The Pairwise Multiple Comparison Multiplicity Problem: An Alternative Approach to Familywise/Comparisonwise Type I Error Control

by

H. J. Keselman, Robert Cribbie

University of Manitoba

and

Burt Holland

Temple University

Abstract

When simultaneously undertaking many tests of significance researchers are faced with the problem of how best to control the probability of committing a Type I error. The familywise approach deals directly with multiplicity problems by setting a level of significance for an entire set (family) of related hypotheses, while the comparison approach ignores the multiplicity issue by setting the rate of error on each individual contrast/test/hypothesis. A new formulation of control presented by Benjamini and Hochberg, their *false discovery rate,* does not provide as stringent control as the familywise rate, but concomitant with this relaxing in stringency is an increase in sensitivity to detect effects, compared to familywise control. Type I error and power rates for four relatively powerful and easy-to-compute pairwise multiple comparison procedures were compared to the Benjamni and Hochberg technique for various one-way layouts using test statistics that do not assume variance homogeneity.

The Pairwise Multiple Comparison Multiplicity Problem:

An Alternative Approach to Familywise/Comparisonwise Type I Error Control

The multiplicity problem in statistical inference refers to selecting the statistically significant findings from a large set of findings (tests) to support one's research hypotheses. Selecting the statistically significant findings from a larger pool of results that also contain nonsignificant findings is problematic since when multiple tests of significance are computed the probability that at least one will be significant by chance alone increases with the number of tests examined. Discussions on how to deal with multiplicity of testing have permeated many literatures (e.g., Psychology, Statistics) for decades and continue to this day. In one camp are those who believe that the occurrence of any false positive must be guarded at all costs (see Games, 1971; Miller, 1981; Ryan, 1959, 1960, 1962; Westfall & Young, 1993). That is, as promulgated by Thomas Ryan, pursuing a false lead can result in the waste of much time and expense, and is an error of inference that accordingly should be stringently controlled. Clark (1976) also believes in stringent Type I error control, however, his rationale, differes from that of Ryan. Specifically, Clark (1976, p. 258) believes that because of the prejudice to only report significant results, Type I errors, once made, "are very difficult to correct." Clark (p. 258), citing Bakan (1966), states "the danger to science of the Type I error is much more serious than the Type II error because highly significant results appear definitive and tend to discourage further investigation. ... A stray Type I error can indeed by catastrophic." In this vein, Greenwald (1975, pp. 13-15) cites examples regarding the difficulty of publishing corrections to previously published findings. Regardless of the rationale that have been suggested for stringent Type I error control, those in this camp deal with the multiplicity issue by setting the level of significance (α) for the entire set of tests computed.

For example, in the pairwise multiple comparison problem, Tukey's (1953) multiple comparison procedure (MCP) uses a critical value wherein the probability of

making at least one Type I error in the set of pairwise contrast tests is equal to α. This type of control has been referred to in the literature as the experimentwise or familywise error rate of 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 familywise control. Multiple comparisonists seem to have settled on the familywise label. Thus, in the remainder of the paper, when we speak about overall error control, we are referring to familywise control. As indicated, for the set of pairwise tests, Tukey's procedure sets a familywise rate of error for the family consisting of all pairwise comparisons.

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

It is not the intention of this paper to advocate one position over the other, but rather, to present and examine another conceptualization of error which could be viewed as a compromise position between these two camps, and consequently may be an approach experimenters who are uncertain of which extreme approach to follow may be more comfortable adopting.

The False Discovery Rate

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 (FDR), presented by Benjamini and Hochberg (1995). FDR is defined by these authors as the

expected proportion of the number of erroneous rejections to the total number of rejections. The motivation for such control, as Shaffer (1995) suggests, stems from a common misconception regarding the overall error rate. That is, some believe that the overall rate applied to a family of hypotheses indicates that on average "only a proportion α of the rejected hypotheses are true ones, i.e., are falsely rejected" (Shaffer, 1995, p. 567). This is clearly a misconception, for as Shaffer notes, if all hypotheses are true, "then 100% of rejected hypotheses are true, i.e., are rejected in error, in those situations in which any rejections occur" (p. 567). Such a misconception, however, suggests setting a rate of error for the proportion of rejections which are erroneous, hence the FDR.

We elaborate on the FDR within the context of pairwise comparisons. Suppose we have J (j $= 1, \ldots,$ J) means, $\mu_1, \mu_2, \ldots, \mu_J$, and our interest is in testing the family of m $=$ J(J $-$ 1)/2 pairwise hypotheses, H_i: $\mu_j - \mu_{j'} = 0$, 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. Benjamini and Hochberg (1995) presented a summarization of the relationship between these random variables which we re-present in Table 1. In terms of the random variable V, the per comparison error rate is $\mathcal{E}$(V/m), while the familywise rate is given by P(V $\geq$ 1). Thus, testing each and every comparison at α guarantees that $\mathcal{E}$(V/m) $\leq \alpha$, while testing each and every comparison at α/m (Bonferroni) guarantees P(V $\geq$ 1) $\leq \alpha$.

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

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

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

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

As Benjamini, Hochberg, and Kling (1994) indicate, this error rate has a number of important properties:

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

(b) When $m_0 < m$, the FDR is smaller than or equal to the familywise rate of error. Specifically, if $V \geq 1$ then $V/R \leq 1$, and if $V = 0$ then $V/R = 0$, and thus $P(V \geq 1) \geq \mathcal{E}(Q)$. This indicates that if the familywise rate is controlled for a procedure, than FDR is as well. Moreover, and most importantly for the purposes of this paper, if one adopts a procedure which provides strong (i.e., over all possible mean configurations) FDR control, rather than strong familywise control, then based on the preceding relationship, *a gain in power can be expected*.

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

In addition, to these characteristics, Benjamini et al. (1994) provide a number of illustrations where FDR control seems more reasonable than familywise or per

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

To date, Benjamini and his collaborators (1994, 1995) have developed MCPs providing FDR control for two simultaneous testing problems, for any number of linear contrasts and for all pairwise contrasts. For both adaptations, they provide empirical verification that the FDRs were less than the desired .05 value for all configurations of means investigated and that the probabilities of detecting true differences were greater with an FDR critical value than with a familywise value. Specifically, in the pairwise multiple comparison problem, they have shown that FDR control results in greater sensitivity to detect true pairwise differences as compared to adopting a familywise Bonferroni critical value [i.e., Hochberg's (1988) step-up procedure] to test for pairwise significance.

As promising as the Benjamini et al. (1994) findings are, they may not be strong enough to warrant a switch to FDR control in the pairwise multiple comparison problem.

That is, applied researchers have available to them many MCPs that can be applied to pairwise comparisons other than Bonferroni type procedures investigated by Benjamini et al., and they may provide as much or more power to detect nonnull pairwise differences as the FDR method does. Furthermore, the Benjamini et al. study compared the two procedures when data were obtained from populations having equivalent variability; accordingly, they used Student's two independent sample *t*-test when examining the pairwise comparisons. In the behavioral sciences, however, population variances are rarely equal (see Keselman, Huberty, Lix, Olejnik, Cribbie, Donahue, Kowalchuk, Lowman, Petoskey, Keselman, and Levin, in press; Wilcox, Charlin, & Thompson, 1986). Therefore, the purpose of our investigation was to compare the Benjamini et al. (1994) approach for testing pairwise comparisons with other popular methods for their Type I error and power rates with statistics that do not assume variance homogeneity. The question we will examine is: Can the FDR procedure for examining pairwise comparisons, when coupled with a heteroscedastic statistic, provide acceptable Type I error control when population variances are unequal, and, if so, does FDR control provide sufficiently superior power, compared to relatively powerful familywise MCPs, to warrant a switch to FDR control?

Pairwise MCPs

The Benjamini et al. (1994) procedure for pairwise comparisons, which is based on FDR control, was compared to the methods due to Hochberg (1988), Shaffer (1986), Hayter (1986), and Welsch (1977a, 1977b).[1] These stepwise familywise MCPs were chosen since they have been shown to be somewhat more powerful procedures relative to other simultaneous methods, such as Tukey's HSD (1953), and because they are relatively easy to compute (see Keselman, 1994; Ramsey 1981, 1993; Seaman, Levin, & Serlin, 1991).

Benjamini and Hochberg's (1995) FDR Pairwise Procedure. In this procedure, the p-values corresponding to the m pairwise statistics for testing the hypotheses H_1, ...

, H_m are ordered from smallest to largest, i.e., $p_1 \leq p_2 \leq , \ldots , \leq p_m$, where $m = J(J-1)/2$. Let k be the largest value of i for which $p_i \leq \frac{i}{m}\alpha$, then reject all H_i, $i = 1, 2, \ldots , k$. According to this procedure one begins by assessing the largest p-value, p_m, proceeding to smaller p-values as long as $p_i > \frac{i}{m}\alpha$. Testing stops when $p_k \leq \frac{k}{m}\alpha$.

Hochberg's (1988) Sequentially Acceptive Step-Up Bonferroni Procedure. In this procedure, the p-values corresponding to the m statistics for testing the hypotheses H_1, $\ldots$, H_m are also ordered from smallest to largest. Then, for any $i = m, m-1, \ldots , 1$, if $p_i \leq \alpha/(m-i+1)$, the Hochberg procedure rejects all $H_{i'}$ ($i' \leq i$). According to this procedure, therefore, one begins by assessing the largest p-value, p_m. If $p_m \leq \alpha$, all hypotheses are rejected. If $p_m > \alpha$, then H_m is accepted and one proceeds to compare $p_{(m-1)}$ to $\alpha/2$. If $p_{(m-1)} \leq \alpha/2$, then all H_i ($i = m-1 , \ldots , 1$) are rejected; if not, then $H_{(m-1)}$ is accepted and one proceeds to compare $p_{(m-2)}$ with $\alpha/3$, and so on.

Shaffer's (1986) Sequentially Rejective Bonferroni Procedure that begins with an omnibus test. Like the preceding procedures, the p-values associated with the test statistics are rank ordered. In Shaffer's procedure, however, one begins by comparing the smallest p-value, p_1, to α/m. If $p_1 > \alpha/m$, statistical testing stops and all pairwise contrast hypotheses (H_i, $1 \leq i \leq m$) are retained; on the other hand if $p_1 \leq \alpha/m$, H_1 is rejected and one proceeds to test the remaining hypotheses in a similar step-down fashion by comparing the associated p-values to α/m^*, where m^* is equal to the maximum number of true null hypotheses, given the number of hypotheses rejected at previous steps. Appropriate denominators for each α-stage test for designs containing up to ten treatment levels can be found in Shaffer's Table 2.

Shaffer (1986) proposed a modification to her sequentially rejective Bonferroni procedure which involves beginning this procedure with an omnibus test. [Though MCPs that begin with an omnibus test frequently are presented with the F test, other omnibus tests (e.g., a range statistic) can also be applied to these MCPs; in our investigation we use the omnibus Welch (1951) test.] If the omnibus test is declared nonsignificant,

statistical testing stops and all pairwise differences are declared nonsignificant. On the other hand, if one rejects the omnibus null hypothesis one proceeds to test pairwise contrasts using the sequentially rejective Bonferroni procedure previously described, with the exception that p_m, the smallest p-value, is compared to a significance level which reflects the information conveyed by the rejection of the omnibus null hypothesis. For example, for $m = 6$, rejection of the omnibus null hypothesis implies at least one inequality of means and, therefore, p_6 is compared to $\alpha/3$, rather than $\alpha/6$.

Hayter's (1986) Two-Stage Procedure. This procedure begins with a test (Welch, 1951) of the omnibus hypothesis [H_o: $\mu_1 = \mu_2 = \cdots = \mu_J$], which if rejected, leads to the stage two tests of the pairwise contrasts using a Studentized range critical value for $J - 1$ means. If the omnibus hypothesis is not rejected, then all pairwise hypothese are retained.

Numerical Illustration

We illustrate the mechanics of the Benjamini and Hochberg (1995) MCP for pairwise comparisons as well as its potential for power superiority in comparison to standard methods such as the usual (α/m) and Hochberg (1988) Bonferroni methods of familywise control, with the data set presented by Toothaker (1991, p. 72). Toothaker selected this example from Miller (1981, p. 82) since it provides a good illustration of power differences between pairwise MCPs. The data set is from a $J = 5$ one-way design where the five group means and variances equal $\bar{X}_1 = 16.1$, $\bar{X}_2 = 17.0$, $\bar{X}_3 = 20.7$, $\bar{X}_4 = 21.1$, and $\bar{X}_5 = 26.5$, and $s_1^2 = 7.2044$, $s_2^2 = 7.2004$, $s_3^2 = 7.2023$, $s_4^2 = 7.1943$, and $s_5^2 = 7.1985$, respectively, and sample sizes are all equal to five.[2] Table 2 contains the values of the contrasts, the standard errors of the contrasts, the values of *t* (defined in the next section), the df for the tests, the p-values, and the Bonferroni, Hochberg and FDR critical constants.

Statistical significance according to the standard Bonferroni procedure is determined by comparing the observed p-values to .05/10, while with the Hochberg procedure the critical constant is .05/(11 – i). Thus, according to the Bonferroni

procedures, group 5 differs from both groups 1 and 2 (5-2, 5-1). On the other hand, adopting FDR control, the largest p-value that is less than or equal to its critical constant is p_6; accordingly, one can reject H_6 as well as H_5, H_4, H_3, H_2, and H_1. That is, referring to Table 2, group 5 differs from groups 1, 2, 3, and 4 (5-4, 5-3, 5-2, 5-1), and group 1 differs from groups 3 and 4 (4-1, 3-1). Thus, with FDR control, four additional contrasts are declared significant.

Test Statistics

In our investigation of the MCPs, the omnibus and pairwise tests were computed with Welch's (1938, 1951) nonpooled statistics. We chose to investigate these statistics since, as indicated in the introduction, the data obtained in psychological experiments frequently does not conform to the homogeneity of variance assumption. Furthermore, prior research indicates that the power to detect effects is not substantially reduced when using these statistics compared to the usual omnibus (ANOVA *F* test) and pairwise test (Student's two-sample *t*) statistics when the homogeneity assumption is satisfied (Brown & Forsythe, 1974; Dijkstra & Werter, 1981; Tomarkin & Serlin, 1986; Wilcox et al. 1986). Thus, there is much to recommend in uniformly adopting a nonpooled testing approach to data analysis (see Lix & Keselman, 1995). Therefore, since we also believe that uniform adoption of the nonpooled statistic in the multiple comparison problem is advantageous, we apply it uniformly in our simulated conditions, that is, even when group sizes and variances are equal.

In our work we assume a one-way model where n_j independent random observations X_{1j}, X_{2j} , ... , X_{n_jj} ($i = 1 , \dots , n_j$) were sampled from population j ($j = 1 , ... , J$). Furthermore, we assume that the X_{ij}s are obtained from a population with mean μ_j and unknown variance σ_j^2, with $\sigma_j^2 \neq \sigma_{j'}^2$ ($j \neq j'$). For this model, let $\bar{X}_j = \Sigma_i X_{ij}/n_j$ and $s_j^2 = \Sigma_i(X_{ij} - \bar{X}_j)^2/(n_j - 1)$, where $\bar{X}_j$ is the estimate of μ_j and s_j^2 is the

usual unbiased estimate of the variance for population j. The omnibus test can be expressed as

$$F = \frac{\sum_{j=1}^{J} w_j(\bar{X}_j - \bar{X})^2/(J-1)}{1+\frac{2(J-2)}{(J^2-1)}\sum_{j=1}^{J}\frac{(1-w_j/W)^2}{n_j-1}} \quad (1)$$

where $w_j = n_j/s_j^2$, $\bar{X} = \sum_{j=1}^{J} w_j\bar{X}_j/W$, and $W = \sum_{j=1}^{J} w_j$. The test statistic is approximately distributed as an F variate and is referred to the critical value $F[(1-\alpha); (J-1), \nu]$, the $(1-\alpha)$-centile of the F distribution, where error degrees of freedom (df) are obtained from

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

Numerical results can be obtained from JMP (Sall & Lehman, 1996).

The pairwise tests can also obviously be obtained with the statistic given in Equation (1), but we present the two-sample version since researchers are more likely to be familar with it. The Welch (1938) two-sample statistic is

$$t = \frac{\bar{X}_j - \bar{X}_{j'} - (\mu_j - \mu_{j'})}{\sqrt{\frac{s_j^2}{n_j} + \frac{s_{j'}^2}{n_{j'}}}} \quad , \quad (3)$$

where error df are obtained from

$$\nu = \frac{\left(\frac{s_j^2}{n_j} + \frac{s_{j'}^2}{n_{j'}}\right)^2}{\frac{(s_j^2/n_j)^2}{n_j - 1} + \frac{(s_{j'}^2/n_{j'})^2}{n_{j'} - 1}}. \quad (4)$$

Methods

In the simulation study six variables were manipulated: (a) the number of levels of the independent variable/the number of pairwise tests in the family, (b) the total sample size, (c) the degree of sample size imbalance, (d) the degree of variance inequality, (e) the pairing of group sizes and variances, and (f) the configuration of population means.

Since Type I error control and power to detect nonnull pairwise differences could be related to the number of comparisons in the family of pairwise tests, we chose, like Seaman et al. (1991), to investigate J = 3, 4, and 5 one-way layouts; for these designs there are 3, 6, and 10 pairwise comparisons, respectively.

For each case of J we varied total sample size (N). Specifically, the group sizes (n, when equal) across each value of J were set at either n = 10, n = 15, or n = 19. For the nonnull mean configurations these sample sizes result in an *a priori* theoretical omnibus test (ANOVA F) power of .60, .80, and .90, respectively (assuming equal σ_j^2s) (see Seaman et al., 1991).

We also varied sample size balance/imbalance. According to a recent survey of the educational and psychological literatures for papers published in 1995-6, unbalanced designs are the norm, not the exception (Keselman, Huberty, Lix, Olejnik, Cribbie, Donahue, Kowalchuk, Lowman, Petoskey, Keselman, and Levin (in press). Furthermore, since the effects of variance heterogeneity are exacerbated by sample size imbalance, we included three cases of balance/imbalance for each layout investigated. In particular, sample sizes were either equal, moderately unequal, or very unequal, where the degree of

balance/imbalance was quantified with a coefficient of sample size variation (SCV); SCV is defined as $(\Sigma_j(n_j - \overline{n})^2/J)^{\frac{1}{2}}/\overline{n}$, where $\overline{n}$ is the average group size. When sample sizes were equal SCV = 0; the moderately unequal cases had values of SCV $\simeq$ .10, while SCV $\simeq$ .40 for the largest case of imbalance investigated. Keselman et al. report that SCV $\simeq$ .40 values, or greater, are common. Sample sizes and values of SCV are enumerated in Table 3 for each case of J.

We also considered variance equality/inequality, having again three cases which were also quantified by a coefficient of (variance) variation (VCV). Across the three layouts, VCV = 0 when the variances were equal, and VCV $\simeq$ .50 and VCV $\simeq$ .60, respectively, for our cases of moderate and more than moderate inequality. The degrees of variance heterogeneity were based on results reported by Keselman et al. (in press). According, to their survey, it is not uncommon in behavioral science investigations for unequal variances to have values of VCV $\simeq$ .50. Our largest case of heterogeneity represents a worst case scenario, though other investigators believe such inequality does occur with some frequency (see Wilcox, 1994). Variances and values of VCV are enumerated in Table 4 for each case of J.

When variances were unequal, they were both positively and negatively paired with the group sizes. For positive (negative) pairings, the group having the fewest (greatest) number of observations was associated with the population having the smallest variance, while the group having the greatest (fewest) number of observations was associated with the population having the largest variance. These conditions were chosen since they typically produce conservative and liberal results, respectively.

The last variable manipulated was the configuration of means. Here also we used the values investigated by Seaman et al. (1991) (their Table 3). Two definitions of power were used in this investigation: (a) the average per-pair power rate, where per-pair power is the probability of detecting a true pairwise difference, and (b) the all-pairs power rate,

or the probability of detecting all true pairwise differences. The any-pairs power rate, or the probability of detecting at least one true pairwise difference was not collected since we agree with Ramsey (1978) and Keselman (1994) in that if determining *any* true difference is of importance, as would be the case in say exploratory research, then this is not really an area where a MCP would be best; an omnibus test and its overall power is likely to be more suited to this type of research endeavor.

To generate pseudo-random normal variates, we used the SAS generator RANNOR (SAS Institute, 1989). If Z_{ij} is a standard normal variate, then $X_{ij} = \mu_j + \sigma_j \times Z_{ij}$ is a normal variate with mean equal to μ_j and variance equal to σ_j^2. Our simulation program was written in SAS/IML (SAS, 1989) code. Five thousand or more replications of each condition were performed using a .05 criterion for each of the procedures investigated.[3]

Results

Our initial analysis of the data examined whether any of the MCPs did not consistently maintain their rates of Type I error close to the nominal 5% value. That is, since the focus of our study was to compare the power of the MCPs, we did not want to include in the power comparison phase of the investigation any procedure whose rates of error were inflated. Not only do these inflated rates indicate the exact conditions of our simulation in which a procedure breaks down but, as well, are suggestive of the vulnerability of the procedure to other degrees of assumption violations not investigated but which possibly could be encountered by psychological researchers. This preliminary analysis indicated, as previously noted, that only Welsch's (1977a) procedure could not consistently maintain Type I error control; on occasion, its rates of Type I error substantially exceeded 5% (e.g., 12.59%). Accordingly, the Welsch procedure was not included in the power phase of our investigation.

Type I error rates. The pattern of Type I error rates for the remaining MCPs were generally well controlled and consistent across the cases of J and N and therefore we only present the rates for J = 5 (collapsing over N). In particular, Figures 1 and 2 contain empirical percentages of Type I error for the complete null (one case) and partial null cases (an average rate over the 10 partial nulls), respectively. In the remainder of the paper, including figures and tables, we use the following abbreviations when referring to the four MCPs: (a) Benjamini and Hochberg (1995)-BH, (b) Hochberg (1988)-H, (c) Hayter (1986)-HR, and (d) Shaffer (1986)-S.

As indicated, the points plotted in Figures 1 and 2 show that the MCPs can maintain their rates of error close to the nominal 5% value over the cases of variance heterogeneity and sample size imbalance investigated.

Power rates[4]. Since rates of Type I error were well controlled in the complete and partial null cases across the values of J, we enumerate average per-pair power rates for the MCPs in Table 5. The all-pairs power rates were similar in pattern to the average per-pair rates and therefore are not tabled. Since experimenters are not likely to know the population state of affairs with regard to degree of heterogeneity, effect size and mean configuration, deriving recommendations based upon specific combinations of these factors is of little use. Accordingly, Table 5 contain power rates which have been averaged over the nonnull mean configurations and cases of VCV and SCV. The most powerful MCP in each of the investigated conditions, enumerated in Table 5, is indicated with underlining.

The average per-pair rates suggest that the MCPs have similar sensitivity to detect pairwise mean differences. Nonetheless, it is also clear that the relative advantage of the most powerful procedure to the others is related to design size, that is, to the number of pairwise comparisons in the family of computed tests. Specifically, for the smallest of the designs investigated, the S/HR MCPs are uniformly most powerful. However, as the

number of tests in the family of pairwise comparisons increased, that is, as the design size increased, the BH procedure generally became the most powerful MCP. That is, for the J = 4 and J = 5 designs, BH was most powerful in three and eight of the nine conditions investigated, respectively. The reader should also note that though power differences are not large, they are average per-pair differences, and as Seaman et al. (1991) note, “When the number of comparisons being studied becomes large, the probability increases that a given comparison will be declared significant by one procedure but not by another." (p. 583)

Discussion

Four stepwise MCPs (Hochberg, 1988; Hayter, 1986; Shaffer, 1986; Welsch, 1977a, 1977b), which provide familywise Type I error control for a set of pairwise comparisons, were compared to the Benjamini and Hochberg (1995) procedure, a procedure which sets neither familywise nor comparisonwise control, but rather sets a false discovery rate of error, for their rates of Type I error and power to detect nonnull pairwise differences. Nonpooled Welch (1938, 1951) statistics were used with the MCPs in order to determine their operating characteristics in balanced/imbalanced one-way layouts when population variances were either equal or unequal.

Results indicate that, with the exception of Welsch's (1977a) procedure, all MCPs were able to maintain reasonably accurate control over Type I errors over all conditions investigated. Though the Benjamini and Hochberg (1995) procedure was uniformly more powerful than Hochberg's (1988) Bonferroni procedure (an expected finding given the results reported by Benjamini et al.), it was not consistently the most powerful of the procedures investigated. That is, when the number of tests in the family of pairwise comparisons was small, that is six or less, Hayter's (1986) two-stage procedure was typically most powerful. However, when the number of pairwise tests exceeded six, the Benjamini and Hochberg (1995) procedure was typically most powerful.

Our results suggest that as the number of pairwise comparisons in the family of tests increases, the Benjamini and Hochberg (1995) FDR approach loses less power in comparison to other pairwise MCPs based on familywise control, and thus it is a relatively more powerful approach to examining pairwise comparisons than its competitors. That is, it appears that compared to familywise controlling MCPs, the FDR approach suffers less of a power loss as the number of tests in the family increases, a desirable feature for those who decry the insensitivity of familywise controlling procedures. To verify this finding we again compared the power rates (average per-pair and all-pairs) of the MCPs, however, we increased the family size of the pairwise comparison tests to 28 by increasing the design size to $J = 8^5$. The average per-pair rates, averaging over 12 nonnull mean configurations, for the BH and HR MCPs were 8.5% and 7.2%, respectively, when $n = 10$. When $n = 15$ and 19, the rates were 18.2% and 12.5% and 26.6% and 18.3%, respectively. The relative differences [(BH − HR)/HR] for the preceding are .18, .46 and .45, respectively. The all-pairs rates followed a similar pattern. It is also important to note that for each of the configurations investigated, the BH average per-pair power value was larger than the HR value and on occassion was greater by at least .14. Thus, as predicted, FDR always had greater sensitivity to detect nonnull pairwise differences.

Based on our findings and those reported in the literature we make the following recommendations for testing pairwise comparisons: for $J = 3$, Fisher (1935); for $J = 4$, Hayter (1986); for $J \geq 5$, Benjamini and Hochberg (1995).

Footnotes

1. The Welsch (1977a) procedure on occasion resulted in very liberal rates of Type I error (e.g., >12%) and therefore was not included in the power phase of our study; consequently, to save space, we do not define the procedure here (See Welsch or Keselman, 1994 for a description of the procedure.)

2. The numerical example is intended to demonstrate the use of the FDR MCP and therefore we were not concerned with the values of the sample sizes and variances, though in our computations we still use the Welch (1938) nonpooled statistic.

3. In order to ensure accurate estimation of the FDR, we required that there be 5000 simulations in which $R > 0$ (Remember $\mathcal{E}(Q)$ is defined to be zero when $R = 0$); accordingly, the familywise MCPs' rates were based on considerably more than 5000 simulations.

4. For $J = 3$, S and HR are equivalent and are in fact equivalent to Fisher's (1935) Two Stage Least Significant Difference MCP. When $J = 3$ Fisher's test provides accurate familywise control when assumptions are satisfied (see Levin, Serlin & Seaman, 1994).

5. For $J = 8$, after confirming Type I error control, we investigated 12 patterns of nonull configurations that were similar to those reported by Seaman et al. (1991). In addition, for each pattern, sample size per group was again set at 10, 15, and 19; additionally, we only investigated the equal n_j/equal σ_j^2 condition.

References

Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, 66, 423-437.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. Journal of the Royal statistical Society B, 57, 289-300.

Benjamini, Y., Hochberg, Y., & Kling, Y. (1994). False discovery rate controlling procedures for pairwise comparisons. Unpublished manuscript.

Brown, M.B., & Forsythe, A.B. (1974). The small sample behavior of some statistics which test the equality of several means. Technometrics, 16, 129-132.

Clark, H. H. (1976). Reply to Wike and Church. Journal of Verbal Learning and Verbal Behavior, 15, 257-261.

Dijkstra, J.B., & Werter, P.S.P.J. (1981). Testing the equality of several means when the population variances are unequal. Communications in Statistics, Simulation and Computation, B10, 557-569.

Fisher, R.A. (1935). The design of experiments. Edinburgh, Scotland: Oliver & Boyd.

Games, P.A. (1971). Multiple comparisons of means. American Educational Research Journal, 8, 531-565.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

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.

Keselman, H.J. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures' means. Journal of Educational Statistics, 19, 127-162.

Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., & Levin, J.R. (in press). Statistical practises of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. Review of Educational Research.

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.

Miller, R.G., Jr. (1981). Simultaneous statistical inference (2nd ed.). New York: Springer-Verlag.

Norusis, M.J. (1993). SPSS for windows: Advanced Statistics Release 6.0. Chicago, Illinois, SPSS Inc.

Ramsey, P. H. (1981). Power of univariate pairwise multiple comparison procedures. Psychological Bulletin, 90, 352-366.

Ramsey, P. H. (1993). Multiple comparisons of independent means. In L. Edwards (Ed.) Applied analysis of variance in the behavioral sciences, 25-61. New York: Marcel Dekker.

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

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

Ryan, T.A. (1960). Significance tests for multiple comparison of proportions, variances and other statistics. Psychological Bulletin, 57, 318-328.

Ryan, T.A. (1962). The experiment as the unit for computing rates of error. Psychological Bulletin, 59, 301-305.

Sall, J., & Lehman, A. (1996). JMP Start Statistics: A Guide to Statistical and Data Analysis Using JMP and JMP IN Software. Cary, NC: SAS Institute.

SAS Institute Inc. (1985). SAS User's Guide: Basics Version 5 Edition. Cary, NC: SAS Institute.

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

Seaman, M.A., Levin, J.R., & Serlin, R.C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. Psychological Bulletin, 110, 577-586.

Shaffer, J.P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.

Shaffer, J.P. (1995). Multiple hypothesis testing: A review. Annual Review of Psychology, 46, 561-584.

Tomarken, A.J., & Serlin, R.C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychological Bulletin, 99, 90-99.

Toothaker, L.E. (1991). Multiple comparisons for researchers. Newbury Park: Sage.

Tukey, J.W. (1953). The problem of multiple comparisons. In H.I. Braun (Ed.) The collected works of John W. Tukey, Volume VIII Multiple comparisons: 1948-1983, 1-300. New York: Chapman & Hall.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika, 29, 350-362.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. Biometrika, 38, 330-336.

Welsch, R.E. (1977a). Stepwise multiple comparison procedures. Journal of the American Statistical Association, 72, 566-575.

Welsch, R.E. (1977b). Tables for stepwise multiple comparison procedures (Working paper No. 949-77). Cambridge: Massachusetts Institute of Technology.

Westfall, P.H. & Young, S.S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustments. New York: Wiley & Sons.

Wilcox, R.R. (1994). A one-way random effects model for trimmed means. Psychometrika, 59, 289-306.

Wilcox, R.R., Charlin, V.L., & Thompson, K.L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and F^* statistics. Communications in Statistics, Simulation and Computation, 15, 933-943.

Wilson, W. (1962). A note on the inconsistency inherent in the necessity to perform multiple comparisons. Psychological Bulletin, 59, 296-300.

Figure Captions

Figure 1. J = 5 Empirical Type I Error Rates (Complete Null Hypothesis)

Figure 2. J = 5 Empirical Type I Error Rates (Averaged Over All Partial Null Hypotheses)

Author Note

This research was supported by a Social Sciences and Humanities Research Council grant (#410-95-0006). Correspondence concerning this article should be addressed to H. J. Keselman, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2.

Table 1. Number of Errors Committed when Testing m Null Hypotheses

	Declared Non-Significant	Declared Significant	Total
True Null Hypotheses	U	V	m_0
Non-true Null Hypotheses	T	S	$m - m_0$
	$m - R$	R	m

Note: From “Controlling the False Discrovery Rate: A Practical and Powerful Approach to Multiple Testing,” by Y. Benjamini and Y. Hochberg, (1995), Journal of the Royal Statistical Society, Series B, 57, p. 289-300. Copyright 1995 by The Royal Statistical Society. Adapted with permission. R is an observable random variable, while U, V, S, and T are unobservable random variables.

Table 2. Critical constants for Various Multiple Comparison Procedures

Contrast	1 vs 5	2 vs 5	3 vs 5	4 vs 5	1 vs 4	1 vs 3	2 vs 4	2 vs 3	1 vs 2	3 vs 4
Value	10.4	9.5	5.8	5.4	5.0	4.6	4.1	3.7	0.9	0.4
S. Error	1.70	1.70	1.70	1.70	1.70	1.70	1.70	1.70	1.70	1.70
t	6.13	5.60	3.42	3.18	2.95	2.71	2.42	2.18	0.53	0.24
df	8	8	8	8	8	8	8	8	8	8
p-value	.000	.001	.009	.013	.019	.027	.042	.061	.610	.820
index i	1	2	3	4	5	6	7	8	9	10
B	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005
H	.005	.0056	.0063	.0071	.0081	.0100	.0125	.0167	.0250	.0500
FDR	.0050	.0100	.0150	.0200	.0250	.0300	.0350	.0400	.0450	.0500

Note: B stands for Bonferroni, H stands for Hochberg's (1988) step-up Bonferroni procedure, and FDR stands for Benjamini and Hochberg's (1994) False Discovery Rate of Type I Error.

Table 3. Sample Sizes and Coefficients of Sample Size Variation (SCV)

J	N	Sample Sizes	SCV
3	30	10, 10, 10	0
		9, 10, 11	.082
		5, 10, 15	.408
	45	15, 15, 15	0
		13, 15, 17	.109
		7, 15, 23	.436
	57	19, 19, 19	0
		17, 19, 21	.086
		9, 19, 29	.430
4	40	10, 10, 10, 10	0
		9, 10, 10, 11	.071
		5, 7, 13, 15	.412
	60	15, 15, 15, 15	0
		13, 15, 15, 17	.094
		7, 11, 19, 23	.422
	76	19, 19, 19, 19	0
		17, 19, 19, 21	.074
		9, 14, 24, 29	.416
5	50	10, 10, 10, 10, 10	0

		9, 10, 10, 10, 11	.063
		5, 6, 10, 14, 15	.405
	75	15, 15, 15, 15, 15	0
		13, 14, 15, 16, 17	.094
		7, 9, 15, 21, 23	.422
	95	19, 19, 19, 19, 19	0
		17, 18, 19, 20, 21	.074
		9, 11, 19, 27, 29	.426

Table 4. Variances and Coefficients of Variance Variation (VCV)

J	Population Variances	VCV
3	1, 1, 1	0
	1, 2, 4	.535
	1, 4, 8	.662
4	1, 1, 1, 1	0
	1, 2, 4, 4	.472
	1, 3, 5, 8	.608
5	1, 1, 1, 1, 1	0
	1, 1, 2, 3, 4	.530
	1, 2, 4, 6, 8	.610

Table 5. Per-Pair Power Percentages Averaged Across Nonnull Configurations

J	N	Condition	BH	HG	SR	HR
3	30	$=n_j$ or $=\Sigma_j$	23.9	22.0	26.5	26.5
		PP	11.5	10.4	12.8	12.8
		NP	11.2	10.3	13.4	13.4
	45	$=n_j$ or $=\Sigma_j$	36.5	34.3	39.8	39.8
		PP	18.3	16.7	20.2	20.2
		NP	17.5	16.3	21.2	21.2
	57	$=n_j$ or $=\Sigma_j$	44.9	42.8	48.0	48.0
		PP	23.8	22.0	26.0	26.0
		NP	23.0	22.0	27.6	27.6
4	40	$=n_j$ or $=\Sigma_j$	16.7	13.1	15.8	16.9
		PP	6.7	5.1	6.4	7.0
		NP	6.7	5.4	6.9	7.6
	60	$=n_j$ or $=\Sigma_j$	28.0	22.8	26.2	27.6
		PP	11.5	8.9	10.9	11.6
		NP	11.0	8.8	11.3	12.3
	76	$=n_j$ or $=\Sigma_j$	36.4	30.5	33.8	35.3
		PP	15.6	12.3	14.7	15.6
		NP	15.0	12.2	15.4	16.5
5	50	$=n_j$ or $=\Sigma_j$	12.0	8.1	9.6	11.1
		PP	4.8	3.3	4.0	4.6
		NP	5.4	3.8	4.6	5.5

	75	$=n_j$ or $=\Sigma_j$	21.4	15.0	17.1	18.9
		PP	8.9	5.9	7.0	8.0
		NP	10.0	6.5	7.8	9.0
	95	$=n_j$ or $=\Sigma_j$	28.7	20.1	23.1	25.0
		PP	12.4	8.4	9.9	11.0
		NP	12.6	9.1	10.8	12.1

Note: $=n_j$ or $=\Sigma_j$: Equal sample sizes or equal variances; PP(NP): Positive (Negative)pairings of sample sizes and variances. Underlined values indicate the MCP with the largest empirical power value.