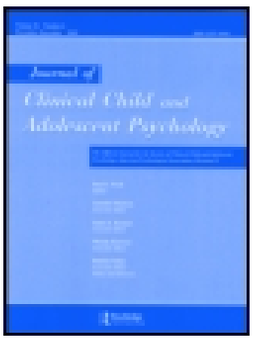

Routledge
Taylor & Francis Group

Journal of Clinical Child & Adolescent Psychology

ISSN: 1537-4416 (Print) 1537-4424 (Online) Journal homepage: http://www.tandfonline.com/loi/hcap20

Pairwise Multiple Comparison Test Procedures: An Update for Clinical Child and Adolescent Psychologists

H. J. Keselman , Robert A. Cribbie & Burt Holland

To cite this article: H. J. Keselman , Robert A. Cribbie & Burt Holland (2004) Pairwise Multiple Comparison Test Procedures: An Update for Clinical Child and Adolescent Psychologists, Journal of Clinical Child & Adolescent Psychology, 33:3, 623-645, DOI: 10.1207/s15374424jccp3303_19

To link to this article: http://dx.doi.org/10.1207/s15374424jccp3303_19

Journal of Clinical Child and Adolescent Psychology
2004, Vol. 33, No. 3, 623–645

METHODOLOGICAL ARTICLE

Pairwise Multiple Comparison Test Procedures: An Update for Clinical Child and Adolescent Psychologists

H. J. Keselman
University of Manitoba

Robert A. Cribbie
York University

Burt Holland
Temple University

Locating pairwise differences among treatment groups is a common practice of applied researchers. Articles published in this journal have addressed the issue of statistical inference within the context of an analysis of variance (ANOVA) framework, describing procedures for comparing means, among other issues. In particular, 1 article (Jaccard & Guilamo-Ramos, 2002b) presented some new methods of performing contrasts of means whereas another presented a framework for obtaining robust tests within this same context (Jaccard & Guilamo-Ramos, 2002a). The purpose of this article is to add to these contributions by presenting some newer methods for conducting pairwise comparisons of means, that is by extending the contributions of the first article and applying the framework of the second article to pairwise multiple comparisons. The newer methods are intended to provide additional sensitivity to detect treatment group differences and provide tests that are robust to the effects of variance heterogeneity, nonnormality, or both.

As noted by Jaccard and Guilamo-Ramos (2002a, 2002b) and Wilcox (2002), researchers conducting studies related to children and adolescents in clinical psychology are often interested in comparing the means of several treatment conditions $(\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_J; j = 1, \ldots, J)$ on a specific dependent measure. When each treatment group mean is compared with every other group mean, the tests are designated *pairwise comparisons.* Pairwise comparisons among treatment group means provide focused questions in studies such as those that compare the (a) social functioning and number of childhood behavior problems of attention deficit hyperactivity disorder–high inattentiveness, attention deficit hyperactivity disorder–low inattentiveness, and nondiagnosed control children (Carlson & Mann, 2002); (b) depression, anxiety, and a number of cognitive distortions of internalizing only, externalizing only, internalizing and externalizing, and control (no internalizing or externalizing) children (Epkins, 2000); (c) parental distress and parental coping of parents of children with attention deficit hyperactivity disorder–inattentive, attention deficit hyperactivity disorder–combined (inattentive and hyperactive), and nondiagnosed control children (Podolski & Nigg, 2001); (d) teacher-rated amount of aggressive–disruptive behavior of children labeled low cognitive ability-elevated inattention, elevated inattention only, low cognitive ability only, or control (no problems) (Bellanti & Bierman, 2000); and (e) parenting stress, depression, and satisfaction and perceptions of child behavior of custodial grandparents seeking outpatient psychological services for their 3- to 12-year-old grandchildren; custodial grandparents not seeking any psychological services for their grandchildren; maternal parents seeking outpatient psychological services for their children (Daly & Glenwick, 2000), to illustrate but a few exemplars.

Work on this article was supported by grants from the Natural Sciences and Engineering Research Council and the Social Sciences and Humanities Research Council of Canada. We would like to thank James Jaccard for his many helpful comments on earlier drafts. The first author would also like to thank the Mathematics Department at Monmouth University (West Long Branch, NJ) for providing a home base during a sabbatical leave.

Requests for reprints should be sent to H. J. Keselman, Department of Psychology, 190 Dysart Road, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2. E-mail: hj_keselman@umanitoba.ca

When computing all pairwise comparisons, the researcher must consider various issues (these issues pertain to other classes of multiple tests as well), including (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. 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 multiple comparison procedures that are available, the Tukey and Scheffé (1959) methods are most preferred (Keselman, Huberty, et al., 1998).

With respect to the selection of a multiple comparison procedure, the researcher must be aware that his or her choice can often significantly affect the results of the experiment. For example, many multiple comparison procedures (e.g., those that are based on traditional test statistics) are inappropriate (and may lead to incorrect decisions) when assumptions of the test statistics are not met (e.g., normality, variance homogeneity). Furthermore, several multiple comparison procedures have recently been proposed that, according to published results 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 (see Wilcox, 2003).

Therefore, the goal of this article is to describe some of the newer multiple comparison procedures 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, the goal is to help popularize newer procedures, procedures that should provide researchers with more robust and more powerful tests of their pairwise comparison null hypotheses. We also discuss the generalization of procedures to factorial designs.

It is also important to note that the multiple comparison procedures that are presented in this article were selected for discussion, by and large, because researchers can, in most cases, obtain numerical results with a statistical package, in particular, through the SAS (1999) system of computer programs. The SAS system (see Westfall, Tobias, Rom, Wolfinger, & Hochberg, 1999) presents a comprehensive, up-to-date array of multiple comparison procedures. It should be noted that for many of the procedures we discuss, as well as many that we do not discuss, numerical solutions can be obtained from SPSS (see Norusis, 2002); we indicate when this is the case. Accordingly, we acknowledge at the beginning of our presentation that some of the material we present follows closely Westfall et al.'s presentation. This article, however, focuses on multiple comparison procedures 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., multiple comparison procedures 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 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 more difficult when multiple tests of significance (e.g., all possible pairwise comparisons between group means) are computed (see Jaccard & Guilamo-Ramos, 2002a, 2002b). 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. 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 either support or refute one's research hypotheses. Selecting the statistically significant findings from a larger pool of results that also contain nonsignificant findings is problematic because 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; Ryan, 1960, 1962; Westfall & Young, 1993). That is, as promul-

gated by 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) multiple comparison procedure 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 article, when we refer to 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). 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 5% 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 may be interested in determining if there are significant pairwise mean differences among 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 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 comparisons can be much lower than α. Clearly, the conclusions of an experiment can be greatly affected by the level of significance and the family of inferences over which Type I error control is imposed.

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, and so on). Specification of a family of comparisons, 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 article family size is all possible pairwise comparisons.) As Westfall et al. (1999) noted, 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. recommended 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). Accordingly, we believe that unless the family of interest is clear-cut and obvious, the analyst is obliged to justify the choice of family size. Definitions of the CWE and FWE rates appear in many sources (e.g., Kirk, 1995; Toothaker, 1991; Tukey, 1953; see Appendix A for error rate definitions).

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

In the past, controlling the FWE rate has been recommended by many researchers (e.g., Hancock & Klockars, 1996; Ryan, 1962; Tukey, 1953). Indeed, ac-

cording to Seaman, Levin, and Serlin (1991) it is "the most commonly endorsed approach to accomplishing Type I error control" (p. 577). Not surprisingly, therefore, Keselman, Huberty, et al. (1998) reported that approximately 85% of researchers conducting pairwise comparisons adopt some form of FWE control. However, it would not surprise us that, particularly when family size is large, the false discovery rate might presently be the preferred method of control (see below).

Although many multiple comparison procedures purport to control FWE, some provide "strong" FWE control whereas 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 ($\mu_1 = \mu_2 = \cdots = \mu_J$) and all possible partial null configurations. (An example of a partial null hypothesis is ($\mu_1 = \mu_2 = \cdots = \mu_{J-1} \neq \mu_J$). 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) noted, 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. Examples of multiple comparison procedures that only weakly control FWE are the Newman (1939)–Keuls (1952) and Duncan (1955) procedures.

False Discovery Rate Control

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, 1960; Toothaker, 1991; Tukey, 1953), although other error rates, such as the false discovery rate, also have been proposed (e.g., Benjamini & Hochberg, 1995). 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 false discovery rate, presented by Benjamini and Hochberg. The false discovery rate is defined by these authors as the expected proportion of the number of erroneous rejections to the total number of rejections (see Appendix A for further details).

Benjamini and Hochberg (1995) provided a number of illustrations in which false discovery rate control seems more reasonable than familywise or comparisonwise control. Exploratory research, for example, would be one area of application for false discovery rate control. That is, in new areas of inquiry in which one is 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 false discovery rate method of control that 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 conceptual formulations does one need more stringent familywise control. Another area in which false discovery rate control might be preferred over familywise control, suggested by Benjamini and Hochberg, would be when two treatments (say, treatments for dyslexia) are being compared in multiple subgroups (say, children of different ages). In studies of this sort, in which an overall decision regarding the efficacy of the treatment is not of interest but 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 the false discovery rate rather than familywise control.

Recently, use of the false discovery rate criterion has become widespread when making inferences in research involving the human genome, where family sizes in the thousands are common. See the review by Dudoit, Shaffer, and Boldrick (2003) and the references contained therein. Another area of research in psychology where false discovery rate controlling procedures have had a significant impact is functional magnetic resonance imaging. In these experiments, researchers conduct numerous (often more than 100,000) significance tests that relate to tests of activation on specific voxels (i.e., areas) within the brain (e.g., Callan et al., 2003).

Because multiple testing with the false discovery rate tends to detect more significant differences than testing with FWE, some researchers may be tempted to automatically prefer false discovery rate control to FWE control. We caution that researchers who use the false discovery rate should be obligated to explain, in terms of the definitions of the two criteria, why it is more appropriate to control the false discovery rate than FWE in the context of their research.

Adjusted *p* Values

As indicated, FWE control is the rate of error control that is currently favored by researchers in most social science contexts. 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 $\tilde{p}_c, c = 1,\ldots,C$ (Westfall et al., 1999; Westfall & Young, 1993). As Westfall and Young noted, "$\tilde{p}_c$ is the smallest significance level for which one still rejects a given hypothe-

sis (H_c) 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 chosen by the experimenter. The advantage of adjusted *p* values for multiple comparison procedures, as with *p* values for tests in comparisonwise contexts, is that they are more informative than merely declaring retain or reject H_0; they are a measure of the weight of evidence for or against the null hypothesis when controlling FWE—for example, if the researcher or reader can conclude that the test is statistically significant at the FWE = .10 level but not at the FWE = .05 level (see Appendix A for an illustration of computing adjusted *p* values). Adjusted *p* values are provided by the SAS Institute (1999) system for many popular multiple comparison procedures (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, is based on the work of Westfall et al. (1999).

According to Westfall et al. (1999), when multiple tests of significance are (to be) examined, power can be defined from four different perspectives: (a) complete power, (b) minimal power, (c) individual power, and (d) proportional power. The definitions they provide are

- Complete Power—*P* (reject all H_cs that are false)
- Minimal Power—*P* (reject at least one H_c that is false)
- Individual Power—*P* (reject a particular H_c that is false)
- Proportional Power (average proportion of false H_cs 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 well-executed research designs. For example, as Westfall et al. noted, if 10 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 multiple comparison procedure critical value. Lastly, proportional power indicates what proportion of false null hypotheses one is likely to detect. We tend to prefer individual power because it focuses equally on all tests and most closely resembles the notion of power when testing hypotheses in isolation from one another.

Some newer power concepts have been developed for use when testing with the false discovery rate criterion. An example, due to Genovese and Wasserman (2002), is the false nondiscovery rate. Define *A* to be the number of accepted hypotheses and *T* the number of accepted hypotheses that are false. Then the false nondiscovery rate = $E(T/A)$ if $A > 0$ and = 0 if $A = 0$.

Types of Multiple Comparison Procedures

Multiple comparison procedures can examine pairwise hypotheses either simultaneously or sequentially. A simultaneous multiple comparison procedure 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* (see Einot & Gabriel, 1975). A sequential (stepwise) multiple comparison procedure 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. Multiple comparison procedures that require a significant omnibus test to conduct pairwise comparisons have been referred to as *protected tests*.

Multiple comparison procedures 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 statistic 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 case, one can have a smaller pairwise difference be significant and a larger pairwise difference be nonsignificant 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 simultaneous testing procedures and stepwise procedures is important to

note. Simultaneous test procedures allow researchers to examine simultaneous intervals around the statistics of interest, whereas stepwise procedures do not (see, however, Bofinger, Hayter, & Liu, 1993).

Test Statistics

This section presents test statistics that can be used to assess the significance of pairwise comparisons. Note, however, that other statistics can also be adopted (see Appendix A). A hypothesis for the pairwise comparison ($H_c : \mu_j - \mu_{j'} = 0$), when group sizes ($n_1$, n_2) are unequal, can be examined with the test statistic:

$$t_c = \frac{\bar{Y}_j - \bar{Y}_{j'}}{\sqrt{\frac{MSE}{n_j} + \frac{MSE}{n_{j'}}}}$$

where $\bar{Y}_j$ is the jth group mean ($j \neq j'$) and mean square error is the usual analysis of variance (ANOVA) estimate of error variance (see Appendix A for further details). Note that this is the usual two-sample Student t test, distributed as a t variate with $n_1 + n_2 - 2$ df. When group sizes are equal, the statistic (with $2[n - 1]$ df) would be

$$t_c = \frac{\bar{Y}_j - \bar{Y}_{j'}}{\sqrt{\frac{2MSE}{n}}}$$

Multiple Comparison Procedures for Normally Distributed Data/Homogeneous Population Variances

Tukey

Tukey (1953) proposed a simultaneous test procedure for all pairwise comparisons in what Toothaker (1991) described as possibly "the most frequently cited unpublished paper in the history of statistics" (p. 41). Tukey's multiple comparison procedure uses a critical value obtained from the Studentized range distribution $q(\alpha, J, \nu)$, where q is a value from the Studentized range distribution based on J means and ν degrees of freedom (see Kirk, 1995). 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

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

Tukey's procedure can be implemented in SAS's (1999) general linear model program (as well as SPSS [Norusis, 2002]). It is also important to note that Tukey's method, as well as other multiple comparison procedures, can be utilized when group sizes are unequal (see Appendix A). Westfall et al. (1999) enumerated the general linear model 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 (see Appendix A for details).

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). 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 $100(1 - \alpha/2)$ 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 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., $\mu_1 = \mu_2 = \cdots = \mu_{J-1} < < \mu_J$ all means but one equal and in which the set of $J - 1$ equal means is quite disparate from the one mean), the rate of familywise Type I error can be much in excess of the level of significance (Hayter, 1986; Hochberg & Tamhane, 1987; Keselman, Keselman, & Games, 1991).

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 $P(q_{(J-1,\, \nu)} > \sqrt{2}\ t_{(\alpha/2,\, \nu)})$. One can see that for $J > 3$ the maximum FWE will exceed the level of significance. In fact, Hayter showed that for $\nu = \infty$, $\alpha = .05$, FWE attains values of .1222 and .9044 for $J = 4$ and $J = 8$, respectively. Thus, this usual form of the least significant difference procedure does not provide a satisfactory two-stage procedure for researchers when $J > 3$.

Accordingly, Hayter (1986) proposed a modification to Fisher's (1935) least significant difference procedure that would provide strong control over FWE. Like the least significant difference procedure, no comparisons are tested unless the omnibus test is sig-

nificant. If the omnibus test is significant, then H_c is rejected if

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

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

It should be noted that many authors recommend Fisher's (1935) 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

Benjamini and Hochberg

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

The Benjamini and Hochberg (1995) procedure has been shown to control the FDR for several situations of dependent tests, that is, for a wide variety of multivariate distributions that make their procedure applicable to most testing situations social scientists might encounter (see Sarkar, 1998; Sarkar & Chang, 1997). In addition, simulation studies comparing the power of the Benjamini and Hochberg procedure to several FWE controlling procedures have shown that as the number of treatment groups increases (beyond $J = 4$), the power advantage of their 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, 2002; Miller, 1981). Therefore, control of the false discovery rate 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 (see Appendix A for details). The MULTTEST program (SAS Institute, 1999) can be used to obtain these adjusted p values (see Westfall et al., 1999). Benjamini and Hochberg (2000) also presented a modified (adaptive) version of their original procedure (see Appendix A for a description of this modification).

Closed Testing Sequential Multiple Comparison Procedures

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 simultaneous test procedures (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 multiple comparison procedures that can be obtained through the SAS system of programs (see Appendix A for greater detail).

As Westfall et al. (1999) noted, it was during the past two decades (prior to their 1999 publication) that a unified approach to stepwise testing evolved. The unifying concept has been the closure principle. Multiple comparison procedures based on this principle have been designated as *closed testing* procedures. 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. A coherent multiple comparison procedure is one that avoids inconsistencies in that it will not reject a hypothesis that is implied by a hypothesis that was not rejected. For example, a procedure that retains the hypothesis $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$ must also retain $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$ (see Appendix A for further details). Because closed testing procedures were not always easy to derive, various authors derived other simplified stepwise procedures that 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 (i.e., Type I errors are being controlled too tightly), these simpler stepwise multiple comparison procedures would not be as powerful as exact closed testing methods. Nonetheless, these methods are still typically more powerful than simultaneous test procedures (e.g., Tukey) and therefore are recommended; furthermore, researchers can obtain numerical results through the SAS system.

One such stepwise method was introduced by Ryan (1960), Einot and Gabriel (1975), and Welsch (1977) and is available through SAS (Ryan–Einot–Gabriel–Welsch [REGWQ]) as well as SPSS. One can better understand the logic of their procedure if we first introduce one of the most popular stepwise strategies for examining pairwise differences between means, the Newman–Keuls 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 a α level of significance. At each stage of testing, only subsets of ordered means that are statistically significant are subjected to further testing (with $\alpha = .05$). Although the Newman–Keuls procedure is very popular among applied researchers, it is becoming increasingly well known that when $J > 3$ it does not control FWE at α in the strong sense (see Hochberg & Tamhane, 1987).

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

- Test all subset ($p = 2, \ldots, J$) hypotheses at $\alpha_p = 1-(1-\alpha)^{\frac{p}{J}}$, for $p = 2, \ldots, J-2$ and at level $\alpha_p = \alpha$ for $p = J-1, J$.
- Start testing with an examination of the complete null hypothesis $\mu_1 = \mu_2 = \cdots = \mu_J$, and, if rejected, step down to examine subsets of $J - 1$ means, $J - 2$ means, and so on.
- Accept all subset hypotheses implied by a homogeneity hypothesis that has not been rejected as null without testing (see Appendix A for further comments).

The REGWQ procedure can be implemented with the SAS GLM program. Westfall et al. (1999) illustrated the additional power that researchers can obtain with this procedure as compared to Tukey's (1953) simultaneous method (see Appendix A for details regarding power analyses). We remind the reader, however, that this procedure cannot be used to construct simultaneous confidence intervals. We also note that there is an F test version of REGWQ available in SPSS (see Norusis, 2002; by right-clicking with one's mouse on the procedure, SPSS will provide a description); however, SAS has abandoned the F test version.

Multiple Comparison Procedures for Nonnormally Distributed Data

An underlying assumption of all of the previously presented multiple comparison procedures 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) suggested that most populations are skewed, contain outliers, or both. Researchers falsely assuming normally distributed data risk obtaining biased Type I error rates, Type II error rates, or both for many patterns of nonnormality, especially when other assumptions are also not satisfied (e.g., variance homogeneity; see Wilcox, 1997; Wilcox & Keselman, 2003).

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, leading to hypothesis tests of such 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 (see Appendix A for greater detail). An example program for all possible pairwise comparisons is given by Westfall et al. (1999).

Similarly, pairwise comparisons of means (or ranks) can be obtained through permutation of the data with the program provided by Westfall et al. (1999). 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 ($Y_{11},\ldots, Y_{n1}, Y_{1J},\ldots, Y_{nJJ}$) 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; see Appendix A for further details).

When users adopt this approach to combat the effects of nonnormality, they should heed the cautionary note provided by Westfall et al. (1999); 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.

Multiple Comparison Procedures 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 non-

robustness of multiple comparison procedures with conventional test statistics (e.g., t, F) and evidence that population variances are commonly unequal (Keselman, Huberty, et al., 1998), researchers who persist in applying multiple comparison procedures with conventional test statistics increase the risk of Type I errors. As Olejnik and Lee (1990) concluded, "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 (Games & Howell, 1976) or Dunnett (1980) procedures for designs with unequal σ_j^2s (e.g., see Kirk, 1995; Norusis, 2002; Toothaker, 1991), sequential procedures can provide more power than simultaneous test procedures while generally controlling the FWE (Hsuing & Olejnik, 1994). However, SPSS does provide simultaneous solutions based on Games–Howell, Dunnett, as well as other solutions, for unequal variances.

The SAS software can once again be used to obtain numerical results. In particular, Westfall et al. (1999) provided SAS programs for logically constrained step-down pairwise tests when heteroscedasticity exists. The macro uses SAS's mixed-model program (PROC MIXED), which allows for a nonconstant error structure across groups. Also, the program adopts the Satterthwaite (1946) solution for error *df*. Westfall et al. reminded the reader that the solution requires large data sets to provide approximately correct FWE control.

It is important to note that other non-SAS solutions are possible in the heteroscedastic case. For example, sequential procedures based on the usual t_c statistic can be easily modified for unequal σ_j^2s (and unequal n_js) 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 $s_{j'}^2$, and n_j and $n_{j'}$ represent the usual unbiased sample variances and group sizes for the jth 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 ν_W *df* (see Appendix A for the ν_W formula).

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 due to Welch (1951). With the Welch procedure (F_W), the omnibus null hypothesis is rejected if $F_W > F_{(J-1,\, \nu_W)}$ (see Appendix A for the formulas for F_W and ν_W). 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 multiple comparison procedures. For example, Keselman et al. (1999) verified that one can use this approach with Hochberg's (1988) step-up Bonferroni multiple comparison procedure (see Westfall et al., 1999) as well as with Benjamini and Hochberg's (1995) false discovery rate method to conduct all possible pairwise comparisons in the heteroscedastic case.

Multiple Comparison Procedures for Nonnormally Distributed Data/Heterogeneous Population Variances

A different type of testing procedure, based on trimmed (or censored) means, has been discussed by Yuen and Dixon (1973), Wilcox (1995, 2002, 2003), and Wilcox and Keselman (2003) 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; see also Wilcox & Keselman, 2003). By substituting robust measures of location and scale for the usual mean and variance, it should be possible to obtain test statistics that are insensitive to the combined effects of variance heterogeneity and nonnormality. 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 (e.g., Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990).

Although 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 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.

Specifically trimmed means are computed by removing a percentage of observations from each of the tails of a distribution (set of observations). To calculate a trimmed mean, simply remove a determined percentage of the observations and compute the mean from the remaining observations. To calculate a Winsorized variance, the smallest nontrimmed score replaces the scores trimmed from the lower tail of the distribution and the largest nontrimmed score replaces the observations removed from the upper tail. The nontrimmed and replaced scores are called *Winsorized scores*. A Winsorized variance is calculated by applying the usual formula for the variance to the Winsorized scores. The Winsorized variance is used because it can be shown that the standard error of a trimmed mean is a function of the Winsorized variance (see Wilcox & Keselman, 2003). A common symmetric trimming percentage is 20%. See Wilcox (2003) for a justification of 20% trimming. Computational definitions of the trimmed mean and Winsorized variance are given in Appendix A.

To test a pairwise comparison null hypothesis, compute $\bar{Y}_t$ and d, the sample trimmed mean and squared standard error of the mean, respectively, for the jth group; label the results $\bar{Y}_{tj}$ and d_j. The robust pairwise test (see Keselman, Lix, & Kowalchuk, 1998) becomes

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

(see Appendix A for the ν_{Wt} formula). 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: $\mu_{tj} = \mu_{tj'}$, where μ_t represents the population trimmed mean.

Yuen and Dixon (1973) and Wilcox (1995) reported 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 multiple comparison procedures (see Keselman, Othman, Wilcox, & Fradette, 2004, for another robust statistic).

A Model-Testing Procedure

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

When testing all pairwise comparisons, intransitive decisions are extremely common with conventional multiple comparison procedures (Dayton, 1998). An intransitive decision refers to declaring a population mean (μ_j) not significantly different from two different population means ($\mu_j = \mu_{j'}$, $\mu_j = \mu_{j''}$), when the latter two means are declared significantly different ($\mu_{j'} \neq \mu_{j''}$). 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 multiple comparison procedure. 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. Mutually exclusive and transitive models are each evaluated using Akaike's criterion, and the model having the minimum criterion is retained as the most probable population mean configuration (see Appendix A for details). 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 in which Groups 1 and 2 are derived from the same population, whereas Groups 3 and 4 each represent distinct populations. The model having the lowest criterion value would be retained as the most probable population model.

Dayton's (1998) model-testing approach has the virtue of avoiding intransitive decisions. It is more powerful in the sense of all-pairs power than Tukey's honestly significant difference, which is not designed to avoid intransitive decisions. One finding reported by Dayton, as well as Huang and Dayton (1995), is that the information criterion approach 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}, his information criterion approach selected the more complicated pattern {1, 2, 3, 4} more than 10% of the time, whereas the information criterion only rarely selected less complicated models (e.g., {12, 34}). This tendency can present a special problem for the complete null case {1234}, in which the infor-

mation criterion approach has a tendency to select more complicated models. Consequently, a recommendation by Huang and Dayton 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 Akaike's (1974) Information Criterion, and the model having the minimum criterion is retained as the most probable population mean configuration (see Appendix A for details). As in the original Dayton procedure, an appropriate omnibus test can also be applied.

Generalizations to Factorial Designs

Though we have presented "newer" pairwise multiple comparison procedures within the context of a one-way completely randomized design, the reader should note that all of the methods we have discussed can be applied to examining pairwise differences in more complex designs, specifically for examining pairwise differences in factorial designs. For example, the multiple comparison procedures that use the Welch (1938) heteroscedastic statistic with trimmed means can be adapted to completely randomized factorial designs and between by within-subjects repeated measures designs. Lix and Keselman (1995) and Keselman, Wilcox, and Lix (2003) demonstrated the use of this statistic for performing pairwise contrasts in these designs. In particular, they demonstrated the computation of pairwise differences on the marginal cell means as well as tetrad (i.e., pairwise by pairwise) contrasts of the interaction effect (see Jaccard & Guilamo-Ramos, 2002a, 2002b).

Researchers can also compute pairwise differences within a particular row or column of their factorial design. Though it is well known that these tests do not examine interaction effects, they nonetheless can provide applied researchers with valuable information (see Jaccard & Guilamo-Ramos, 2002a, 2002b). Accordingly, we note that appropriate multiple comparison procedures are available for this task (see Appendix A and Cheung & Chan, 1996).

Summary

Selecting an appropriate multiple comparison procedure 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 to the selection process. In addition, the selection of a proper multiple comparison procedure 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 multiple comparison procedures have been proposed that improve on one or more aspects of previously recommended multiple comparison procedures. 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 also 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 that can be obtained through a statistical package are likely to be adopted by researchers. Accordingly, we emphasized many of the procedures that are available through the SAS system because Westfall et al. (1999) have provided many useful programs for obtaining numerical solutions. In conclusion, we encourage 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 multiple comparison procedures so the reader can check his or her facility to work with the SAS/Westfall et al. (1999) programs and to demonstrate through example the differences between their operating characteristics.

Consider a study of shyness and social anxiety in which the primary outcome variable was a fear of negative evaluation (i.e., apprehension about negative evaluations by others; Watson & Friend, 1969). One hundred fifty adolescents read one of five scenarios describing a social situation and then indicated the extent to which they would be concerned about a negative evaluation in that situation using a modified version of Leary's (1983) brief version of the Watson and Friend Fear of Negative Evaluation scale. The measure asked respondents to rate 12 statements taken from this scale but that were adapted for situational analysis (e.g., "In this situation, I would be worried about what kind of impression I would make"). Each statement was rated on a 5-point, Likert type scale of 1 (*strongly disagree*), 2 (*moderately disagree*), 3 (*neither agree nor disagree*), 4 (*moderately agree*), or 5 (*strongly agree*). Higher scores indicate a response pattern consistent with greater fear of negative evaluation. The responses were averaged across the 12 items.

The five scenarios described a social interaction in a school setting, and the only attribute that varied was the description of the main person the adolescent would be interacting with in the scenario. The person was described as either being two grade levels below the respondent, one grade level below the respondent, the same grade level as the respondent, one grade level above the respondent, or two grade levels above the respondent. The investigator was interested in the extent to which adolescents had greater fear of negative evaluations as a function of the age or grade level of the person they were interacting with.

The data ($n_1 = n_2 = \cdots = n_J = 20$) presented in Table 1 were randomly generated by us, though they could represent the outcomes of the problem just described.

Table 2 contains adjusted p values and FWE (α = .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 multiple comparison procedures that we discussed previously. In particular, of the 10 comparisons, 5 were found to be statistically significant with the Tukey (1953), Hayter (1986), REGWQ, bootstrap, stepdown bootstrap, and permutation procedures: $\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_1 - \mu_5$, $\mu_2 - \mu_4$, and $\mu_2 - \mu_5$. The Benjamini and Hochberg (1995) and adaptive Benjamini and Hochberg (2000) multiple comparison procedures, on the other hand, resulted in 6 statistically significant comparisons; $\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_1 - \mu_5$, $\mu_2 - \mu_4$, $\mu_2 - \mu_5$, and $\mu_3 - \mu_5$. (Numerical results for the adaptive Benjamini and Hochberg procedure were not obtained through SAS; they were obtained through hand calculations.) Clearly the procedures based on the more liberal false discovery rate found more comparisons to be statistical significant than the FWE controlling multiple comparison procedures.

We also investigated the 10 pairwise comparisons with the trimmed means and model-testing procedures; the results for the trimmed means analysis are also reported in Table 2. Numerical results for robust estimation and testing were obtained through SPSS (Norusis, 2002; see Appendix B for the SPSS program). In particular, we computed the group trimmed means $\bar{Y}_{t1} = 1.49$, $\bar{Y}_{t2} = 1.91$, $\bar{Y}_{t3} = 2.53$, $\bar{Y}_{t4} = 2.9$ $\bar{Y}_{t5} = 3.15$) as well as the group Winsorized standard deviations ($\hat{\sigma}_{W1} = .532$, $\hat{\sigma}_{W2} = .607$, $\hat{\sigma}_{W3} = .306$ $\hat{\sigma}_{W4} = .655$, $\hat{\sigma}_{W5} = .550$). One can use the SAS/IML program referred to by Keselman et al. (2003) to obtain the trimmed means and Winsorized standard deviations. These values (as well as the effective sample sizes $h_1 = \cdots = h_5 = 12$) were then read into an SPSS file and we allowed SPSS to calculate nonpooled t statistics (t_{Wt} and ν_{Wt}) and their corresponding p values (through the ONEWAY program). The results reported in Table 2 indicate that with this approach, seven comparisons were found to be statistically significant; that is, all comparisons except $\mu_1 - \mu_2$, $\mu_3 - \mu_4$, and $\mu_4 - \mu_5$.

With regard to the model-testing approach we examined the 2^{J-1} models of nonoverlapping subsets of ordered means and used the minimum Akaike (1974) Information Criterion 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 "winning" models combine one pair, but there are other models that are plausible given the data available. That is, there are 3

Table 1. *Data Values and Summary Statistics (Means and Standard Deviations) for the Empirical Example*

Two Grades Below	One Grade Below	Same Grade	One Grade Above	Two Grades Above
1.07	2.01	2.47	3.65	2.94
1.77	1.30	2.82	3.58	3.04
1.58	3.21	2.52	1.91	3.01
1.51	1.79	2.83	4.21	2.06
0.32	1.99	2.60	3.18	3.32
0.39	3.09	2.25	2.27	1.87
1.55	1.83	3.16	2.71	2.32
3.11	1.75	2.55	3.71	4.00
1.65	2.56	2.55	3.63	5.21
1.55	1.22	2.75	1.10	2.25
1.53	2.00	1.55	4.17	3.65
0.87	2.85	0.59	2.45	4.23
2.94	0.89	3.14	3.17	3.76
1.10	1.78	3.36	2.74	3.87
1.44	1.38	2.83	1.68	2.63
0.37	1.75	1.72	3.50	2.87
0.58	1.09	2.14	2.42	2.56
2.61	2.92	4.38	2.08	3.52
2.39	3.08	1.96	1.01	3.91
2.27	1.02	2.11	3.41	2.69
M 1.530	1.974	2.513	2.828	3.186
SD 0.823	0.742	0.772	0.948	0.836

Table 2. *Adjusted p Values and FWE (α = .05) Significant (*) Comparisons*

$\mu_j - \mu_{j'}$	Tukey	Hayter	REGWQ	Boot	Stepb	Perm	BH	BH–A	TM
1 vs 2	.4368			.4370	.2373	.4334	.1146		.2046
1 vs 3	.0026	*	*	.0025	.0018	.0026	.0008	*	.0010
1 vs 4	<.0001	*	*	<.0001	<.0001	<.0001	.0002	*	.0009
1 vs 5	<.0001	*	*	<.0001	<.0001	<.0001	.0002	*	<.0001
2 vs 3	.2466			.2489	.1446	.2453	.0603		.0392
2 vs 4	.0130	*	*	.0127	.0080	.0127	.0030	*	.0129
2 vs 5	.0001	*	*	.0002	.0001	.0002	.0002	*	.0016
3 vs 4	.7488			.7499	.2967	.7472	.2314		.2038
3 vs 5	.0847			.0851	.0506	.0854	.0197	*	.0298
4 vs 5	.6525			.6533	.2967	.6509	.1956		.5048

Note: Tukey–Tukey (1953); Hayter–Hayter (1986); REGWQ = Ryan (1960)–Einot & Gabriel (1975)–Welsch (1977); Boot = Bootstrap; Stepb = Stepwise bootstrap; Perm = Permutation—Westfall et al. (1999); BH = Benjamini & Hochberg (1995); BH–A = Adaptive—Benjamini & Hochberg (2000); TM = trimmed means (and Winsorized variances) used with a nonpooled *t* test and BH critical constants. Raw trimmed mean *p* values: .1842, .0003, .0002, <.0001, .0275, .0065, .0006, .1631, .0179, .5048.

to 4 similar models that differ with respect to whether one or another pair of means is combined, or none: {1, 2, 34, 5}, {1, 2, 3, 45}, {1, 2, 3, 4, 5}, and {12, 34, 5} with criterion values of 252.24, 252.65, 252.66, and 253.21, respectively. If one must choose a best model, then it is {1, 2, 34, 5} (minimum criterion value; results were obtained through hand calculations. However, a GAUSS program is available from the Department of Measurement and Statistics, University of Maryland Web site). Though this ambiguity might seem like a negative feature of the model-testing approach, Dayton would maintain that being able to enumerate a set of conclusions (i.e., competing models) provides a broader, more comprehensive perspective regarding group differences than does the traditional approach.

Recommendations

We conclude by suggesting "optimal" multiple comparison strategies researchers may follow with the following caveat in mind. That is, it is not possible to recommend one specific pairwise multiple comparison procedure across all situations that applied researchers may encounter; no one procedure would be "best." As stated earlier, the choice among multiple comparison procedures rests on issues such as (a) the type of error control one wants to set, (b) the balance between Type I and Type II (hence power) errors that is sought, (c) whether the classical assumptions of normality and homoscedasticity hold, and (d) the availability of statistical software to obtain numerical results—that is, an ease of computation issue. Thus, we enumerate a number of scenarios that might describe one's data and within each of these contexts suggest a multiple comparison procedure that we like.

Scenario 1

Scenario 1 is a design with four or five groups in which the sample sizes are somewhat small and power is a bit of a problem. The researcher has a strong enough theory to motivate including all the groups in the study, but not so strong a theory as to have a priori predictions about which groups will differ in which ways. Normality and variance homogeneity assumptions are tenable. The researcher does not want to adopt a method that provides weak Type I error (FWE) control and would prefer one that provides strong(er) control. That is, the researcher seeks to limit the overall rate of Type I error at some reasonable value (e.g., 5%) and achieve as much power as possible to detect pairwise differences given the small sample sizes available.

Recommendations. When there are four or five groups there would be a total of 6 or 10 pairwise comparisons, respectively. In the first case we would recommend the REGWQ procedure. As we noted previously, it is a stepwise procedure that should provide more power to detect pairwise differences than the simultaneous test procedure due to Tukey (1953). Also, numerical computations can be implemented with the SAS GLM program. However, in the second case, in which many comparisons will be computed, researchers can optimize their likelihood of detecting nonnull pairwise differences by adopting the false discovery rate method of control due to Benjamini and Hochberg (1995, 2000). Numerical results can either be obtained through SAS's MULTTEST or as we have demonstrated with our numerical example. For researchers who want to avoid the possibility of intransitive outcomes, Dayton's (1998) model-testing approach can be adopted. Remember that this approach provides, according to Dayton, a holistic examination of the pairwise multiple comparison problem by revealing a set of competing models that purport to describe the "true" state of nature. Lastly, if the researcher wants to set simultaneous intervals around pairwise differences, then he or she must select a simultaneous procedure—Tukey.

Scenario 2

Scenario 2 is the set-up described in Scenario 1; however, normality and variance homogeneity assumptions are not tenable.

Recommendations. Our recommendations for this scenario are basically the ones we offered in Scenario 1; however, rather than adopting the usual statistic (i.e., Student's *t* test) with the usual least squares estimators for mean and variance, researchers should use a heteroscedastic statistic, such as Welch's (1938) *t* test, with robust estimators, that is trimmed means and Winsorized variances. This robust procedure can then be applied with the REGWQ or Benjamini and Hochberg (1995, 2000) procedures. Though the major statistical software packages do not perform robust analyses, a numerical solution can easily be obtained. Keselman et al. (2003) provided a SAS/IML (SAS Institute, 1989) program that can be used to obtain numerical results for omnibus and focused (e.g., pairwise comparisons) tests in independent and correlated groups designs (see also Appendix C). Adopting robust procedures is particularly crucial when group sizes are unequal. Unequal group sizes, for example, exacerbate the effects of variance heterogeneity. One can set simultaneous confidence intervals around pairwise differences involving robust statistics (i.e., trimmed means; see Wilcox, 2003).

Scenarios 3 and 4

Scenarios 3 and 4 are identical to the Scenarios 1 and 2 but for exactly three groups.

Recommendations. For Scenario 3 we would recommend the Fisher (1935)–Hayter (1986) two-stage procedure. In the three-group problem, one can in fact use Fisher's least significant difference procedure, setting a CWE rate of $\alpha = .05$ on each pairwise difference, and the FWE will not exceed .05. Accordingly, this procedure should optimize one's chances of finding true pairwise differences among the population means. If simultaneous confidence intervals are of interest, then researchers could adopt Tukey's (1953) procedure (see Westfall et al., 1999). Again, researchers have the alternative choice of adopting Dayton's (1998) model-testing approach. For Scenario 4, we also recommend the Fisher two-stage test; however, as in Scenario 2, we recommend that researchers adopt a heteroscedastic statistic based on trimmed means and Winsorized variances (see Cribbie & Keselman, 2003a, 2003b). Again, if one wants to set simultaneous confidence intervals when assumptions (normality and homogeneity) are not satisfied, robust procedures should be used (e.g., see Wilcox, 2003).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC–19,* 716–723.

Bellanti, C. J., & Bierman, K. L. (2000). Disentangling the impact of low cognitive ability and inattention on social behavior and peer relationships. *Journal of Clinical Child Psychology, 29,* 66–75.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series 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.

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.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement of visual speech gestures. *Neuroreport, 14,* 2213–2218.

Carlson, C. L., & Mann, M. (2002). Sluggish cognitive tempo predicts a different pattern of impairment in the attention deficit hyperactivity disorder, predominantly inattentive type. *Journal of Clinical Child and Adolescent Psychology, 31,* 123–129.

Cheung, S. H., & Chan, W. S. (1996). Simultaneous confidence intervals for pairwise multiple comparisons in a two-way unbalanced design. *Biometrics, 52,* 463–472.

Copenhaver, M. D., & Holland, B. (1988). Computation of the distribution of the maximum Studentized range statistic with application to multiple significance testing of simple effects. *Journal of Statistical Computation and Simulation, 30,*1–15.

Cribbie, R. A., & Keselman, H. J. (2003a). Pairwise multiple comparisons: A model comparison approach versus stepwise procedures. *British Journal of Mathematical and Statistical Psychology, 56,* 167–182.

Cribbie, R. A., & Keselman, H. J. (2003b). The effects of nonnormality on parametric, nonparametric and model comparison approaches to pairwise comparisons. *Educational and Psychological Measurement, 63,* 615–635.

Daly, S. L., & Glenwick, D. S. (2000). Personal adjustment and perceptions of grandchild behavior in custodial grandmothers. *Journal of Clinical Child Psychology, 29,* 108–118.

Dayton, C. M. (1998). Information criteria for the paired-comparisons problem. *The American Statistician, 52,* 144–151.

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science, 18,* 71–103.

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.

Epkins, C. C. (2000). Cognitive specificity in internalizing and externalizing problems in community and clinic-referred children. *Journal of Clinical Child Psychology, 29,* 199–208.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh, England: 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.

Genovese, C., & Wasserman, L. (2002). Operating characteristics and extensions of the False Discovery Rate procedure. *Journal of the Royal Statistical Society, Series B*, *64,* 1151–1160.

Gross, A. M. (1976). Confidence interval robustness with long-tailed symmetric distributions. *Journal of the American Statistical Association, 71,* 409–416.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* New York: Wiley.

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. (1984). A proof of the conjecture that the Tukey–Kramer multiple comparisons procedure is conservative. *Annals of Statistics, 12,* 61–75.

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: Wiley.

Holland, B., & Cheung, S. H. (2002). Family size robustness criteria for multiple comparison procedures. *Journal of the Royal Statistical Society, Series B, 64,* 63–77.

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.

Huber, P. J. (1981). *Robust statistics.* New York: Wiley.

Jaccard, J., & Guilamo-Ramos, V. (2002a). Analysis of variance frameworks in clinical child and adolescent psychology: Advanced issue and recommendations. *Journal of Clinical Child Psychology, 31,* 278–294.

Jaccard, J., & Guilamo-Ramos, V. (2002b). Analysis of variance frameworks in clinical child and adolescent psychology: Issues and recommendations. *Journal of Clinical Child and Adolescent Psychology, 31,* 130–146.

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., Games, P. A., & Rogan, J. C. (1979). Protecting the overall rate of Type 1 errors for pairwise comparisons with an omnibus test statistic. *Psychological Bulletin, 86,* 884–888.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A, Donahue, B., et al. (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.

Keselman, H. J., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample *t* test. *Psychological Science, 15,* 47–51.

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology, 40,* 586–596.

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.

Kramer, C. Y. (1956). Extension of the multiple range test to group means with unequal numbers of replications. *Biometrics, 12,* 307–310.

Leary, M. R. (1983). A brief version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin, 9,* 371–375.

Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin, 115,* 153–159.

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin, 117,* 547–560.

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., Jr. (1981). *Simultaneous statistical inferencen*(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. (2002). *SPSS 11.0 Guide to data analysis.* Upper Saddle River, NJ: 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 non-normal variation. *Biometrika, 23,* 114–133.

Podolski, C.-L., & Nigg, J. T. (2001). Parent stress and coping in relation to child ADHD severity and associated child disruptive behavior problems. *Journal of Clinical Child Psychology, 30,* 503–513.

Rothman, K. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1,* 43–46.

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.

Sarkar, S. K. (1998). Probability inequalities for ordered MTP2 random variables: A proof of the Simes conjecture. *Annals of Statistics, 26,* 494–504.

Sarkar, S. K., & Chang, C.-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association, 92,* 1601–1608.

SAS Institute Inc. (1989). *SAS/IML software: Usage and reference, version 6* (1st ed.). Cary, NC: Author.

SAS Institute Inc. (1999). *SAS/STAT user's guide, Version 7.* Cary, NC: Author.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *The American Statistician, 44,* 174–180.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

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.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing.* New York: Wiley.

Toothaker, L. E. (1991). *Multiple comparisons for researchers.* Newbury Park, CA: Sage.
Tukey, J. W. (1953). *The problem of multiple comparisons.* Unpublished manuscript, Princeton University.
Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.), *Contributions to probability and statistics.* Stanford University Press, Stanford.
Watson, D., & Friend, R. (1969). Measurement of social- evaluative anxiety. *Journal of Consulting and Clinical Psychology, 33,* 448–457.
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., Tobias, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (1999). *Multiple comparisons and multiple tests.* Cary, NC: SAS Institute.
Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment.* New York: Wiley.
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). *Introduction to robust estimation and hypothesis testing.* San Diego: Academic Press.
Wilcox, R. R. (2002). Understanding the practical advantages of modern ANOVA methods. *Journal of Clinical Child and Adolescent Psychology, 31,* 399–412.
Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* New York: Academic.
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.
Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8,* 254–274.
Yuen, K. K., & Dixon, W. J. (1973). The approximate behavior of the two-sample trimmed *t*. *Biometrika, 60,* 369–374.

Appendix A

False Discovery Rate Control

Suppose we have J means, $\mu_1, \mu_2, \ldots, \mu_J$, 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 $E(V/C)$, whereas the familywise rate is given by $P(V \geq 1)$. Thus, testing each and every comparison at α guarantees that $E(V/C) \leq \alpha$, whereas, according to the Bonferroni inequality, testing each and every comparison at level α/C guarantees that $P(V \geq 1) \leq \alpha$.

According to Benjamini and Hochberg (1995), the proportion of errors committed by falsely rejecting null hypotheses can be expressed through the random variable $Q = V/R$, that is, the proportion of rejected hypotheses that 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.) False discovery rate was defined by Benjamini and Hochberg as the mean of Q, that is

$$E(Q) = E\left(\frac{V}{R}\right), \text{ or } E(Q) = E\left(\frac{\text{Number of false rejections}}{\text{Number of rejections}}\right)$$

That is, the false discovery rate is the expected proportion of false discoveries or false positives.

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

(a) If $\mu_1 = \mu_2 = \cdots = \mu_J$, then all C pairwise comparisons truly equal zero, and therefore the false discovery rate is equivalent to the familywise rate; that is, in the case of the complete null being true, false discovery rate 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) In testing the family of pairwise hypotheses, of which m_0 are true, when $m_0 < C$, the false discovery rate is smaller than or equal to the familywise rate of error. The false discovery rate 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 the false discovery rate is as well. Moreover, if one adopts a procedure that provides false discovery rate control, rather than strong (i.e., over all possible mean configurations) familywise control, then based on the preceding relation, 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 false discovery rate and the familywise value and thus a greater likelihood of increased power by adopting false discovery rate control.

Adjusted *p* Values

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, where C stands for the number of comparisons (tests). Note that this is equivalent to rejecting any H_{0c} for which $C \cdot p_c$ is less than or equal to α, where p_c is the usual (nonmultiplicity adjusted) p value. Therefore, Bonferroni adjusted p values are

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

Westfall et al. (1999) provided programs (2.4 and 2.5) that convert unadjusted p_c values to multiplicity adjusted $\tilde{p}_c$.

Error Rate Definitions

For completeness, we provide definitions of these rates of error. The comparisonwise error rate is defined as

$$CWE = P(\text{Reject } H_0 \,|\, H_0 \text{ is true})$$

That is, the comparisonwise error rate is the usual probability of rejecting a null hypothesis (H_0), given that (|) the null hypothesis is true. On the other hand, the FWE rate for multiple tests of significance in which some hypotheses ($H_{0j1}, H_{0j2}, \ldots, H_{0jm}$) are true and the remaining ($k - m$) are false is given by

$$FWE = P(\text{Reject at least one of } H_{0j1}, H_{0j2}, \ldots, H_{0jm} \,|\, H_{0j1}, H_{0j2}, \ldots, H_{0jm} \text{ all are true})$$

A Statistical Model for the Problem

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

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

where Y_{ij} is the score of the ith participant ($i = 1, \ldots, n$) in the jth ($j = 1, \ldots, J$) group ($\Sigma_j n = N$), μ_j is the jth group mean, and ε_{ij} is the random error for the ith participant in the jth 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^2s) are equal. Relevant sample estimates include

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

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

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

The Unequal Group Size Version of Tukey's (1953) Multiple Comparison Procedure—The Tukey–Kramer (1956) Method

A pairwise test statistic uses the unequal sample size case t test formula presented in the body of the text and significance is assessed by comparing

$$|t_c| > q_{(J,\Sigma_j(n_j-1))} \Big/ \sqrt{2}$$

Hayter (1984) proved that the Tukey–Kramer multiple comparison procedure only approximately controls the FWE—the rate is slightly conservative; that is, the true rate of Type I error will be less than the significance level. The general linear model procedure in SAS will automatically compute the Kramer version of Tukey's test when group sizes are unequal.

It should be noted that pairwise comparisons can be computed with statistics other than Student's t test. In particular, one may use a Studentized range (q) or F test. For a pairwise comparison, these tests, respectively, would be

$$q_c = \frac{\bar{Y}_j - \bar{Y}_{j'}}{\sqrt{\left[\frac{MSE}{n_j} + \frac{MSE}{n_{j'}}\right] \Big/ 2}}$$

and

$$F_c = \frac{\left(\bar{Y}_j - \bar{Y}_{j'}\right)^2}{\frac{MSE}{n_j} + \frac{MSE}{n_{j'}}}$$

Statistical packages use different test statistics and authors of texts and articles do as well. We choose the t test because of its almost universality. However, by understanding the relation between these tests, researchers can easily convert from one to the other. The relations one should know are

$$t_\nu^2 = F_{1,\nu}$$
$$t_\nu = \sqrt{F_{1,\nu}}$$
$$t_\nu = q_{(J,\nu)} / \sqrt{2}$$
$$F_{1,\nu} = q_{(J,\nu)}^2 / 2$$

Power Analysis for the Tukey Multiple Comparison Procedure

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 SASs 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) provided a macro (%Individual Power) for obtaining numerical results. These authors also provide a macro (%SimPower) that computes complete, minimal, and proportional power.

Adjusted *p* Values for the Benjamini and Hochberg (1995) False Discovery Rate Multiple Comparison Procedure

Statistical significance can be assessed once again with adjusted *p* values. For the Benjamini and Hochberg (1995) false discovery rate method of control, the adjusted *p* values are

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

Note that min stands for minimum; thus min(*a, b*) means select *a* if it is the minimum value or *b* if it is the minimum value. Westfall et al. (1999) provided a program (2.11) that converts unadjusted p_c values to adjusted false discovery rate adjusted values. With adjusted $\tilde{p}$ values researchers can present a fuller interpretation of their results. For example, the first 5 of 10 false discovery rates adjusted $\tilde{p}$ values reported by Westfall et al. in the output from program 2.11 are 0.00100, 0.02900, 0.04400, 0.07225, and 0.09960. With this information, a researcher can state that three of the comparisons were significant at the $\alpha = .05$ level (i.e., 0.00100, 0.02900, 0.04400), whereas two comparisons were significant at $\alpha = .10$ the level (i.e., 0.07225 and 0.09960).

We illustrate this algorithm with the numerical example raw *p* values (see Table 2). After rank ordering the raw *p* values, we see that the comparison of Group 4 versus Group 5 has the largest raw *p* value, .5048, hence $\tilde{p}_{(10)} = .5048$. The second largest raw *p* value is for the comparison of Groups 1 and 2 and equals .1842.

Accordingly, $\tilde{p}_{(9)} = \min[\tilde{p}_{(10)}, (10/9)(p_{(9)})] = \min[.5048, (10/9)(.1842)] = .2046$. Based on the next largest raw *p* value (.1631—Group 3 vs. Group 4), $\tilde{p}_{(8)} = \min[\tilde{p}_{(9)}, (10/8)(p_{(9)})] = \min[.2046, (10/8)(.1631)] = .2038$ and so on.

Benjamini and Hochberg's (2000) Modified (Adaptive) Multiple Comparison Procedure

Benjamini and Hochberg (2000) also presented a modified (adaptive) version of their original procedure that utilizes the data to estimate the number of true H_cs. (The adaptive Benjamini and Hochberg procedure has only been demonstrated, *not proven*, to control false discovery rate, 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 false discovery rate is controlled at a level less than that specified (α).

To compute the modified Benjamini and Hochberg (2000) procedure, the p_c values are ordered (smallest to largest) $p_1, \ldots, p_c$, and for any $c = C, C - 1, \ldots, 1$, if $p_c \leq \alpha(c/C)$, reject all $H_{c'}$ ($c' \leq c$), as in the Benjamini and Hochberg (1995) procedure. If all H_cs are retained, testing stops. If any H_c is rejected with the criterion of the Benjamini and Hochberg procedure, then testing continues by estimating the slopes $S_c = (1 - p_c)/(C + 1 - c)$, where $c = 1, \ldots, C$. Then, for any $c = C, C - 1, \ldots, 1$, if $p_c \leq \alpha(c/\hat{C}_T)$, reject all $H_{c'}$ ($c' \leq c$), where $\hat{C}_T = \min[(1/S^*) + 1, C]$, $[x]$ 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 modified Benjamini and Hochberg (2000) procedure, noted by both Benjamini and Hochberg and Holland and Cheung (2002), 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 modified Benjamini and Hochberg procedure 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 modified Benjamini and Hochberg procedure would involve comparing $p_6 = .1239$ to $\alpha(c/C) = .05(6/6) = .05$. Because $.1239 > .05$, the procedure would continue by comparing $p_5 = .0490$ to $\alpha(c/C) = .05(5/6) = .0417$. Again, because $.0490 > .0417$, the procedure would continue by comparing $p_4 = .0145$ to $\alpha(c/C) = .05(4/6) = .0333$. Because $.0145 < .0333$, 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 $c = 1, \ldots, C$. The calculated slopes for this example are as follows: $S_1 = .1664$, $S_2 = .1991$, $S_3 = .2475$, $S_4 = .3285$, $S_5 = .4755$, 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 $\hat{C}_T = \min[(1/S^*) + 1, C] = \min[(1/6) + 1, 6] = \min[1.1667, 6] = 1$. Therefore, the modified Benjamini and Hochberg procedure would compare $p_6 = .1239$ to $\alpha(c/\hat{C}_T) = .05(6/1) = .30$.

Because .1239, < .30, but .1239 > α, H_6 would not be rejected and the procedure would continue by comparing $p_5 = .0490$ to $\alpha(c/\hat{C}_T) = .05(5/1) = .25$. Because .0490 < .25 and .0490 < α, 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

These methods are designated as closed testing procedures because they address families of hypotheses that are closed under intersection (∩). 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" (Westfall et al., 1999, 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 14 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 $H_{1,2,3,4}$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.
- 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^{++}.

Coherent Results

According to Marcus, Peritz, and Gabriel (1976) and as enumerated by Westfall et al. (1999), 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" Westfall et al. (1999, p. 151).

REGWQ

Westfall et al. (1999) indicated "by using the Ryan–Einot–Gabriel–Welsch Q procedure, strong control is conservatively ensured by testing *directly* all subset homogeneity hypotheses, and *indirectly* all subset intersection hypotheses" (p. 154). 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 population means corresponding to the smallest and largest means in the set.

Power Analysis for REGWQ

Researchers can use Westfall et al.'s (1999) macro (%SimPower) to examine complete, minimal, and proportional power.

Bootstrapped Adjusted *p* Values

The empirical distribution, say $\hat{F}$, is obtained by sampling, *with replacement,* the pooled sample residuals $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{\mu}_{ij} = Y_{ij} - \overline{Y}_j$. 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.

Westfall et al.'s (1999) PROC MULTTEST computes adjusted *p* values. 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, $p_1^*, p_2^*, \ldots, p_C^*$ values are obtained from the pairwise tests.
- This 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.

A Permutation Multiple Comparison Procedure

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 permutation p value can be computed as $p = P(\bar{Y}_1^* - \bar{Y}_2^* \geq \bar{Y}_1 - \bar{Y}_2)$ (this is for an upper-tailed test; this statement is modified for a lower- or two-tailed test). 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 permutated data.

Welch's (1938) Formula for ν_W

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

In the formula ν_W, s_j^2 and n_j stand for the jth ($j \neq j'$) unbiased group variance and sample size, respectively.

Welch's (1951) Formula for F_W and ν_W

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

where $w_j = n_j / s_j^2$, and $\tilde{Y} = \Sigma w_j \bar{Y}_j / \Sigma w_j$. 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

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

Trimmed Means and Winsorized Variances

Let $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$ 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, 2002) suggested 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

$$\bar{Y}_W = \frac{1}{n}\left[(g+1)Y_{(g+1)} + Y_{(g+2)} + \ldots + Y_{(n-g-1)} + (g+1)Y_{(n-g)}\right]$$

and the sample Winsorized sum of squared deviations is

$$SSD_W = (g+1)(Y_{(g+1)} - \bar{Y}_W)^2 + (Y_{(g+2)} - \bar{Y}_W)^2 + \ldots + (Y_{(n-g-1)} - \bar{Y}_W)^2 + (g+1)(Y_{(n-g)} - \bar{Y}_W)^2$$

Accordingly, the sample Winsorized variance is $\hat{\sigma}_W^2 = SSD_W/(n-1)$ and the squared standard error of the trimmed mean is estimated as (Staudte & Sheather, 1990)

$$d = \frac{(n-1)\hat{\sigma}_W^2}{h(h-1)}$$

Error Degrees of Freedom ν_{Wt} When Using Trimmed Means

$$\nu_{Wt} = \frac{(d_j + d_{j'})^2}{\frac{d_j^2}{h_j - 1} + \frac{d_{j'}^2}{h_{j'} - 1}}$$

Dayton's (1998) Model-Testing Approach with Akaike's (1974) Information Criterion

According to Akaike's (1974) Information Criterion, the model having the minimum criterion is retained as the most probable population mean configuration, where

$$AIC = SS_W + \Sigma_j n_j(\bar{Y}_j - \bar{Y}_{mj})^2 + 2q$$

$\bar{Y}_{mj}$ is the estimated sample mean for the jth group (given the hypothesized population mean configuration for the mth model), SS_W is the ANOVA pooled within group sum of squares, and q is the degrees of freedom for the model.

For heterogeneous variances, Akaike's (1974) Information Criterion is:

$$AIC = -2\{(-N/2)(\ln(2\pi)+1) - 1/2(\Sigma n_j \ln(S))\} + 2q$$

where S is the biased variance (i.e., SS/N) for the jth group, substituting the estimated group mean (given the hypothesized mean configuration for the mth model) for the actual group mean in the calculation of the variance.

Simple Main-Effect Testing

In a two-way layout consisting of r rows and c columns, let μ_{ij} denote the mean response of the cell in row i and column j. Interaction is said to exist if the pattern of differences among the means of the columns differs from row to row and equivalently if the pattern of differences among the means of the rows differs from column to column. The presence of interaction can be detected as nonparallel traces in a standard interaction plot, or more formally as a significant F test for interaction in an analysis of variance.

When interaction is present and the levels of the factors are qualitative (rather than quantitative), one cannot compare the column means averaged over the r rows (equivalently, the row means averaged over the c columns). In this situation it would be incorrect to employ the usual Tukey procedure to detect pairwise column differences. Instead, a correct analysis covering the simultaneous comparison of all simple effects consists of examining the simple effects in each row, that is, pairwise cell mean differences of the form $\mu_{ij} - \mu_{ij'}$. These are all $c(c-1)/2$ pairwise mean differences in each of rows $i = 1,2, \ldots, r$, or $rc(c-1)/2$ comparisons in all. If one wants to control the FWE at α for this superfamily, an extension of the Tukey procedure to handle simple effect testing was given by Copenhaver and Holland (1988) for the case of balanced sampling and extended by Cheung and Chan (1996) to situations with unbalanced sampling (heterogeneous cell sample sizes).

Appendix B

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

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

DATA;
INPUT IV DV;
CARDS;
   1.00  1.07
   1.00  2.27
         ⋮
   5.00  2.94
   5.00  2.69
;

*** 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;
  MEANS 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 .0917
2 .0003
3 .0001
4 .0001
5 .0422
6 .0015
7 .0001
8 .2314
9 .0118
10 .1760
;

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;
```

Appendix C
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 (the last case in our data set) and below the mean (the first case in our data set) by $\sqrt{\hat{\sigma}^2_W \left[\frac{n-1}{2}\right]}$; where n represents the number of participants before trimming and $\hat{\sigma}^2_W$ is the 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 (referred to in SPSS as "Equal Variances Not Assumed") ***

IV	DV
1.00	-.15
1.00	1.49
1.00	1.49
1.00	1.49
1.00	1.49
1.00	1.49
1.00	1.49
1.00	1.49
1.00	1.49
1.00	1.49
1.00	1.49
1.00	3.13
2.00	.04

```
2.00    1.91
2.00    1.91
2.00    1.91
2.00    1.91
2.00    1.91
2.00    1.91
2.00    1.91
2.00    1.91
2.00    1.91
2.00    1.91
2.00    3.78
3.00    1.63
3.00    2.53
3.00    2.53
3.00    2.53
3.00    2.53
3.00    2.53
3.00    2.53
3.00    2.53
3.00    2.53
3.00    2.53
3.00    2.53
3.00    2.53
3.00    3.44
4.00    .91
4.00    2.93
4.00    2.93
4.00    2.93
4.00    2.93
4.00    2.93
4.00    2.93
4.00    2.93
4.00    2.93
4.00    2.93
4.00    2.93
4.00    4.95
5.00    1.46
5.00    3.15
5.00    3.15
5.00    3.15
5.00    3.15
5.00    3.15
5.00    3.15
5.00    3.15
5.00    3.15
5.00    3.15
5.00    4.84
```

```
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 Benjamini and Hochberg *p* values from the raw *p* values generated previously

```
DATA THREE;
INPUT TEST PVAL;
DATALINES;
1 .18418
2 .00030
3 .00019
4 .00001
5 .02746
6 .00646
7 .00062
8 .16306
9 .01788
10 .50477
;
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;
```

Received April 16, 2003
Accepted March 23, 2004