) \\ &\vdots \\ \tilde{p}_{(C-c)} &= \min(\tilde{p}_{(C-c+1)}, [C/(C-c)] \cdot p_{(C-c)}) \\ &\vdots \\ \tilde{p}_{(1)} &= \min(\tilde{p}_{(2)}, Cp_{(1)})\end{aligned}$$

SASs (1999) MULTTEST program can be used to obtain these adjusted p-values (See Westfall et al, 1999, p. 35).

BH-A. Benjamini and Hochberg (in press) also presented a modified (adaptive) version of their original procedure that utilizes the data to estimate the number of true H_c s. [The adaptive BH procedure has only been demonstrated, not proven, to control FDR, and only in the independent case.] With the original procedure, when the number of true null hypotheses (C^T) is less than the total number of hypotheses, the FDR rate is controlled at a level less than that specified (α).

To compute the BH-A procedure, the p_c -values are ordered (smallest to largest) p_1, \dots, p_C , and for any $c = C, C-1, \dots, 1$, if $p_c \leq \alpha(c/C)$, reject all $H_{c'}$ ($c' \leq c$), as in the BH procedure. If all H_c s are retained, testing stops. If any H_c is rejected with the criterion of the BH procedure, then testing continues by estimating the slopes $S_c = (1-p_c) / (C+1-c)$, where $c = 1, \dots, C$. Then, for any $c = C, C-1, \dots, 1$, if $p_c \leq \alpha(c/C^T)$, reject all $H_{c'}$ ($c' \leq c$), where $C^T = \min[(1/S^*) + 1, C]$, $[x]$ is largest integer less than or equal to x and S^* is the minimum value of S_c such that $S_c < S_{c-1}$. If all $S_c > S_{c-1}$, S^* is set at C .

One disadvantage of the BH-A procedure, noted by both Benjamini and Hochberg (in press) and Holland and Cheung (2000), is that it is possible for an H_c to be rejected with $p_c > \alpha$. Therefore, it is suggested, by both authors, that H_c only be rejected if: a) the hypothesis satisfies the rejection criterion of the BH-A; and b) $p_c \leq \alpha$. To illustrate this procedure, assume a researcher has conducted a study with $J = 4$ and $\alpha = .05$. The ordered p-values associated with the $C = 6$ pairwise comparisons are: $p_1 = .0014$, $p_2 = .0044$, $p_3 = .0097$, $p_4 = .0145$, $p_5 = .0490$, and $p_6 = .1239$. The first stage of the BH-A procedure would involve comparing $p_6 = .1239$ to $\alpha(c/C) = .05(6/6) = .05$. Since $.1239 > .05$, the procedure would continue by comparing $p_5 = .0490$ to $\alpha(c/C) = .05(5/6) = .0417$. Again, since $.0490 > .0417$, the procedure would continue by comparing $p_4 = .0145$ to $\alpha(c/C) = .05(4/6) = .0333$. Since $.0145 < .0333$, H_4 would be rejected. Because at least one H_c was rejected during the first stage, testing continues by estimating each of the slopes, $S_c = (1-p_c)/(C-c+1)$, for $c = 1, \dots, C$. The calculated slopes for this example are: $S_1 = .1664$, $S_2 = .1991$, $S_3 = .2475$, $S_4 = .3285$, $S_5 = .4755$ and $S_6 = .8761$. Given that all $S_c > S_{c-1}$, S^* is set at $C = 6$. The estimated number of true nulls is then determined by $C^T = \min [(1/S^*) + 1, C] = \min[(1/6) + 1, 6] = \min[1.1667, 6] = 1$. Therefore, the BH-A procedure would compare $p_6 = .1239$ to $\alpha(c/C^T) = .05(6/1) = .30$. Since $.1239 < .30$, but $.1239 > \alpha$, H_6 would not be rejected and the procedure would continue by comparing $p_5 = .0490$ to $\alpha(c/C^T) = .05(5/1) = .25$. Since $.0490 < .25$ and $.0490 < \alpha$, H_5 would be rejected; in addition, all $H_{c'}$ would also be rejected (i.e., H_1, H_2, H_3 , and H_4).

CLOSED TESTING SEQUENTIAL MCPs

As we indicated previously, researchers can adopt stepwise procedures when examining all possible pairwise comparisons, and typically they provide greater sensitivity to detect differences than do STPs, e.g., Tukey's (1953) method, while still maintaining strong FWE control. In this section, we present some theory and methods related to closed testing sequential MCPs that can be obtained through the SAS system of programs.

As Westfall et al. (1999, p. 149) note, it was in the past two decades that a unified approach to stepwise testing has evolved. The unifying concept has been the closure principle. MCPs based on this principle have been designated as closed testing procedures. These methods are designated as closed testing procedures because they address families of hypotheses that are closed under intersection (\cap). By definition, a closed family "is one for which any subset intersection hypothesis involving members of the family of tests is also a member of the family" (p. 150).

To illustrate, suppose that one wants to test all possible pairwise comparisons among four means; that is, six pairwise tests. The closed set is formed by taking all possible intersections among the pairwise hypotheses. An important point to remember is that a hypothesis that is formed by an intersection of two or more hypotheses is true if and only if all of the components are true. For example, if we intersect $H_{2,3}: \mu_2 = \mu_3$ with say $H_{2,4}: \mu_2 = \mu_4$, we obtain $H_{2,3,4}: \mu_2 = \mu_3 = \mu_4$ because if $\mu_2 = \mu_3$ and $\mu_2 = \mu_4$ then it must be the case that $\mu_2 = \mu_3 = \mu_4$. Forming all possible intersections we get fourteen hypotheses in the closed family:

- The six pairwise homogeneity hypotheses- $H_{1,2}: \mu_1 = \mu_2$, $H_{1,3}: \mu_1 = \mu_3$, $H_{1,4}: \mu_1 = \mu_4$, $H_{2,3}: \mu_2 = \mu_3$, $H_{2,4}: \mu_2 = \mu_4$, $H_{3,4}: \mu_3 = \mu_4$,

- The four three means homogeneity hypotheses- $H_{1,2,3}: \mu_1 = \mu_2 = \mu_3$, $H_{1,2,4}: \mu_1 = \mu_2 = \mu_4$, $H_{1,3,4}: \mu_1 = \mu_3 = \mu_4$, $H_{2,3,4}: \mu_2 = \mu_3 = \mu_4$,
- The one four means homogeneity hypothesis- $H_{1,2,3,4}: \mu_1 = \mu_2 = \mu_3 = \mu_4$, and
- The three subset intersection hypotheses- $H_{(1,2)\cap(3,4)}: \mu_1 = \mu_2$ and $\mu_3 = \mu_4$, $H_{(1,3)\cap(2,4)}: \mu_1 = \mu_3$ and $\mu_2 = \mu_4$, $H_{(1,4)\cap(2,3)}: \mu_1 = \mu_4$ and $\mu_2 = \mu_3$.

Because of the hierarchical structure of the hypotheses, there are a number of important implications related to the stepwise testing format. Specifically, if $H_{(1,2)\cap(3,4)}: \mu_1 = \mu_2$ and $\mu_3 = \mu_4$ is true, then it follows that both $H_{1,2}: \mu_1 = \mu_2$ and $H_{3,4}: \mu_3 = \mu_4$ are necessarily true. That is, the truth of $H_{(1,2)\cap(3,4)}$ implies the truths of $H_{1,2}$ and $H_{3,4}$. These types of implications for closed testing procedures are referred to as the coherence property of these methods. Coherence states that if H^+ implies H^{++} , then whenever H^+ is retained so must H^{++} .

The closed testing principle has led to a way of performing multiple tests of significance such that FWE is strongly controlled with results that are coherent. In particular, according to Marcus, Peritz, and Gabriel (1976) and as enumerated by Westfall et al. (1999, p. 151), the following procedure guarantees coherence and strong FWE control: First, "Test every member of the closed family by a (suitable) α level test (α is CWE controlled not FWE controlled). Second, a hypothesis can be rejected provided (1) its corresponding test was significant at α , and (2) every other hypothesis in the family that implies it has also been rejected by its corresponding α level test."

Because closed testing procedures were not always easy to derive, various authors derived other simplified stepwise procedures which are computationally simpler, though at the expense of providing smaller α values than what theoretically could be obtained with a closed testing procedure.

Naturally, as a consequence of having smaller α values, these simpler stepwise MCPs would not be as powerful as exact closed testing methods. Nonetheless, these methods are still typically more powerful than STPs (e.g., Tukey) and therefore are recommended and furthermore, researchers can obtain numerical results through the SAS system.

REGWQ. One such method was introduced by Ryan (1960), Einot and Gabriel (1975) and Welsch (1977) and is available through SAS. One can better understand the logic of the REGWQ procedure if we first introduce one of the most popular stepwise strategies for examining pairwise differences between means, the NK procedure.

In this procedure, the means are rank ordered from smallest to largest and the difference between the smallest and largest means is first subjected to a statistical test, typically with a range statistic (Q), at an α level of significance. If this difference is not significant, testing stops and all pairwise differences are regarded as null. If, on the other hand, this first range test is statistically significant, one 'steps-down' to examine the two $J - 1$ subsets of ordered means, that is, the smallest mean versus the next-to-largest mean and the largest mean versus the next-to-smallest mean, with each tested at an α level of significance. At each stage of testing, only subsets of ordered means that are statistically significant are subjected to further testing. Although the NK procedure is very popular among applied researchers, it is becoming increasingly well known that when $J > 3$ it does not limit the FWE to α (See Hochberg & Tamhane, 1987, p. 69).

Ryan (1960) and Welsch (1977), however, have shown how to adjust the subset levels of significance in order to provide strong FWE control. Specifically, in order to strongly control FWE a researcher must:

- Test all subset ($p = 2, \dots, J$) hypotheses at $\alpha_p = 1 - (1 - \alpha)^{\frac{p}{J}}$, for $p = 2, \dots, J - 2$ and at level $\alpha_p = \alpha$ for $p = J - 1, J$.
- Testing starts with an examination of the complete null hypothesis $\mu_1 = \mu_2 = \dots = \mu_J$ and if rejected one steps down to examine subsets of $J - 1$ means, $J - 2$ means, and so on.
- All subset hypotheses implied by a homogeneity hypothesis that has not been rejected are accepted as null without testing.

Westfall et al. (1999, p. 154) indicate “by using REGWQ, strong control is conservatively ensured by testing *directly* all subset homogeneity hypotheses, and *indirectly* all subset intersection hypotheses.” Additionally, a nice feature about using a range procedure when testing homogeneity hypotheses is that when a subset homogeneity hypothesis is rejected one can automatically reject the equality of the smallest and largest means in the set.

REGWQ can be implemented with the SAS GLM program. Moreover, users can use Westfall et al.'s (1999, p. 154) macro (%SimPower) to examine complete, minimal and proportional power. They illustrate, as well, the additional power that researchers can obtain with REGWQ as compared to Tukey's (1953) simultaneous method.

We remind the reader, however, that this procedure can not be used to construct simultaneous confidence intervals.

CLOSED TESTING PROCEDURES THAT INCORPORATE LOGICAL DEPENDENCIES

Westfall et al. (1999) also provide a SAS macro for computing closed testing procedures that incorporate logical dependencies among the hypotheses.

The macro is quite general in that it not only can be used with pairwise comparisons but as well with any collection of linear combinations. The program also correctly incorporates the correlational structure of the tests. The macro is based on the work of Shaffer (1986) and Westfall (1997) and works in the following manner:

- Order the pairwise $|t_{c,s}|$ from most to least significant, that is, $|t_1| \geq |t_2| \geq \dots \geq |t_C|$ (Remember there is a corresponding order for the associated $H_{c,s}$, that is, H_1, \dots, H_C).
- If $|t_1| \geq c_{\alpha,1}$ reject H_1 , where $c_{\alpha,1}$ is the $1 - \alpha$ quantile of the distribution of $\max T_C$. If H_1 is not rejected, fail to reject it as well as H_2, \dots, H_C and stop testing. On the other hand, if H_1 is rejected, continue to the next step.
- If $|t_2| \geq c_{\alpha,2}$ reject H_2 , where $c_{\alpha,2}$ is the maximum of all quantiles of $\max_{S_2} T_C$. Note that the “maximum is taken over all sets of null hypotheses S_2 that (1) contain H_2 and (2) whose joint truth does not contradict falseness of H_1 ” (Westfall et al., 1999, p. 169). If H_2 is not rejected, fail to reject it as well as H_3, \dots, H_C and stop testing. On the other hand, if H_2 is rejected, continue to the next step.
- If $|t_3| \geq c_{\alpha,3}$ reject H_3 , where $c_{\alpha,3}$ is the maximum of all quantiles of $\max_{S_3} T_C$. Note that the “maximum is taken over all sets of null hypotheses S_3 that (1) contain H_3 and (2) whose joint truth does not contradict falseness of either H_1 or H_2 ” (Westfall et al., 1999, p. 169). If H_3 is not rejected, fail to reject it as well as H_4, \dots, H_C and stop testing. On the other hand, if H_3 is rejected, continue to the next step.
- Repeat testing in the fashion described until testing stops, or until all pairwise nulls have been rejected.

Numerical computations for this closed test stepwise MCP can be obtained with macros (%SimTests, %Estimates, %Contrasts) provided by Westfall et al. (1999, pp. 93, 169-171).

MCPs FOR NORMALLY DISTRIBUTED DATA/ HETEROGENEOUS POPULATION VARIANCES

The previously presented procedures assume that the population variances are equal across treatment conditions. Given available knowledge about the nonrobustness of MCPs with conventional test statistics (e.g., t, F), and evidence that population variances are commonly unequal (Keselman et al., 1998; Wilcox, 1988), researchers who persist in applying MCPs with conventional test statistics increase the risk of Type I errors. As Olejnik and Lee (1990) conclude, "most applied researchers are unaware of the problem [of using conventional test statistics with heterogeneous variances] and probably are unaware of the alternative solutions when variances differ" (p. 14).

Although recommendations in the literature have focused on the Games-Howell (1976), or Dunnett (1980) procedures for designs with unequal σ_j^2 s (e.g., see Kirk, 1995; Toothaker, 1991), sequential procedures can provide more power than STPs while generally controlling the FWE (Hsuing & Olejnik, 1994; Kromrey & La Rocca, 1994).

The SAS software can once again be used to obtain numerical results. In particular, Westfall et al. (1999, pp. 206-207) provide SAS programs for logically constrained step-down pairwise tests when heteroscedasticity exists. The macro uses SASs mixed-model program (PROC MIXED) which allows for a non constant error structure across groups. As well, the program adopts the Satterthwaite (1946) solution for error df. Westfall et al. remind the reader that

the solution requires large data sets in order to provide approximately correct FWE control.

It is important to note that other non SAS solutions are possible in the heteroscedastic case. For completeness, we note how these can be obtained. Specifically, sequential procedures based on the usual t_c statistic can be easily modified for unequal σ_j^2 s (and unequal n_j s) by substituting Welch's (1938) statistic, $t_w(\nu_w)$ for $t_c(\nu)$, where

$$t_w = \frac{\bar{Y}_j - \bar{Y}_{j'}}{\sqrt{\frac{s_j^2}{n_j} + \frac{s_{j'}^2}{n_{j'}}}},$$

and s_j^2 and $s_{j'}^2$ represent the sample variances for the j th and j' th group, respectively. This statistic is approximated as a t variate with critical value $t_{(1-\alpha/2), \nu_w}$, the $100(1 - \alpha/2)$ quantile of Student's t distribution with df

$$\nu_w = \frac{\left\{ \frac{s_j^2}{n_j} + \frac{s_{j'}^2}{n_{j'}} \right\}^2}{\frac{(s_j^2/n_j)^2}{n_j - 1} + \frac{(s_{j'}^2/n_{j'})^2}{n_{j'} - 1}}.$$

For procedures simultaneously comparing more than two means or when an omnibus test statistic is required (protected tests), robust alternatives to the usual ANOVA F statistic have been suggested. Possibly the best known robust omnibus test is that due to Welch (1951). With the Welch procedure, the omnibus null hypothesis is rejected if $F_w > F_{(J-1, \nu_w)}$ where:

$$F_W = \frac{\sum_{j=1}^J w_j (\bar{Y}_j - \tilde{Y})^2 / (J-1)}{1 + \frac{2(J-2)}{(J^2-1)} \sum_{j=1}^J \frac{(1-w_j/\sum w_j)^2}{n_j-1}},$$

and $w_j = n_j/s_j^2$, $\tilde{Y} = \sum w_j \bar{Y}_j / \sum w_j$. The statistic is approximately distributed as an F variate and is referred to the critical value, $F_{(1-\alpha, J-1, \nu_W)}$, the $100(1 - \alpha)$ quantile of the F distribution with $J - 1$ and ν_W df, where

$$\nu_W = \frac{J^2-1}{3 \sum_{j=1}^J \frac{(1-w_j/\sum w_j)^2}{n_j-1}}.$$

The Welch test has been found to be robust for largest to smallest variance ratios less than 10:1 (Wilcox, Charlin & Thompson, 1986).

Based on the preceding, one can use the nonpooled Welch test and its accompanying df to obtain various stepwise MCPs. For example, Keselman et al. (1998) verified that one can use this approach with Hochberg's (1988) step-up Bonferroni MCP (see Westfall et al., 1999, pp. 32-33) as well as with Benjamini and Hochberg's (1995) FDR method to conduct all possible pairwise comparisons in the heteroscedastic case.

MCPs FOR NONNORMALLY DISTRIBUTED DATA

An underlying assumption of all of the previously presented MCPs is that the populations from which the data are sampled are normally distributed. Although it may be convenient (both practically and statistically) for researchers to assume that their samples are obtained from normally distributed populations, this assumption may rarely be accurate (Micceri, 1989; Pearson, 1931; Wilcox,

1990) (Tukey (1960) suggests that most populations are skewed and/or contain outliers.). Researchers falsely assuming normally distributed data risk obtaining biased Type I and/or Type II error rates for many patterns of nonnormality, especially when other assumptions are also not satisfied (e.g., variance homogeneity) (See Wilcox, 1997).

Bootstrap and Permutation Tests. The SAS system allows users to obtain both simultaneous and stepwise pairwise comparisons of means with methods that do not presume normally distributed data. In particular, users can use either bootstrap or permutation methods to compute all possible pairwise comparisons.

Bootstrapping allows users to create their own empirical distribution of the data and hence adjusted p-values are accordingly based on the empirically obtained distribution, not a theoretically presumed distribution. For example, the empirical distribution, say \hat{F} , is obtained by sampling, with replacement, the pooled sample residuals $\hat{\epsilon}_{ij} = Y_{ij} - \hat{\mu}_j = Y_{ij} - \bar{Y}_j$. That is, rather than assume that residuals are normally distributed, one uses empirically generated residuals to estimate the true shape of the distribution. From the pooled sample residuals one generates bootstrap data.

First, remember that adjusted p-values are calculated as $\tilde{p}_c = P(\max_c |T_c| \geq |t_c|)$. As Westfall et al. (1999, p. 229) note, in many cases, this is equivalent to $\tilde{p}_c = P(\min_c P_c \leq p_c)$. Their PROC MULTTEST computes adjusted p-values in this fashion. With this in mind, bootstrapping of adjusted p-values with their MULTTEST program is performed in the following manner:

- Bootstrap data, Y_{ij}^* , is generated by sampling with replacement from the pooled sample of residuals.

- Based on the bootstrapped data, p_1^* , p_2^* , ..., p_c^* values are obtained from the pairwise tests.
- The above process is repeated many times (PROC MULTTEST allows the user to set the number of replications.).
- For stepwise testing, PROC MULTTEST uses minima over appropriate restricted subsets to obtain the adjusted p-values.

An example program for all possible pairwise comparisons is given by Westfall et al. (1999, p. 229). As these authors note, PROC MULTTEST does not allow users to incorporate logical constraints as does their %SimTests program (Remember %SimTests assumes data are normally distributed.)

As well, pairwise comparisons of means (or ranks) can be obtained through permutation of the data with the program provided by Westfall et al. (1999, pp. 233-234). Permutation tests also do not require that the data be normally distributed. Instead of resampling with replacement from a pooled sample of residuals, permutation tests take the observed data ($Y_{11}, \dots, Y_{n_1,1}, \dots, Y_{1J}, \dots, Y_{n_J,J}$) and randomly redistributes them to the treatment groups, and summary statistics (i.e., means or ranks) are then computed on the randomly redistributed data. The original outcomes (all possible pairwise differences from the original sample means) are then compared to the randomly generated values (e.g., all possible pairwise differences in the permutation samples). That is, if $\bar{Y}_1^* - \bar{Y}_2^*$ is the difference between the first two treatment group means based on a permutation of the data, then a permutational p-value can be computed as $p = P(\bar{Y}_1^* - \bar{Y}_2^* \geq \bar{Y}_1 - \bar{Y}_2)$. Accordingly, for pairwise comparisons, the adjusted p-values are calculated as $\tilde{p}_c = P(\min_c P_c^* \leq p_c)$, where the P_c^* are computed from the permuted data.

When users adopt this approach to combat the effects of nonnormality they should take heed of the cautionary note provided by Westfall et al. (1999, p. 234), namely, the procedure may not control the FWE when the data are heterogeneous, particularly when group sizes are unequal. Thus, we introduce another approach, pairwise comparisons based on robust estimators and a heteroscedastic statistic, an approach that has been demonstrated to generally control the FWE when data are nonnormal and heterogeneous even when group sizes are unequal.

MCPs FOR NONNORMALLY DISTRIBUTED DATA/ HETEROGENEOUS POPULATION VARIANCES

Trimmed Means Approach

A different type of testing procedure, based on trimmed (or censored) means, has been discussed by Yuen and Dixon (1973) and Wilcox (1995, 1997), and is purportedly robust to violations of normality.

That is, it is well known that the usual group means and variances, which are the basis for all of the previously described procedures, are greatly influenced by the presence of extreme observations in distributions. In particular, the standard error of the usual mean can become seriously inflated when the underlying distribution has heavy tails. Accordingly, adopting a nonrobust measure "can give a distorted view of how the typical individual in one group compares to the typical individual in another, and about accurate probability coverage, controlling the probability of a Type I error, and achieving relatively high power" (Wilcox, 1995, p. 66). By substituting robust measures of location and scale for the usual mean and variance, it should be possible to obtain test

statistics which are insensitive to the combined effects of variance heterogeneity and nonnormality.

While a wide range of robust estimators have been proposed in the literature (see Gross, 1976), the trimmed mean and Winsorized variance are intuitively appealing because of their computational simplicity and good theoretical properties (Wilcox, 1995). The standard error of the trimmed mean is less affected by departures from normality than the usual mean because extreme observations, that is, observations in the tails of a distribution, are censored or removed. Furthermore, as Gross (1976) noted, "the Winsorized variance is a consistent estimator of the variance of the corresponding trimmed mean" (p. 410). In computing the Winsorized variance, the most extreme observations are replaced with less extreme values in the distribution of scores.

Trimmed means are computed by removing a percentage of observations from each of the tails of a distribution (set of observations). Let $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ represent the ordered observations associated with a group. Let $g = [\gamma n]$, where γ represents the proportion of observations that are to be trimmed in each tail of the distribution and $[x]$ is notation for the largest integer not exceeding x . Wilcox (1995) suggests that 20% trimming should be used. The effective sample size becomes $h = n - 2g$. Then the sample trimmed mean is

$$\bar{Y}_t = \frac{1}{h} \sum_{i=g+1}^{n-g} Y_{(i)} .$$

An estimate of the standard error of the trimmed mean is based on the Winsorized mean and Winsorized sum of squares. The sample Winsorized mean is

$$\bar{Y}_w = \frac{1}{n}[(g + 1)Y_{(g+1)} + Y_{(g+2)} + \dots + Y_{(n-g-1)} + (g + 1)Y_{(n-g)}],$$

and the sample Winsorized sum of squared deviations is

$$\begin{aligned} SSD_w &= (g + 1)(Y_{(g+1)} - \bar{Y}_w)^2 + (Y_{(g+2)} - \bar{Y}_w)^2 + \dots + (Y_{(n-g-1)} - \bar{Y}_w)^2 \\ &\quad + (g + 1)(Y_{(n-g)} - \bar{Y}_w)^2. \end{aligned}$$

Accordingly, the squared standard error of the mean is estimated as (Staudte & Sheather, 1990)

$$d = \frac{SSD_w}{h(h-1)}.$$

To test a pairwise comparison null hypothesis compute \bar{Y}_t and d for the j th group, label the results \bar{Y}_{tj} and d_j . The robust pairwise test (see Keselman, Lix & Kowalchuk, 1998) becomes

$$t_w = \frac{\bar{Y}_{tj} - \bar{Y}_{tj'}}{\sqrt{d_j + d_{j'}}},$$

with estimated df

$$\nu_w = \frac{(d_j + d_{j'})^2}{d_j^2/(h_j-1) + d_{j'}^2/(h_{j'}-1)}.$$

When trimmed means are being compared the null hypothesis relates to the equality of population trimmed means, instead of population means. Therefore, instead of testing $H_0: \mu_j = \mu_{j'}$, a researcher would test the null hypothesis, $H_0:$

$\mu_{tj} = \mu_{tjr}$, where μ_t represents the population trimmed mean (Many researchers subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least squares parameters when they are dealing with populations that are nonnormal in form.).

Yuen and Dixon (1973) and Wilcox (1995) report that for long tailed distributions, tests based on trimmed means and Winsorized variances can be much more powerful than tests based on the usual mean and variance. Accordingly, when researchers feel they are dealing with nonnormal data they can replace the usual least squares estimators of central tendency and variability with robust estimators and apply these estimators in any of the previously recommended MCPs.

A MODEL TESTING PROCEDURE

The procedure to be described takes a completely different approach to specifying differences between the treatment group means. That is, unlike previous approaches which rely on a test statistic to reject or accept pairwise null hypotheses, the approach to be described uses an information criterion statistic to select a configuration of population means which most likely corresponds with the observed data. Thus, as Dayton (1998, p. 145) notes, “model-selection techniques are not statistical tests for which type-I error control is an issue.”

When testing all pairwise comparisons, intransitive decisions are extremely common with conventional MCPs (Dayton, 1998). An intransitive decision refers to declaring a population mean (μ_j) not significantly different from two different population means ($\mu_j = \mu_{jr}$, $\mu_j = \mu_{jrr}$), when the latter two means are declared significantly different ($\mu_{jr} \neq \mu_{jrr}$). For example, a researcher conducting all pairwise comparisons ($J = 4$) may decide not to reject any hypotheses implied

by $\mu_1 = \mu_2 = \mu_3$ or $\mu_3 = \mu_4$, but reject $\mu_1 = \mu_4$ and $\mu_2 = \mu_4$, based on results from a conventional MCP. Interpreting the results of this experiment can be ambiguous, especially concerning the outcome for μ_3 .

Dayton (1998) proposed a model testing approach based on Akaike's (1974) Information Criterion (AIC). Mutually exclusive and transitive models are each evaluated using AIC, and the model having the minimum AIC is retained as the most probable population mean configuration, where:

$$\text{AIC} = \text{SS}_w = \sum_j n_j (\bar{Y}_j - \bar{Y}_{mj})^2 + 2q,$$

\bar{Y}_{mj} is the estimated sample mean for the j th group (given the hypothesized population mean configuration for the m th model), SS_w is the ANOVA pooled within group sum of squares and q is the degrees of freedom for the model. For example, for $J = 4$ (with ordered means) there would be $2^{J-1} = 8$ different models to be evaluated ($\{1234\}$, $\{1, 234\}$, $\{12, 34\}$, $\{123, 4\}$, $\{1, 2, 34\}$, $\{12, 3, 4\}$, $\{1, 23, 4\}$, $\{1, 2, 3, 4\}$). To illustrate, the model $\{12, 3, 4\}$ postulates a population mean configuration where groups one and two are derived from the same population, while groups three and four each represent independent populations. The model having the lowest AIC value would be retained as the most probable population model.

Dayton showed that the AIC model-testing approach was more powerful than Tukey's HSD (all-pairs power) for many population mean configurations, and more importantly, eliminates intransitive decisions. One finding reported by Dayton, as well as Huang and Dayton (1995), is that the AIC has a slight bias for selecting more complicated models than the true model. For example, Dayton found that for the mean pattern $\{12, 3, 4\}$, AIC selected the more complicated pattern $\{1, 2, 3, 4\}$ more than ten percent of the time, whereas AIC only rarely

selected less complicated models (e.g., {12, 34}). This tendency can present a special problem for the complete null case {1234}, where AIC has a tendency to select more complicated models. Consequently, a recommendation by Huang and Dayton (1995) is to use an omnibus test to screen for the null case, and then assuming rejection of the null, apply the Dayton procedure.

Dayton's (1998) model testing approach can be modified to handle heterogeneous treatment group variances. Like the original procedure, mutually exclusive and transitive models are each evaluated using AIC, and the model having the minimum AIC is retained as the most probable population mean configuration. For heterogeneous variances:

$$\text{AIC} = -2 \left\{ (-N/2) (\ln(2\pi) + 1) - 1/2 (\sum n_j \ln(S)) \right\} + 2q,$$

where S is the biased variance for the j th group, substituting the estimated group mean (given the hypothesized mean configuration for the m th model) for the actual group mean in the calculation of the variance. As in the original Dayton procedure, an appropriate omnibus test can also be applied.

SUMMARY

Selecting an appropriate MCP requires an extensive assessment of available information regarding the testing situation. Information about the importance of Type I errors, power, computational simplicity, and so on, are extremely important in selecting a MCP. In addition, the selection of a proper MCP is dependent on data conforming to validity assumptions, such as normality

and variance homogeneity. Routinely selecting a procedure without careful consideration of available information and alternatives can severely reduce the reliability and validity of the results.

Recently, several pairwise MCPs have been proposed that improve on one or more aspects of previously recommended MCPs. In particular, stepwise procedures that control the overall rate of Type I error in a strong sense, as well as methods resulting from bootstrapping and permuting the data and methods that substitute robust estimators and heteroscedastic test statistics for the usual estimators and statistics are now available. In addition to defining these 'newer' methods, we as well indicated statistical software that can be used to obtain numerical results. Indeed, our guiding principle for selecting which procedures to review was based on our belief that only procedures which can be obtained through a statistical package are likely to be adopted by researchers. Accordingly, we emphasized many of the procedures which are available through the SAS system because Westfall et al. (1999) have provided many useful programs for obtaining numerical solutions. In conclusion, by writing this chapter we hope our summary encourages researchers to switch from older methods for assessing pairwise multiple comparisons to the newer approaches we have reviewed.

NUMERICAL EXAMPLE

We present a numerical example for the previously discussed MCPs so that the reader can check his/her facility to work with the SAS/Westfall et al. (1999) programs and to demonstrate through example the differences between their operating characteristics. In particular, the data ($n_1 = n_2 = \dots = n_J = 20$) presented in Table 1 was randomly generated by us though they could represent

the outcomes of a problem solving task where the five groups were given different clues to solve the problem; the dependent measure was the time, in seconds, that it took to solve the task. The bottom two rows of the table contain the group means and standard deviations, respectively.

Table 2 contains adjusted p-values and FWE ($\alpha = .05$) significant (*) values for the 10 pairwise comparisons for the five groups. The results reported in Table 2 conform, not surprisingly, to the properties of the MCPs that we discussed previously. In particular, of the ten comparisons, five were found to be statistically significant with the Tukey (1953), Hayter (1986), Bootstrap, Stepdown Bootstrap and permutation procedures: $\bar{Y}_1 - \bar{Y}_3$, $\bar{Y}_1 - \bar{Y}_4$, $\bar{Y}_1 - \bar{Y}_5$, $\bar{Y}_2 - \bar{Y}_5$, and $\bar{Y}_3 - \bar{Y}_5$. In addition to these five comparisons, the REGWQ MCP found one additional comparison ($\bar{Y}_4 - \bar{Y}_5$) to be statistically significant. On the other hand, the logically constrained approach only detected three significant pairwise comparisons, namely $\bar{Y}_1 - \bar{Y}_4$, $\bar{Y}_1 - \bar{Y}_5$, and $\bar{Y}_2 - \bar{Y}_5$. That is, as Westfall et al. (1999, p. 171) note, the logically constrained tests need not be more powerful than REGWQ. The BH MCP, on the other hand, resulted in seven statistically significant comparisons; $\bar{Y}_1 - \bar{Y}_3$, $\bar{Y}_1 - \bar{Y}_4$, $\bar{Y}_1 - \bar{Y}_5$, $\bar{Y}_2 - \bar{Y}_4$, $\bar{Y}_2 - \bar{Y}_5$, $\bar{Y}_3 - \bar{Y}_5$ and $\bar{Y}_4 - \bar{Y}_5$. The adaptive BH procedure (BH-A) resulted in one additionally significant comparison- $\bar{Y}_2 - \bar{Y}_3$ (Numerical results were not obtained through SAS; they were obtained through hand-calculations.). Clearly the procedures based on the more liberal FDR found more comparisons to be statistical significant than the FWE controlling MCPs.

We also investigated the ten pairwise comparisons with the trimmed means and model testing procedures; the results for the trimmed means analysis are also reported in Table 2 (see Appendix B for the SPSS program). Numerical results for robust estimation and testing were obtained through SPSS (Norusis, 1997). In particular, we computed the group trimmed means ($\bar{Y}_{t1} = 14.92$,

$\bar{Y}_{t2} = 16.67$, $\bar{Y}_{t3} = 18.50$, $\bar{Y}_{t4} = 19.08$ and $\bar{Y}_{t5} = 21.08$) as well as the group Winsorized standard deviations ($d_1^{\frac{1}{2}} = 2.11$, $d_2^{\frac{1}{2}} = 1.68$, $d_3^{\frac{1}{2}} = 3.75$, $d_4^{\frac{1}{2}} = 2.33$ and $d_5^{\frac{1}{2}} = 4.24$). These values (as well as the effective sample sizes- $n_1 = \dots = n_5 = 12$) were then read into an SPSS file and we then allowed SPSS to calculate nonpooled t-statistics (t_W and ν_W) and their corresponding p-values (through the ONEWAY program). The results reported in Table 2 indicate that with this approach six comparisons were found to be statistically significant: $\bar{Y}_1 - \bar{Y}_2$, $\bar{Y}_1 - \bar{Y}_3$, $\bar{Y}_1 - \bar{Y}_4$, $\bar{Y}_1 - \bar{Y}_5$, $\bar{Y}_2 - \bar{Y}_4$ and $\bar{Y}_2 - \bar{Y}_5$. Clearly, other MCPs had greater power to detect more pairwise differences. However, the reader should remember that robust estimation should result in more powerful tests when data are nonnormal as well as heterogeneous (see Wilcox, 1997), which was not the case with our numerical example data. Furthermore, trimmed results were based on 12 subjects per group, not 20.

With regard to the model testing approach we examined the $2^J - 1$ models of nonoverlapping subsets of ordered means and used the minimum AIC value to find the best model that "is expected to result in the smallest loss of precision relative to the true, but unknown, model (Dayton, 1998, p. 145)." From the 16 models examined, the best model is {1,2,3,4,5} (Results were obtained through hand calculations. However, a GAUSS program will soon be available from the Department of Measurement & Statistics, University of Maryland web-site). Accordingly, this approach provides identical findings to the BH-A procedure, namely, $\bar{Y}_1 - \bar{Y}_3$, $\bar{Y}_1 - \bar{Y}_4$, $\bar{Y}_1 - \bar{Y}_5$, $\bar{Y}_2 - \bar{Y}_3$, $\bar{Y}_2 - \bar{Y}_4$, $\bar{Y}_2 - \bar{Y}_5$, $\bar{Y}_3 - \bar{Y}_5$ and $\bar{Y}_4 - \bar{Y}_5$ would be judged to be different from one another.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, AC-19, 716-723.
- Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B, 57, 289-300.
- Benjamini, Y. & Hochberg, Y. (in press). On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics.
- Benjamini, Y., Hochberg, Y. & Kling, Y. (1994). False discovery rate controlling procedures for pairwise comparisons. Unpublished manuscript.
- Bofinger, E., Hayter, A. J., & Liu, W. (1993). The construction of upper confidence bounds on the range of several location parameters. Journal of the American Statistical Association, 88, 906-911.
- Dayton, C. M. (1998). Information criteria for the paired-comparisons problem. The American Statistician, 52, 144-151.
- Duncan, D. B. (1955). Multiple range and multiple F tests. Biometrics, 11, 1-42.
- Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. Journal of the American Statistical Association, 75, 796-800.
- Einot, I. & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. Journal of the American Statistical Association, 70, 574-583.
- Fisher, R. A. (1935). The design of experiments. Edinburgh: Oliver & Boyd.
- Games, P. A. (1971). Multiple comparisons of means. American Educational Research Journal, 8, 531-565.
- Games, P. A. & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances. Journal of Educational Statistics, 1,

113-125.

- Gross, A. M. (1976). Confidence interval robustness with long-tailed symmetric distributions. Journal of the American Statistical Association, 71, 409-416.
- Hancock, G. R. & Klockars, A. J. (1996). The quest for α : Developments in multiple comparison procedures in the quarter century since Games (1971). Review of Educational Research, 66, 269-306.
- Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. Journal of the American Statistical Association, 81, 1000-1004.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75, 800-802.
- Hochberg, Y. & Tamhane, A. C. (1987). Multiple comparison procedures. New York: John Wiley.
- Holland, B. & Cheung, S. H. (2000). Family size robustness criteria for multiple comparison procedures. Manuscript submitted for publication.
- Hsuing, T. & Olejnik, S. (1994). Power of pairwise multiple comparisons in the unequal variance case. Communications in Statistics: Simulation and Computation, 23, 691-710.
- Huang, C. J. & Dayton C. M. (1995). Detecting patterns of bivariate mean vectors using model-selection criteria. British Journal of Mathematical and Statistical Psychology, 48, 129-147.
- Keselman, H. J., Cribbie, R. A. & 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.
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C. &

- Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. Review of Educational Research, 68, 350-386.
- Keselman, H. J., Keselman, J. C. & Games, P. A. (1991). Maximum familywise Type I error rate: The least significant difference, Newman-Keuls, and other multiple comparison procedures. Psychological Bulletin, 110, 155-161.
- Keselman, H. J., Lix, L. M. & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. Psychological Methods, 3, 123-141.
- Keuls, M. (1952). The use of the "Studentized range" in connection with an analysis of variance. Euphytica, 1, 112-122.
- Kirk, R. E. (1995). Experimental design: Procedures for the behavioral sciences. Toronto: Brooks/Cole Publishing Company.
- Kromrey, J. D. & La Rocca, M. A. (1994). Power and Type I error rates of new pairwise multiple comparison procedures under heterogeneous variances. Journal of Experimental Education, 63, 343-362.
- Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. Psychological Bulletin, 115, 153-159.
- Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika, 63, 655-660.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.
- Miller, R. G. (1981). Simultaneous statistical inference (2nd Ed.). New York: Springer-Verlag.
- Newman, D. (1939). The distribution of the range in samples from a normal

- population, expressed in terms of an independent estimate of standard deviation. Biometrika, 31, 20-30.
- Norusis, M. J. (1997). SPSS 9.0 Guide to data analysis. New Jersey: Prentice Hall.
- Olejnik, S. & Lee, J. (1990, April). Multiple comparison procedures when population variances differ. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.
- Pearson, E. S. (1931). The analysis of variance in cases of nonnormal variation. Biometrika, 23, 114-133.
- Petrinovich, L. F. & Hardyck, C. D. (1969). Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. Psychological Bulletin, 71, 43-54.
- Rothman, K. (1990). No adjustments are needed for multiple comparisons. Epidemiology, 1, 43-46.
- Ryan, T. A. (1959). Multiple comparisons in psychological research. Psychological Bulletin, 56, 26-47.
- Ryan, T. A. (1960). Significance tests for multiple comparison of proportions, variances, and other statistics. Psychological Bulletin, 57, 318-328.
- Ryan, T. A. (1962). The experiment as the unit for computing rates of error. Psychological Bulletin, 59, 305.
- Ryan, T. A. (1980). Comment on "Protecting the overall rate of Type I errors for pairwise comparisons with an omnibus test statistic". Psychological Bulletin, 88, 354-355.
- SAS Institute Inc. (1999). SAS/STAT user's guide, Version 7. Cary, NC: SAS Institute Inc.
- Saville, D. J. (1990). Multiple comparison procedures: The practical solution. The American Statistician, 44, 174-180.

- Scheffé, H. (1959). The analysis of variance. Wiley, 1959.
- Seaman, M. A., Levin, J. R. & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. Psychological Bulletin, 110, 577-586.
- Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.
- Staudte, R. G., & Sheater, S. J. (1990). Robust estimation and testing. New York: Wiley.
- Toothaker, L. E. (1991). Multiple comparisons for researchers. Newbury Park, CA: Sage Publications Inc.
- Tukey, J. W. (1953). The problem of multiple comparisons. Unpublished manuscript, Princeton University, Department of Statistics.
- Tukey, J.W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds) Contributions to probability and Statistics, Stanford, CA: Stanford University Press.
- Welch, B. L. (1938). The significance of the difference between two means when population variances are unequal. Biometrika, 38, 330-336.
- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. Biometrika, 38, 330-336.
- Welsch, R. E. (1977). Stepwise multiple comparison procedures. Journal of the American Statistical Association, 72, 566-575.
- Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. Journal of the American Statistical Association, 92, 299-306.
- Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (1999). Multiple comparisons and multiple tests. Cary, NC: SAS Institute, Inc.
- Westfall, P. H., & Wolfinger, R. D. (1997). Multiple tests with discrete

- distributions. The American Statistician, 51, 3-8.
- Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment. New York: Wiley.
- Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James' second order method. British Journal of Mathematical and Statistical Psychology, 41, 109-117.
- Wilcox, R. R. (1990). Comparing the means of two independent groups. Biometrics Journal, 32, 771-780.
- Wilcox, R. R. (1995). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. British Journal of Mathematical and Statistical Psychology, 48, 99-114.
- Wilcox, R. R. (1997). Three multiple comparison procedures for trimmed means. Biometrical Journal, 37, 643-656.
- Wilcox, R. R., Charlin, V. L. & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F* statistics. Communications in Statistics: Simulation and Computation, 15, 933-943.
- Wilson, W. (1962). A note on the inconsistency inherent in the necessity to perform multiple comparisons. Psychological Bulletin, 59, 296-300.
- Yuen, K. K. & Dixon, W. J. (1973). The approximate behavior of the two-sample trimmed t. Biometrika, 60, 369-374.

Appendix A

Below we include the SAS (1999) syntax for our hypothetical data set.

*****DATA INPUT***;**

```
DATA;
INPUT IV DV;
CARDS;
  1.00 17
  1.00 15
  :
  5.00 21
  5.00 18
;
```

***** REGWQ MULTIPLE COMPARISON PROCEDURE***;**

```
PROC GLM;
CLASS IV;
MODEL DV=IV;
CONTRAST '1V2' IV 1 -1 0 0 0;
CONTRAST '1V3' IV 1 0 -1 0 0;
CONTRAST '1V4' IV 1 0 0 -1 0;
CONTRAST '1V5' IV 1 0 0 0 -1;
CONTRAST '2V3' IV 0 1 -1 0 0;
CONTRAST '2V4' IV 0 1 0 -1 0;
CONTRAST '2V5' IV 0 1 0 0 -1;
CONTRAST '3V4' IV 0 0 1 -1 0;
CONTRAST '3V5' IV 0 0 1 0 -1;
CONTRAST '4V5' IV 0 0 0 1 -1;
MEANS IV/REGWQ;
RUN;
```

***** TUKEY'S HSD PROCEDURE WITH ADJUSTED P-VALUES ***;**

```
PROC GLM;
CLASS IV;
MODEL DV=IV;
CONTRAST '1V2' IV 1 -1 0 0 0;
CONTRAST '1V3' IV 1 0 -1 0 0;
CONTRAST '1V4' IV 1 0 0 -1 0;
CONTRAST '1V5' IV 1 0 0 0 -1;
CONTRAST '2V3' IV 0 1 -1 0 0;
```

```
CONTRAST '2V4' IV 0 1 0 -1 0;  
CONTRAST '2V5' IV 0 1 0 0 -1;  
CONTRAST '3V4' IV 0 0 1 -1 0;  
CONTRAST '3V5' IV 0 0 1 0 -1;  
CONTRAST '4V5' IV 0 0 0 1 -1;  
LSMEANS IV/PDIFF ADJUST=TUKEY;  
RUN;
```

***** BOOTSTRAP MULTIPLE COMPARISON PROCEDURE***;**

```
PROC MULTTEST BOOTSTRAP SEED=121211 N=50000;  
CLASS IV;  
TEST MEAN(DV);  
CONTRAST '1V2' 1 -1 0 0 0;  
CONTRAST '1V3' 1 0 -1 0 0;  
CONTRAST '1V4' 1 0 0 -1 0;  
CONTRAST '1V5' 1 0 0 0 -1;  
CONTRAST '2V3' 0 1 -1 0 0;  
CONTRAST '2V4' 0 1 0 -1 0;  
CONTRAST '2V5' 0 1 0 0 -1;  
CONTRAST '3V4' 0 0 1 -1 0;  
CONTRAST '3V5' 0 0 1 0 -1;  
CONTRAST '4V5' 0 0 0 1 -1;  
ODS SELECT CONTINUOUS PVALUES;  
RUN;
```

***** STEP-DOWN BOOTSTRAP MULTIPLE COMPARISON PROCEDURE***;**

```
PROC MULTTEST STEPBOOT SEED=121211 N=50000;  
CLASS IV;  
TEST MEAN(DV);  
CONTRAST '1V2' 1 -1 0 0 0;  
CONTRAST '1V3' 1 0 -1 0 0;  
CONTRAST '1V4' 1 0 0 -1 0;  
CONTRAST '1V5' 1 0 0 0 -1;  
CONTRAST '2V3' 0 1 -1 0 0;  
CONTRAST '2V4' 0 1 0 -1 0;  
CONTRAST '2V5' 0 1 0 0 -1;  
CONTRAST '3V4' 0 0 1 -1 0;  
CONTRAST '3V5' 0 0 1 0 -1;  
CONTRAST '4V5' 0 0 0 1 -1;  
ODS SELECT CONTINUOUS PVALUES;  
RUN;
```

***** PERMUTATION RESAMPLING MULTIPLE COMPARISON
PROCEDURE***;**

```
PROC MULTTEST PERMUTATION SEED=121211 N=50000;
  CLASS IV;
  TEST MEAN(DV);
  CONTRAST '1V2' 1 -1 0 0 0;
  CONTRAST '1V3' 1 0 -1 0 0;
  CONTRAST '1V4' 1 0 0 -1 0;
  CONTRAST '1V5' 1 0 0 0 -1;
  CONTRAST '2V3' 0 1 -1 0 0;
  CONTRAST '2V4' 0 1 0 -1 0;
  CONTRAST '2V5' 0 1 0 0 -1;
  CONTRAST '3V4' 0 0 1 -1 0;
  CONTRAST '3V5' 0 0 1 0 -1;
  CONTRAST '4V5' 0 0 0 1 -1;
  ODS SELECT PVALUES;
RUN;
```

***** BH MULTIPLE COMPARISON PROCEDURE WITH ADJUSTED P-
VALUES***;**

```
DATA ONE;
INPUT TEST PVAL;
DATALINES;
1 .1406
2 .0007
3 .0002
4 .00005
5 .0469
6 .0185
7 .00005
8 .7002
9 .0065
10 .0185
;
```

```
DATA TWO;
SET ONE;
RENAME PVAL=RAW_P;
PROC MULTTEST PDATA=ONE FDR OUT=OUTP;
PROC SORT DATA=OUTP OUT=OUTP;
BY RAW_P;
PROC PRINT DATA=OUTP;
RUN;
```

***** LOGICALLY CONSTRAINED MULTIPLE COMPARISON PROCEDURE***;**

%MACRO CONTRASTS;

**C = {1 -1 0 0 0, 1 0 -1 0 0, 1 0 0 -1 0, 1 0 0 0 -1, 0 1 -1 0 0,
0 1 0 -1 0, 0 1 0 0 -1, 0 0 1 -1 0, 0 0 1 0 -1, 0 0 0 1 -1};**

C = C`;

CLAB={"1-2", "1-3", "1-4", "1-5", "2-3", "2-4", "2-5", "3-4", "3-5", "4-5"};

%MEND;

%MACRO ESTIMATES;

MSE=10.87;

DF=95;

ESTPAR={15.0, 16.55, 18.65, 19.05, 21.55};

COV=MSE*(1/20)*I(5);

%MEND;

%SIM TESTS(SEED=121223, TYPE=LOGICAL);

Appendix B
SPSS/SAS Program for Trimmed Means Results

Note: To obtain results we used the following procedure. We created a data set with “raw” data that had the desired properties. That is, we made all the cases in a given group equal to the trimmed mean, except for two that were above (in our data set the second-to-last case) and below the mean (in our data set the last case) by $\text{SQRT}[(N - 1)/2] \cdot \text{SD}$; thus we had a group with n (12) cases with a given mean and SD (Square root of Winsorized variance). If you have large sample sizes, you can shortcut this method by using a weighting variable to use just one case with the mean value, with a weight of $n - 2$. The deviation cases would have weights of 1 (This method was suggested by David Nichols, Principal Support Statistician and Manager of Statistical Support, SPSS Inc.).

After computing the trimmed mean and Winsorized variance for each group, the following data was read into an SPSS data file (via their drop down menu system) and then was used to compute Welch (1938) nonpooled test statistics and p-values

IV	DV
1.00	14.917
1.00	14.917
1.00	14.917
1.00	14.917
1.00	14.917
1.00	14.917
1.00	14.917
1.00	14.917
1.00	14.917
1.00	14.917
1.00	14.917
1.00	18.680
1.00	11.153
2.00	16.667
2.00	16.667
2.00	16.667
2.00	16.667
2.00	16.667
2.00	16.667
2.00	16.667
2.00	16.667
2.00	16.667
2.00	16.667
2.00	16.667
2.00	19.653
2.00	13.681

3.00 18.5
3.00 18.5
3.00 18.5
3.00 18.5
3.00 18.5
3.00 18.5
3.00 18.5
3.00 18.5
3.00 18.5
3.00 25.198
3.00 11.802
4.00 19.083
4.00 19.083
4.00 19.083
4.00 19.083
4.00 19.083
4.00 19.083
4.00 19.083
4.00 19.083
4.00 19.083
4.00 19.083
4.00 19.083
4.00 19.083
4.00 19.083
4.00 23.143
4.00 15.023
5.00 21.083
5.00 21.083
5.00 21.083
5.00 21.083
5.00 21.083
5.00 21.083
5.00 21.083
5.00 21.083
5.00 21.083
5.00 21.083
5.00 21.083
5.00 21.083
5.00 28.595
5.00 13.571

ONEWAY DV BY IV

/CONTRAST 1 -1 0 0 0
/CONTRAST 1 0 -1 0 0
/CONTRAST 1 0 0 -1 0
/CONTRAST 1 0 0 0 -1
/CONTRAST 0 1 -1 0 0
/CONTRAST 0 1 0 -1 0
/CONTRAST 0 1 0 0 -1

```
/CONTRAST 0 0 1 -1 0  
/CONTRAST 0 0 1 0 -1  
/CONTRAST 0 0 0 1 -1.
```

The following SAS program was then used to compute adjusted BH p-values from the raw p-values generated above

```
DATA THREE;  
INPUT TEST PVAL;  
DATALINES;  
1 .0080  
2 .0010  
3 .0004  
4 .0004  
5 .0600  
6 .0010  
7 .0010  
8 .5530  
9 .0490  
10 .0740  
;  
DATA FOUR;  
SET THREE;  
RENAME PVAL=RAW_P;  
PROC MULTTEST PDATA=ONE FDR OUT=OUTP;  
PROC SORT DATA=OUTP OUT=OUTP;  
BY RAW_P;  
PROC PRINT DATA=OUTP;  
RUN;
```


Table 1. Data values and summary statistics (Means and Standard Deviations)

J_1	J_2	J_3	J_4	J_5
17	17	17	20	20
15	14	14	15	23
17	15	19	18	18
13	12	15	20	26
22	18	15	14	17
14	18	19	18	14
12	16	20	18	16
15	18	18	16	32
14	16	22	21	21
16	20	16	23	23
14	15	13	15	29
15	19	16	22	21
17	16	23	18	26
11	16	22	19	22
14	19	24	19	22
13	13	20	23	17
17	18	18	23	27
15	16	25	18	18
12	19	13	18	21
17	16	24	23	18
15.00	16.55	18.65	19.05	21.55
2.47	2.11	3.79	2.82	4.64

Note: The first row following the double lines contains the means while the second row following the double lines contains the group standard deviations.

Table 2. Adjusted p-values and FWE significant (*) Comparisons

$Y_i - Y_j$	Tukey	Hayter	REGWQ	LC	Boot	Stepb	Perm	BH	BH-A	TM
1 vs 2	.5740			.5041	.5714	.2567	.5690	.1562		.0133
1 vs 3	.0063	*	*	.0535	.0063	.0048	.0065	.0018	*	.0020
1 vs 4	.0017	*	*	.0356	.0018	.0016	.0018	.0007	*	.0020
1 vs 5	<.0001	*	*	.0002	<.0001	<.0001	<.0001	.0003	*	.0020
2 vs 3	.2676			.2111	.2658	.1209	.2695	.0586	*	.0750
2 vs 4	.1251			.2111	.1238	.0744	.1240	.0264	*	.0020
2 vs 5	<.0001	*	*	.0056	.0001	.0001	.0001	.0003	*	.0020
3 vs 4	.9953			.7868	.9951	.7012	.9959	.7002		.5530
3 vs 5	.0500	*	*	.1693	.0494	.0332	.0490	.0130	*	.0700
4 vs 5	.1251		*	.1786	.1238	.0744	.1240	.0264	*	.0822

Note: Tukey-Tukey (1953); Hayter-Hayter (1986); REGWQ-Ryan (1960)-Einot & Gabriel (1975)-Welsch (1977); LC-Logically constrained tests (Westfall et al., 1999, pp. 168-170); Boot (Bootstrap)/Stepb(Stepwise bootstrap)/Perm (Permutation)-Westfall et al.; BH-Benjamini & Hochberg (1995); BH-A(Adaptive)-Benjamini & Hochberg (in press); TM-Trimmed means (and Winsorized variances) used with a nonpooled t-test and BH critical constants. Raw p-values (1 vs 2, ..., 4 vs 5) for the SAS (1999) procedures are .1406, .0007, .0002, <.0001, .0469, .0185, <.0001, .7022, .0065 and .0185. The corresponding values for the trimmed means tests are .008, .001, .000, .000, .060, .001, .001, .553, .049 and .074.

