Taylor & Francis
Taylor & Francis Group

Pairwise Multiple Comparisons: New Yardstick, New Results
Author(s): Robert A. Cribbie

Source: *The Journal of Experimental Education*, Vol. 71, No. 3 (Spring, 2003), pp. 251-265
Published by: Taylor & Francis, Ltd.
Stable URL: http://www.jstor.org/stable/20152711
Accessed: 07-06-2017 14:22 UTC

The Journal of Experimental Education, 2003, 71(3), 251–265

Pairwise Multiple Comparisons: New Yardstick, New Results

ROBERT A. CRIBBIE
York University (Canada)

ABSTRACT. Behavioral science researchers often wish to compare the means of several treatment conditions on a specific dependent measure. The author used a Monte Carlo study to compare familywise error controlling multiple comparison procedures (MCPs; e.g., Tukey, Bonferroni) with MCPs that were not developed to control the familywise error rate on the probability of correctly identifying the true underlying population mean configuration (true model rate). Recently proposed MCPs that are not intended to control the familywise error rate had consistently larger true model rates than did familywise error controlling MCPs. Furthermore, of the familywise error controlling MCPs investigated, the popular Tukey and Bonferroni MCPs had consistently lower true model rates than did other familywise error controlling MCPs.

Key Words: false discovery rate, model testing, multiple comparison procedures, true model rate

A RESEARCHER CONDUCTING A STUDY with more than two groups is often interested in determining if there are statistically significant mean differences among the groups on a dependent measure. For example, a researcher may wish to determine if overall course ratings differ for lecture, seminar, or computer-mediated classroom formats. In that type of experiment, researchers often want to know if there are any statistically significant differences between any pair of formats. For example, do the ratings of students in lecture-format classes differ from the ratings of students in seminar-format classes? Hypotheses concerning differences between two group means are referred to as *pairwise comparisons*.

Researchers testing pairwise null hypotheses are faced with important decisions regarding error control, including the selection of a significance level (α), a unit of analysis, and a specific multiple comparison procedure (MCP). First, α specifies the probability of rejecting a true null hypothesis (i.e., 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 and the probability of rejecting a false null hypothesis (i.e., power) decreases (assuming that all other factors, e.g., sample size, are held constant). 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, such as $\alpha = .05$ (see e.g., Shaffer, 1995). When multiple hypotheses are being tested, researchers must specify not only α but also the unit of analysis over which Type I error control will be applied and the MCP that will be used.

Unit of Analysis

Several different units of analysis (i.e., error rates) have been proposed in the multiple comparison literature. Although the majority of the literature has focused on the comparisonwise and familywise error rates (e.g., Ryan, 1959; Toothaker, 1991; Tukey, 1953), other error rates, such as the false discovery rate (Benjamini & Hochberg, 1995), have also been proposed.

Comparisonwise Error Rate

The comparisonwise error rate (α_C) is defined as the expected proportion of falsely rejected null hypotheses. In conducting all pairwise comparisons in a one-way completely randomized design, the comparisonwise error rate can be defined as

$$\alpha_C = \alpha = P\text{ (Reject H}_c \mid \text{H}_c \text{ is true)},$$

where H_c represents the $c = 1, \ldots, C$, pairwise null hypotheses (H_c: $\mu_j = \mu_{j'}, j \neq j'$). In a review of the *Bulletin of the Psychonomic Society,* Gaito and Nobrega (1981) found that in almost half (49.1%) of the studies in which multiple comparisons were performed α_C control was applied, although the authors of a more recent evaluation of the educational and psychological literature found that only about 10% of researchers performing pairwise comparisons applied α_C control (Keselman et al., 1998). A number of authors in the MCP field have recommended α_C control over other proposed error rates (e.g., Carmer & Walker, 1985; Davis & Gaito, 1984; Rothman, 1990; Saville, 1990; Wilson, 1962), providing a few simple, but convincing, arguments.

Saville (1990) argued that the natural unit of analysis is the comparison. Researchers conducting pairwise comparisons are often interested in which treat-

Address correspondence to Robert A. Cribbie, Department of Psychology, York University, Toronto, Ontario M3J 1P3, Canada. E-mail: cribbie@yorku.ca

ment groups differ, not in the overall pattern of treatment group differences. Thus, the objectives of the research are the same if C pairwise comparisons are tested in one experiment or if one comparison is tested in each of C experiments. Therefore, it does not make sense to control α over the entire experiment. As Saville (1990) stated, "an experiment is no more a natural unit than a project consisting of several experiments or a research program consisting of several projects" (p. 177).

A second argument in favor of α_C control is that real differences between treatment groups are more likely with a greater number of treatment groups (Duncan, 1955; Hancock & Klockars, 1996). In real experiments the likelihood that all treatment groups are unequal is greater than the likelihood that all treatment groups are equal (Saville, 1990). Therefore, emphasis in experiments should be not on controlling for unlikely Type I errors but in obtaining the most power for detecting even small differences between treatments.

Furthermore, suppose Experimenter 1 is testing for differences between two treatments, A and B, and finds them to be significantly different. Experimenter 2, also testing for differences between treatments A and B, decides to explore differences between the first two treatments and treatment C as well. Invoking control of α over all pairwise comparisons, Experimenter 2 finds the same treatment effect for the difference between treatments A and B as does Experimenter 1 but reports no statistically significant difference between those treatments. Experimenter 2, restricting the maximum rate of Type I errors at α over all three comparisons, has less power for detecting individual treatment effects and consequently finds results inconsistent with those of Experimenter 1. That inconsistency is widely recognized by multiple comparison researchers and was highlighted by Wilson (1962) and Saville (1990). Saville declared that a procedure is inconsistent if the probability of finding two populations significantly different is affected by the number of populations or by the values of the sample means from the other populations. To avoid inconsistency between experiments researchers are advised to use consistent procedures in their analyses (i.e., α_C control), and inconsistent results in the literature can easily be discovered and discarded through replication (Davis & Gaito, 1984).

Last, one of the primary advantages of α_C control is convenience. One evaluates each of the C pairwise treatment effects by using any appropriate test statistic and comparing that statistic with an α-level critical value. Hancock and Klockars (1996) summarized the convenience of α_C control by stating that "if this perspective were unilaterally adopted, virtually all multiple comparisons would be easily conducted with t-tests using liberal critical values, and the MCP researcher would be unemployed" (p. 272).

However, comparisonwise α control is not without drawbacks. The primary disadvantage of α_C control with pairwise comparisons is that the probability of making at least one Type I error increases with the number of comparisons, approaching $1-(1-\alpha)^C$ where$1-(1-\alpha)^C$ would be the probability of making at least one Type I error assuming all comparisons were independent. Therefore, the more treatment groups that a researcher includes in his or her experiment the more likely it is that one or more comparisons will be statistically significant simply by chance.

Familywise Error Rate

The familywise error rate (α_F) is defined as the probability of falsely rejecting one or more hypotheses in a family of hypotheses. Many researchers have recommended controlling α_F (e.g., Hancock & Klockars, 1996; Petrinovich & Hardyck, 1969; Ryan, 1959, 1962; Tukey, 1953), which is "the most commonly endorsed approach to accomplishing Type I error control" (Seaman, Levin, & Serlin, 1991, p. 577). Keselman et al. (1998) reported that approximately 85% of researchers conducting pairwise comparisons adopt some form of α_F control.

The main advantage of α_F control is that the probability of making a Type I error does not increase with the number of comparisons conducted in the experiment. In the previous example used to demonstrate the consistency of α_C control, I also demonstrated the advantage of α_F control. If Experimenter 1 conducts one comparison at $\alpha = .05$ and Experimenter 2 conducts three comparisons, each at $\alpha = .05$, Experimenter 1 has a 5% chance of making at least one Type I error, whereas Experimenter 2 has almost a 15% chance of making a Type I error. The increased risk of Type I errors with α_C control affects both the conclusions of individual experiments as well as the theories based on the results of those experiments. As Petrinovich and Hardyck (1969) proclaimed, "it is better to punish truth than to let falsehood gain respectability" (p. 52). Journal editors in the behavioral sciences have adopted this principle almost exclusively to minimize the number of errors in the published literature (Petrinovich & Hardyck, 1969; Sato, 1996).

One of the main disadvantages of procedures that control α_F is that α_C decreases, often substantially, as the number of treatment groups increases. Therefore, MCPs that control α_F have reduced power for detecting treatment effects when there are many comparisons, increasing the potential for inconsistent results between experiments (Miller, 1981). As Holland and Cheung (2002) contend, "procedures that control [α_F] lose their attractiveness in large data environments with a wide range of possible family sizes" (p. 2)

False Discovery Rate (FDR)

Benjamini and Hochberg (1995) presented a compromise between strict α_F control and liberal α_C control, namely, the false discovery rate. The FDR is defined as the expected ratio (Q) of the number of erroneous rejections (V) to the total number of rejections ($R = V + S$), where S represents the number of true rejections (Benjamini, Hochberg, & Kling, 1994). Therefore, $E(Q) = E(V/[V + S]) = E(V/R)$.

The relationship between FDR control and other error rates was summarized by Benjamini et al. (1994). If all null hypotheses are true, FDR = α_F. On the other hand, if a subset of the null hypotheses is false FDR $\leq \alpha_F$, resulting in an increase in power relative to α_F control. The basis for FDR control is that as the number of comparisons increases, the expected number of false null hypotheses also increases, and thus a procedure that offers greater power as the number of comparisons increases is justified.

MCPs

With a pairwise MCP, one can test hypotheses either simultaneously, sequentially, or through model selection. In a simultaneous MCP, all comparisons are conducted regardless of whether an omnibus test, or any other comparison, is statistically significant. Examples of simultaneous pairwise MCPs include the comparisonwise error control approach, Tukey's (1953) MCP, and Bonferroni's (1937) MCP.

Comparisonwise Error Control (CEC). A simple method for conducting all pairwise comparisons is to reject Hc if $tc \geq t(\alpha, \nu)$, where tc represents the value from an appropriate two-sample test statistic (e.g., two independent-samples *t* test [Welch, 1938]; nonpooled *t* test) and ν represents the error degrees of freedom. The CEC procedure controls $\alpha C = \alpha$, but αF can greatly exceed α as the number of comparisons increases. Note that this procedure differs from Fisher's least significant difference procedure in that no omnibus test is used.

Tukey. In Tukey's honestly significant difference (HSD) MCP, a critical value obtained from the Studentized Range (q) distribution is used, and control of α_F can be obtained. Tukey rejects H_c if $t_c \geq q(\alpha, J, \nu)/(2)^{1/2}$. With unequal sample sizes, the Tukey–Kramer modified statistic is adopted.

Bonferroni. With the Bonferroni (1937) procedure, which also controls α_F, H_c is rejected if $t_c \geq t(\alpha_{pc}, \nu)$ where $\alpha_{pc} = \alpha/C$.

In a sequential MCP, either the significance of an omnibus test or the significance of other comparisons (or both) is considered in evaluating the significance of a particular comparison. Examples of sequential MCPs include Hayter's (1986) MCP, the REGWQ MCP (Einot & Gabriel, 1975; Ryan, 1960; Welsch, 1977) and the FDR MCP (Benjamini & Hochberg, 1995).

Hayter. Hayter (1986) proposed a modification to the popular Fisher (1935) least significant difference procedure that provides consistent control over α_F. H_c is rejected if the omnibus test is statistically significant and $t_c \geq q(\alpha, J\text{–}1, \nu)/(2)^{1/2}$.

REGWQ. Ryan (1960), Einot and Gabriel (1975), and Welsch (1977) proposed a sequential MCP that controls α_F at α. The REGWQ procedure sequentially tests all ordered mean differences for stretch sizes (inclusive range of the ranks of the means) p = J, J–1, . . . , 2, and rejects H_c if

$$t_c \geq q\ (\alpha_p, p, \nu)/(2)^{1/2},$$

where $\alpha_p = \alpha$ for p = J, J–1, and $\alpha_p = 1 - (1 - \alpha)^{p/J}$, for p = J–2, . . ., 2. If any H_cs are retained for p = p', then all H_cs contained in that stretch are retained and are not tested at later stages (i.e., p < p'). If all H_cs are retained for p = p', then all H_cs with p < p' are retained.

FDR. Benjamini and Hochberg (1995) proposed controlling the FDR, as a compromise between strict α_F and liberal α_C control. With the FDR MCP, the pairwise test statistics are first ordered from largest to smallest, and testing proceeds from the largest to smallest test statistic (c = C, C –1, . . ., 1). H_c is rejected if $t_c \geq t\ (\alpha_c, \nu)$, where $\alpha_c = \alpha$ (c/C). In addition, if H_c is rejected, then all $H_{c'}$ (c' ≤ c) are also rejected.

In model testing procedures, possible population mean configurations are compared and the configuration (model) that best accounts for differences among the populations is selected. For example, Dayton (1998) recently introduced a protected model testing procedure that provides a very logical theory for testing pairwise multiple comparisons and that can be applied with unequal variances, nonnormal data, or both.

MTP. Dayton (1998) proposed a model testing procedure (MTP) based on Akaike's information criteria (AIC). One evaluates mutually exclusive and transitive models by using AIC (assuming that an appropriate omnibus null hypothesis has been rejected) and retains the model that has the minimum AIC as the most probable population mean configuration, where

$$\text{AIC} = SS_e + \sum_{j=1}^{J} n_j\ (\bar{x}_j - \bar{x}_{mj})^2 + 2q$$

SS_e is the error sums of squares from an omnibus analysis of variance, ($\bar{x}_j$ is the estimated sample mean for the jth group, $\bar{x}_{mj}$ is the estimated sample mean for the jth group given the hypothesized population mean configuration for the mth model, and q is the number of independent parameters evaluated in the model. A protected version of the MTP procedure (where the MTP procedure is conducted only if an omnibus test of mean equality is statistically significant, and the complete null mean configuration is retained if the omnibus test is nonsignificant) has been recommended by Dayton to improve the accuracy of the procedure when all population means are equal, and a slightly modified version of the AIC statistic is adopted with heterogeneous variances (see Dayton). To illustrate the procedure, with $J = 4$ (and ordered means) there would be $2^{J-1} = 8$ mutually exclusive and transitive models to be evaluated ([1234], [1, 234], [12, 34], [123, 4], [1, 2, 34], [12, 3, 4], [1, 23, 4], [1, 2, 3, 4]). To illustrate, the model [12, 3, 4] postulates a population mean configuration where Groups 1 and 2 are derived from the same population, but Groups 3 and 4 each represent independent populations. The model having the lowest AIC value would be retained as the most probable population model.

Comparing Error Rates and MCPs

The most common approach for comparing the performance of pairwise MCPs is through Monte Carlo studies that evaluate the Type I error control and power of the procedures under various testing conditions. More specifically, the approach commonly adopted by multiple comparison researchers in recommending a specific MCP is to use a two-step approach: a) determine if the MCP provides consistent control of the familywise error rate; and b) assuming the familywise error rate is controlled at approximately α, recommend the MCP with the most power. Power is often defined with respect to *per-pair power* (the average probability of rejecting a false H_c) or *all-pairs power* (the average probability of rejecting all false H_cs). Utilizing this approach results in researchers adopting only procedures that control the familywise error rate, including popular procedures such as the Bonferroni and Tukey MCPs. On the other hand, procedures such as the CEC and the FDR are not recommended because they do not control α_F at α. However, a serious limitation of this approach is that recommendations concerning MCPs are made while considering the Type I error control and the power of the procedures separately instead of simultaneously. However, an innovative approach to comparing pairwise MCPs is the comparison of true model rates (Cribbie & Keselman, in press), which can be defined as the average probability of identifying the correct underlying population mean configuration (typically the goal of researchers using MCPs). With that approach, recommendations concerning MCPs are made while simultaneously considering the Type I error control and power of the procedures. In other words, to identify the true underlying population mean configuration, one should use a procedure that does not reject any true pairwise H_cs and rejects all false pairwise H_cs.

Therefore, my goal in this research was to compare the true model rates of popular familywise error controlling MCPs such as the Tukey and Bonferroni procedures with that of less popular (or recently proposed) MCPs such as the CEC, Hayter, REGWQ, FDR, and MTP, which might potentially provide a better chance of identifying the true underlying population mean configuration.

Method

I used a simulation study to compare the true model rates of the CEC, Tukey, Bonferroni, Hayter, REGWQ, FDR and protected MTP procedures. For the MTP procedure, the true model rate is defined as the probability of selecting the correct model with the AIC statistic, and for the remaining MCPs, the true model rate is defined as the probability of rejecting all false pairwise null hypotheses and not rejecting any true pairwise null hypotheses. In addition to manipulating the MCPs, the following six variables were manipulated in this study: (a) number of levels of the independent variable ($J = 4$ and $J = 7$); (b) total sample size (average $n_j = 10$, 15, and 25); (c) degree of sample size imbalance (equal n_j, moderately unequal n_j, and extremely unequal n_j; see Table 1 for the specific sample size conditions); (d) degree of variance inequality (equal variances, largest to smallest variance ratio of 4:1, and largest to smallest variance ratio of 8:1; see Table 1 for the specific variance conditions); (e) pairings of unequal group sizes and variances (positively paired, where the largest sample size is paired with the largest variance and smallest sample size with the smallest variance, and negatively paired, where the largest sample size is paired with the smallest variance and the smallest sample size is paired with the largest variance); and (f) population mean configuration (complete null, partial null, and complete nonnull mean configurations were used in the study;

TABLE 1
Sample Sizes and Population Variances Used in the Simulation Study

J	Sample sizes	Population variances
4	10, 10, 10, 10	1, 1, 1, 1
	9, 10, 10, 11	1, 2, 4, 4
	5, 8, 12, 15	1, 3, 5, 8
	15, 15, 15, 15	
	13, 15, 15, 17	
	7, 12, 18, 23	
	25, 25, 25, 25	
	20, 25, 25, 30	
	10, 20, 30, 40	
7	10, 10, 10, 10, 10, 10, 10	1, 1, 1, 1, 1, 1, 1
	9, 9, 10, 10, 10, 11, 11	1, 1, 2, 2, 3, 3, 4
	5, 6, 8, 10, 12, 14, 15	1, 2, 2, 4, 7, 7, 8
	15, 15, 15, 15, 15, 15, 15	
	13, 14, 15, 15, 15, 16, 17	
	7, 9, 12, 15, 18, 21, 23	
	25, 25, 25, 25, 25, 25, 25	
	20, 22, 24, 25, 26, 28, 30	
	10, 15, 20, 25, 30, 35, 40	

see Table 2 for the specific mean configurations). The Welch (1938, 1951) heteroscedastic test statistics were used for all pairwise and omnibus tests, respectively. The sample size and variance conditions investigated in this study were expected to parallel data characteristics encountered by applied researchers (see Keselman et al., 1998). The nonnull mean configurations adopted in this study provided a priori omnibus powers of 50%, 70%, and 90% with n_js of 10, 15, and 25, respectively, assuming equal sample sizes and variances. Those values are expected to represent the power achieved in many behavioral science investigations.

The simulation program was written in SAS/IML (SAS Institute, Inc., 1989). Pseudorandom normal variates were generated with the SAS generator RANNOR (SAS Institute, Inc., 1985). If Z_{ij} is a standard normal deviate, then $X_{ij} = \mu_j + (\sigma_j Z_{ij})$ is a normal variate with mean μ_j and variance σ^2_j. Five thousand replications were performed for each condition, using a nominal significance level of .05.

Results

To facilitate a discussion of the findings, I reduced the true model rates for the sample size equality and inequality and variance equality and inequality conditions into the three following classifications: (a) equal sample sizes or variances, (b) positively paired sample sizes and variances, and (c) negatively paired sample sizes and variances. Because the pattern of results was similar across the three

TABLE 2
Population Mean Configurations Used in the Simulation Study

			Population means			
μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
J = 4						
0.000	0.000	0.000	0.000			
0.000	0.000	0.000	0.917			
0.000	0.000	0.477	0.954			
0.000	0.000	0.791	0.791			
0.000	0.353	0.706	1.059			
J = 7						
0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.970
0.000	0.000	0.000	0.000	0.000	0.750	0.750
0.000	0.000	0.000	0.366	0.732	0.732	0.732
0.000	0.000	0.450	0.450	0.450	0.900	0.900
0.000	0.169	0.338	0.507	0.676	0.845	1.014

sample size conditions, only the results for the largest sample size condition are presented. True model rates under the complete null were, as expected, close to 95% for all procedures, except for the CEC procedure, which does not control for multiplicity and therefore had deflated true model rates under the complete null (see Tables 3 and 4). True model rate results for all mean configuration conditions as well as for the nonnull mean configuration conditions are discussed next.

Overall True Model Rates

True model rates for all mean configurations are presented in Table 3 ($J = 4$) and Table 4 ($J = 7$). For $J = 4$ and equal sample size or variances, true model rates were considerably higher for the MTP procedure (39.83%) than for any of the remaining procedures (26.18% to 31.20%). Furthermore, the results for the liberal Type I error controlling procedures (CEC, FDR) were higher than were the rates for the α_F controlling procedures, with true model rates for the Tukey and Bonferroni α_F controlling procedures (26.86% and 26.18%, respectively) the lowest of all procedures. When variances and sample sizes were unequal, the MTP procedure again had considerably higher true model rates than any of the remaining procedures, with rates reaching more than 10% higher than those of the Bonferroni and Tukey α_F controlling procedures. Moreover, the CEC procedure performed comparatively worse with unequal sample sizes and variances.

For $J = 7$, the true model rates were highest for the FDR, MTP, and REGWQ procedures, although the differences were not substantial. As expected, true model rates for all procedures decreased from $J = 4$ to $J = 7$, as the probability

TABLE 3
True Model Rate Percentages for $J = 4$ and $N = 100$

	All mean configurations			Nonnull mean configurations			Complete null mean configurations		
	$= n_j/\sigma^2_j$	PP	NP	$= n_j/\sigma^2_j$	PP	NP	$= n_j/\sigma^2_j$	PP	NP
CEC	31.06	21.14	19.30	18.85	6.34	3.85	79.02	80.35	81.22
Bonferroni	26.18	20.30	19.70	8.76	1.38	0.54	95.87	95.99	96.36
Tukey	26.86	20.33	19.68	9.91	1.70	0.75	94.68	94.85	95.39
Hayter (W)	29.47	21.35	20.21	13.10	2.92	1.43	94.96	95.08	95.33
REGWQ(W)	30.46	21.83	20.56	14.22	3.41	1.63	95.43	95.54	96.27
FDR	31.20	22.22	20.85	15.09	3.85	2.11	95.66	95.72	95.83
MTP(W)	39.83	27.09	31.08	26.04	10.11	15.10	94.90	95.08	94.52

Note. N = total sample size; $= n_j/\sigma^2_j$ = equal sample sizes or variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively; (W) = Welch omnibus test.

TABLE 4
True Model Rate Percentages for $J = 7$ and $N = 175$

	All mean configurations			Nonnull mean configurations			Complete null mean configurations		
	= n_j/σ^2_j	PP	NP	= n_j/σ^2_j	PP	NP	= n_j/σ^2_j	PP	NP
CEC	16.55	11.49	10.81	8.49	2.44	1.34	56.86	56.72	58.16
Bonferroni	17.76	16.17	16.05	2.13	0.24	0.05	95.90	95.80	96.17
Tukey	17.99	15.98	15.83	2.74	0.33	0.07	94.24	94.25	94.63
Hayter (W)	18.65	16.23	15.97	3.29	0.43	0.10	95.47	95.24	95.30
REGWQ(W)	20.27	16.68	16.33	5.06	0.81	0.19	96.30	96.04	97.03
FDR	21.08	16.88	16.27	6.12	1.08	0.33	95.88	95.87	95.99
MTP (W)	21.90	16.42	16.85	7.28	0.70	1.22	94.24	94.46	93.81

Note. N = total sample size; = n_j/σ^2_j = equal sample sizes or variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively; (W) = Welch omnibus test.

of detecting the correct pattern of mean differences became much more difficult with 21 pairwise comparisons than with 6 pairwise comparisons. However, true model rates for the CEC procedure were especially affected by the increased number of levels of the independent variable, especially with unequal sample sizes and variances.

True Model Rates for the Nonnull Cases

True model rates for only the nonnull mean configurations are presented in Table 3 ($J = 4$) and Table 4 ($J = 7$). For $J = 4$, true model rates with equal sample sizes or variances were considerably higher for the MTP procedure (26.04%) than for any of the remaining procedures (8.76% to 18.85%). Similar to the overall true model rate results, the results for the liberal Type I error controlling procedures (CEC, FDR) for the nonnull cases were higher than were the rates for the α_F controlling procedures, with true model rates for the Tukey and Bonferroni α_F controlling procedures the lowest of all procedures. When sample sizes and variances were unequal, the Type I error rates of all procedures decreased substantially. However, the true model rates for the MTP were again notably higher than the rates for the remaining procedures, and the true model rates for the liberal MCPs (CEC, FDR) were higher than the rates for the α_F controlling procedures.

For $J = 7$ the true model rates, as for $J = 4$, were severely deflated for the nonnull mean configuration conditions, with the pattern of nonnull true model rates again similar to that for all conditions. The true model rates were highest for the

FDR, MTP, CEC, and REGWQ procedures; the rates for the Tukey (2.74%) and Bonferroni (2.13%) with equal sample sizes or variances were less than half of the rates for the FDR, CEC, and MTP procedures.

Discussion

Researchers in the behavioral sciences are often confronted with the task of evaluating mean differences in studies with more than two levels of the independent variable. In that situation, researchers are confronted with the task of selecting not only an appropriate error rate to control but also a multiple comparison procedure that controls the selected error rate. Those decisions are anything but clear. With respect to selecting an appropriate error rate, there are important rationales for adopting comparisonwise error, familywise error, or false discovery rate control, with the final decision often resting on the nature of the research. Given that fact, the current trend toward adopting MCPs that only control the familywise error rate seems nonsensical. On the other hand, what does make sense is to inform researchers about the consequences of their decisions, such as the likelihood of correctly detecting the underlying population mean configuration with a specific strategy.

With respect to selecting an appropriate MCP, the decision process is no less convoluted. For example, a researcher analyzing data from a one-way between-subjects design with SPSS 10.0 (SPSS Inc., 2000) is provided with a selection of no less than 18 available multiple comparison procedures. It is then the responsibility of the researcher to select the best procedure for analyzing his or her data. According to reviews of the literature, researchers faced with that decision often select popular procedures such as the Tukey procedure (Keselman et al., 1999). Energetic researchers interested in comparing the performance of available MCPs (and thus making an informed decision) are faced the problem that MCPs are almost always evaluated separately for Type I error control and power. Even though the goal of researchers is typically to identify the true pattern of mean differences among their treatment conditions, the MCP literature is limited in that the reader is forced to independently try to fuse together Type I error and power results for specific procedures. For example, researchers are often instructed by MCP researchers, journal editors, textbook authors, and others to avoid MCPs such as the CEC procedure because there is an increase in the probability of Type I errors relative to familywise error controlling procedures (which is true) and to use MCPs such as the Tukey procedure because it provides good familywise error control (which is also true under some conditions). However, what the researcher is rarely told is what strategy provides the best opportunity to identify the true pattern of mean differences among the groups. For example, is the power advantage of the CEC procedure over the Tukey procedures enough to offset the increased probability of Type I errors?

In this study, I used the true model rate in an attempt to provide a clearer picture of the performance of available multiple comparison procedures. The true model rate provides a direct test of the probability of correctly identifying the true underlying population mean configuration. With four levels of the independent variable, the model testing procedure recently proposed by Dayton (1998) had the highest true model rates of any of the procedures, with the FDR and CEC also having very respectable true model rates. The familywise error controlling procedures had the lowest true model rates of any of the investigated procedures, and the differences were magnified when only the nonnull cases were investigated. Those results are consistent with the results of Cribbie and Keselman (in press), who compared the MTP with familywise error controlling MCPs of Shaffer (1986), Hayter (1986), and Hochberg (1988) with three, four, and five levels of the independent variable. Furthermore, the popular Bonferroni and Tukey procedures had the lowest true model rates of any of the familywise error controlling procedures. That is, even though the Bonferroni and Tukey procedures might be less likely to lead to a Type I error in the conventional approach to MCPs, the overall chance that the researcher will find the true pattern of relationships in the data is reduced in comparison with that of other available procedures.

As the number of levels of the independent variable increased from four to seven, it became much more difficult for any of the procedures to identify the true underlying mean configuration. However, the familywise error controlling procedures continued to have lower true model rates than the FDR and MTP procedures, and the Tukey and Bonferroni procedures continued to have lower true model rates than the remaining familyise error controlling procedures (Hayter, REGWQ). Therefore, even though familywise error controlling procedures are supposed to be more reliable with an increased number of levels of the independent variable because they reduce the risk of Type I error inflation, more liberal procedures such as the FDR MCP, and model testing procedures such as the MTP, outperform the familywise error controlling procedures with respect to identifying the true underlying population mean configuration.

In summary, researchers are faced with important decisions regarding the selection of a multiple comparison procedure. In making those decisions, it is crucial for researchers to consider the nature of the study, including the relative importance of Type I errors, power, and the probability of correctly identifying the true population mean configuration. The results of the present study indicate that regardless of data characteristics (e.g., variance equality and inequality) or the MCP adopted, the probability of detecting the true underlying population mean configuration is very difficult with currently available procedures and is evidence of a need for more research into procedures that will provide a better opportunity of correctly detecting population mean configurations. However, given that a primary goal of researchers is to increase the probability of correctly identifying the true population mean configuration (i.e., not committing any Type I or Type II errors), the results of the current study lend support for the adoption of recently proposed procedures such as the model testing procedure (Dayton, 1998) and the false discovery rate controlling procedure (Benjamini & Hochberg, 1995). Those procedures consistently provide the researcher with a better chance of determining the correct underlying pattern of the population means as compared with that of familywise error controlling procedures. On the other hand, if the nature of a study makes it mandatory that strict familywise error control be imposed, researchers are advised to use the REGWQ MCP instead of the popular Bonferroni or Tukey MCPs, because the REGWQ procedure provides a better opportunity of detecting the true population mean configuration.

REFERENCES

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.

Bonferroni, C. E. (1937). Teoria statistica delle classi e calcolo delle probabilita. In *Volume in onore di Ricarrdo dalla Volta* (pp. 1–62). Firenze, Italy: Universita di Firenza.

Carmer, S. G., & Walker, W. M. (1985). Pairwise multiple comparisons of treatment means in agronomic research. *Journal of Agronomic Education, 14,* 19–26.

Cribbie, R. A., & Keselman, H. J. (in press). Pairwise multiple comparison procedures: A model testing approach versus stepwise procedures. *British Journal of Mathematical and Statistical Psychology.*

Davis, C., & Gaito, J. (1984). Multiple comparison procedures within experimental research. *Canadian Psychology, 25,* 1–13.

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

Duncan, D. B. (1955). Multiple range and multiple *F* tests. *Biometrics, 11,* 1–42.

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.

Gaito, J., & Nobrega, J. A. (1981). A note on multiple comparisons as an ANOVA problem. *Bulletin of the Psychonomic Society, 17,* 169–170.

Hancock, G. R., & Klockars, A. J. (1996). The quest for α: Developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research, 66,* 269–306.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association, 81,* 1000–1004.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75,* 800–802.

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

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., & 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.

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

Petrinovich, L. F., & Hardyck, C. D. (1969). Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. *Psychological Bulletin, 71,* 43–54.

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

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

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

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

Sato, T. (1996). Type I and Type II error in multiple comparisons. *Journal of Psychology, 130,* 293–302.

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. *Annual Review of Psychology, 46,* 561–584.

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

Tukey, J. W. (1953). *The problem of multiple comparisons.* Unpublished manuscript, Department of Statistics, Princeton University, Princeton, NJ.

Welch, B. L. (1938). The significance of the difference between two means when 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. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association, 72,* 566–575.

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