British Journal of Mathematical and Statistical Psychology (2003), 56, 167–182

www.bps.org.uk

Pairwise multiple comparisons: A model comparison approach versus stepwise procedures

Robert A. Cribbie[1]* and H. J. Keselman[2]
[1]York University, Canada
[2]University of Manitoba, Canada

Researchers in the behavioural sciences have been presented with a host of pairwise multiple comparison procedures that attempt to obtain an optimal combination of Type I error control, power, and ease of application. However, these procedures share one important limitation: intransitive decisions. Moreover, they can be characterized as a piecemeal approach to the problem rather than a holistic approach. Dayton has recently proposed a new approach to pairwise multiple comparisons testing that eliminates intransitivity through a model selection procedure. The present study compared the model selection approach (and a protected version) with three powerful and easy-to-use stepwise multiple comparison procedures in terms of the proportion of times that the procedure identified the true pattern of differences among a set of means across several one-way layouts. The protected version of the model selection approach selected the true model a significantly greater proportion of times than the stepwise procedures and, in most cases, was not affected by variance heterogeneity and non-normality.

1. Introduction

Over the past few decades researchers have been presented with a myriad of new procedures and approaches for testing pairwise comparisons in one-way completely randomized designs. With few exceptions, the methods previously proposed have focused on strategies for achieving a better balance between Type I error control, power, and ease of applicability.

However, as Dayton (1998) explains, multiple comparison procedures (MCPs) share one important limitation: there is a high probability that conclusions from the study will contain intransitive decisions. For example, a researcher conducting all pairwise comparisons in a one-way randomized design with three groups ($j = 1, \ldots, J$, where $J = 3$) may decide not to reject hypotheses implied by $\mu_1 = \mu_2$ or $\mu_2 = \mu_3$, but to reject

*Requests for reprints should be addressed to Dr Robert Cribbie, Department of Psychology, York University, Toronto, ON M3J 1P3, Canada (e-mail: cribbie@yorku.ca).

$\mu_1 = \mu_3$, based on the results from a typical MCP. Interpreting the results of this experiment can be ambiguous, especially concerning the outcome for μ_2. Accepting intransitive decisions in multiple comparisons testing has become commonplace with researchers, even though this violates an underlying principle of statistical hypothesis testing, the presumption of distinct populations. For example, a test of the omnibus null hypothesis $H_0: \mu_1 = \mu_2 = \mu_3$ evaluates whether all samples are drawn from the same population against the alternative that all samples (or a subset of the samples, e.g., one pair) were drawn from distinct populations. In the above example, concluding that the second sample was drawn from two separate populations not only violates the presumption of distinct populations, but also makes it difficult for applied researchers to generate logical conclusions concerning the nature of relationships (i.e., population differences) investigated. (Of course, researchers should be aware that failing to reject a hypothesis does not require that they accept it as true; i.e., they have the option of suspending judgement.)

A strategy recently proposed by Dayton (1998) attempts to rectify the problem of intransitive decisions by investigating all possible transitive population models (i.e., mean configurations) in order to identify the 'true' pattern of differences among the set of means. Again considering a design with $J = 3$, a researcher would be faced with comparing (and selecting the best of) the $k = 2^{J-1} = 2^{3-1} = 4$ transitive population models, instead of determining if any or all of the $m = \binom{J}{2} = 3$ pairwise comparisons are significant. With $J = 3$ the researcher would be comparing the models $\{\mu_1 \mu_2 \mu_3\}$, $\{\mu_1, \mu_2 \mu_3\}$, $\{\mu_1 \mu_2, \mu_3\}$, and $\{\mu_1, \mu_2, \mu_3\}$, where means separated by commas represent distinct populations (i.e., populations with unequal means). In addition to eliminating intransitive decisions, Dayton's approach takes a more 'holistic' approach to the testing of multiple comparisons. That is, the model comparison approach allows researchers to examine, and thus compare, the relative competitiveness of various models.

2. Review of the traditional approach

Researchers conducting all m pairwise comparisons with traditional MCPs (e.g., Tukey's honestly significant difference (HSD)) are faced with important decisions regarding Type I error control as a result of conducting multiple (and related) tests of significance. First, a significance level, or decision criterion, must be specified. Researchers have made a practice of selecting some accepted level of significance (e.g., $\alpha = .05$), even though it is important to acknowledge that the selection of α should be based on the nature of the research.

Second, researchers must also specify the unit of analysis over which Type I error control will be applied. The comparisonwise error rate (α_{PC}) sets the probability of falsely rejecting the null hypothesis for each comparison at α, and has been supported and recommended by a number of authors (e.g., Carmer & Walker, 1985; Davis & Gaito, 1984; Rothman, 1990; Saville, 1990; Wilson, 1962). The primary disadvantage of α_{PC} control is that the probability of making at least one Type I error increases with the number of comparisons, approaching $1 - (1 - \alpha)^m$. Critics of α_{PC} (e.g., Ryan, 1959, 1962) have often recommended control of the familywise error rate (α_{FW}). With α_{FW} control, the probability of falsely rejecting one or more hypotheses in a family of hypotheses is set at α. Controlling α_{FW} has been recommended by many researchers (e.g., Hancock & Klockars, 1996; Petrinovich & Hardyck, 1969; Ryan, 1959, 1962;

Tukey, 1953) and is 'the most commonly endorsed approach to accomplishing Type I error control' (Seaman, Levin, & Serlin, 1991, p. 577). Keselman *et al.* (1998) reported that approximately 85% of researchers conducting pairwise comparisons adopt some form of α_{FW} control. The main advantage of procedures that provide α_{FW} control is that the probability of making a Type I error does not increase with the number of comparisons conducted in the experiment. Although many MCPs purport to control α_{FW}, it is important to distinguish between those procedures that provide 'strong' α_{FW} control (α_{FW} is maintained at α when all population means are equal, as well as when multiple subsets of the population means are equal) and procedures that provide 'weak' α_{FW} control (α_{FW} is maintained at α only when all population means are equal).

After establishing a level of significance and error rate, an appropriate pairwise MCP must be selected. Traditional pairwise MCPs can test hypotheses either simultaneously or in a series of steps (stepwise). A simultaneous MCP conducts all comparisons regardless of whether the omnibus test, or any other comparison, is significant (or not significant) at a constant critical value. Examples of simultaneous MCPs include procedures proposed by Tukey (1953), Scheffe (1953) and Bonferroni (1937). A stepwise (or sequential) MCP considers either the significance of the omnibus test or the significance of other comparisons (or both) in generating critical values. MCPs that require a significant omnibus test in order to conduct pairwise comparisons are referred to as protected tests. MCPs that consider the significance of other comparisons when evaluating the significance of a particular comparison can be either step-down or step-up procedures. Step-down MCPs begin by testing the largest pairwise mean difference (i.e., the mean difference resulting in the largest numerical value). Non-significance of this mean difference implies non-significance for smaller pairwise mean differences. Step-up procedures begin by testing the smallest pairwise mean difference. Significance of this difference can imply significance for larger pairwise mean differences. Three stepwise procedures were investigated in this study.

Shaffer (1986) presented a step-down MCP that made use of the fact that the maximum number of true null hypotheses at any stage of testing is often less than $m - i + 1$ ($i = 1, \ldots, m$), the denominator used by Holm (1979) in determining a critical alpha ($\alpha_i = \alpha/(m - i + 1)$) for evaluating the significance of each of the m pairwise comparisons. Shaffer provides tables for $J = 3, \ldots, 10$, for the maximum number of true nulls given rejections at previous stages of sequential testing. Therefore, given ordered (smallest to largest) p-values, $p_{(1)} < p_{(2)} < \ldots < p_{(m)}$, from an appropriate two-sample test statistic, a researcher would reject $H_0: \mu_j = \mu_{j'}$ ($j \neq j'$) if $p_i \leq \alpha/C_k$, where C_k is the maximum number of true nulls possible at the ith stage of testing. If any H_0 is not rejected, testing stops and hypotheses associated with the remaining larger p_is are declared non-significant. Shaffer also suggested incorporating an omnibus test with her original procedure, where rejection of the omnibus test is considered in deriving C_k at the first stage of testing. Non-significance of the omnibus test results in no further testing. The protected version of the Shaffer procedure is hereafter referred to as PSHR.

Hochberg (1988) proposed a step-up procedure (HBG) that combined Simes's (1986) inequality with Holm's (1979) testing procedure. The p-values are ordered (smallest to largest) and for any $i = m, m - 1, \ldots, 1$, if $p_i \leq \alpha/(m - i + 1)$, the HBG procedure rejects all hypotheses where $i' \leq i$.

Hayter (1986) proposed a modification (HTR) to Fisher's (1935) LSD procedure that would provide strong α_{FW} control. Like the LSD procedure, no comparisons are tested unless the omnibus test is significant. If the omnibus test is significant, then a pairwise

H_0 is rejected if

$$|t| \geq \frac{q(\alpha, J-1, df_e)}{2^{1/2}},$$

where t represents Student's two-sample test statistic and $q(\alpha, J-1, df_e)$ is the α-level critical value from the Studentized range distribution with $J-1$ numerator degrees of freedom and error degrees of freedom (df_e) from an appropriate omnibus test (e.g., ANOVA F; Welch, 1951).

Each of the above procedures has been shown to control α_{FW} in the 'strong' sense when the validity assumptions (e.g., variance homogeneity, normality) of traditional test statistics (e.g., t, F) have been satisfied. In addition, power is an important consideration when selecting a MCP. Three popular conceptualizations of power with pairwise comparisons are any-pair, all-pairs and average per-pair power. Any-pair power is the probability of detecting any true pairwise mean difference, all-pairs power is the probability of detecting all true pairwise mean differences, and average per-pair power is the average probability of rejecting a true pairwise mean difference across all pairwise comparisons. When the rate of Type I errors is comparable across procedures, researchers can compare MCPs with respect to power. For example, although the Bonferroni and Scheffe MCPs provide strong α_{FW} control, they are not recommended for testing all pairwise comparisons because they are often substantially less powerful than other available MCPs (that also provide strong α_{FW} control). In addition, although Tukey's HSD is the most powerful of the simultaneous MCPs, it is also not recommended over the stepwise procedures due to a reduction in power under most testing situations.

3. The model testing approach

Dayton's (1998) model testing procedure (MTP) is based on the Akaike (1974) information criterion (AIC). Dayton also examined Schwartz's (1978) information criterion, but found it did not perform as well as AIC. Mutually exclusive and transitive models are each evaluated using AIC, and the model having the minimum AIC (i.e., the minimum loss of precision relative to the true model) is retained, where

$$\text{AIC} = -2\{-(N/2)\ln(2\pi) - (N/2)\ln(S_w^2)\} + SS_w + \sum_{j=1}^{J} n_j(\overline{X}_j - \overline{X}_{kj})^2 + 2q,$$

S_w^2 is the biased within-cell variance (i.e., SS_w/N), SS_w is the within-group sums of squares, n_j is the number of subjects in the jth group, $\overline{X}_j$ is the mean of the jth group, $\overline{X}_{kj}$ is the estimated sample mean for the jth group (given the hypothesized population mean configuration for the kth model), and q is the number of independent parameters estimated in fitting the model. In addition, Dayton (1998) showed that the MTP can be modified to handle heterogeneous treatment group variances. Like the original procedure, mutually exclusive and transitive models are each evaluated using AIC, and the model having the minimum AIC is retained. For heterogeneous variances,

$$\text{AIC} = -2\left\{(-N/2)(\ln(2\pi) + 1) - \frac{1}{2}\left(\sum_{j=1}^{J} n_j \ln(S_j^2)\right)\right\} + 2q,$$

where N is the total number of subjects in the experiment ($\Sigma_j n_j$) and S_j^2 is the biased variance for the jth group, substituting the estimated group mean (given the

hypothesized mean configuration for the kth model) for the actual group mean in the calculation of the variance. The heterogeneous variance AIC statistic adopted in this paper is referred to by Dayton (1998) as the unrestricted heterogeneous model (in contrast to the restricted heterogeneous model also presented by Dayton). Note that this form of the AIC does not pool variances, which is comparable to the approach adopted by heteroscedastic test statistics (e.g., Welch, 1938). Both the homogeneous variance and heterogeneous variance versions of the AIC statistics assume that the errors are normally distributed.

As stated previously, the MTP redefines the traditional view of pairwise multiple comparisons. Consider $J = 4$, where $k = 2^{J-1} = 2^{4-1} = 8$ transitive models are being compared ($\{1234\}, \{1,234\}, \{12,34\}, \{123,4\}, \{1,2,34\}, \{12,3,4\}, \{1,23,4\}, \{1,2,3,4\}$). Using the MTP approach, a researcher would select the model with the minimum AIC value and discuss the implications of that decision within the realm of his/her a priori theory. Furthermore, as indicated, researchers can compare the AIC values (e.g., across homogeneous and heterogeneous models), thus assessing the relative competitiveness of the models. Note that there are no decisions regarding the level of significance, the error rate, or the definition of power (per-pair, all-pairs, etc.) with which to compare the procedures. In fact, the definitions of Type I and Type II error can be discarded in favour of a rate referred to here as the 'true-model rate', representing the proportion of times that the AIC statistic selects the true population model (although the true-model rate could perhaps be conceptualized as a blending of the classical Type I error and power rates, because hypotheses concerning population means are not tested with this approach we, and others (Dayton, 1998), prefer to conceptualize the true-model rate with respect to model comparisons). Dayton showed that the true-model rate for the MTP was larger than that for Tukey's HSD across many population mean configurations.

One finding reported by Dayton is that the AIC has a slight bias for selecting more complicated models than the true model. For example, Dayton found that for the mean pattern $\{12,3,4\}$, AIC selected the more complicated pattern $\{1,2,3,4\}$ more than 10% of the time, whereas AIC only rarely selected less complicated models (e.g., $\{12,34\}$). This tendency can present a special problem for the complete null case, and consequently it is recommended that an omnibus test be used to screen for the complete null. Rejection of the omnibus test would result in comparing the k models, whereas not rejecting the omnibus test would result in accepting the complete null population model.

To summarize, the MTP has important advantages in being able to eliminate intransitive decisions and provide a more holistic approach to summarizing mean differences in studies where pairwise multiple comparisons are performed. However, before recommending the MTP to applied researchers it is important to evaluate how the MTP performs (with respect to the true-model rate) relative to other available MCPs.

4. Method

A Monte Carlo study was used to compare the true-model rate of the MTP with that of the PSHR, HBG and HTR stepwise MCPs. In addition to the original MTP, we also investigated a protected version of the MTP using Welch's (1951) omnibus test (WMTP). For the MTP and WMTP, the true-model rate is defined as the probability of selecting the correct population model with the AIC statistic. For the PSHR, HBG and HTR

Table 1. Sample sizes and population variances used in the Monte Carlo study

J	Sample sizes	Population variances
3	10, 10, 10	1, 1, 1
	9, 10, 11	1, 2, 4
	5, 10, 15	1, 4, 8
	15, 15, 15	
	13, 15, 17	
	7, 15, 23	
	19, 19, 19	
	17, 19, 21	
	9, 19, 29	
4	10, 10, 10, 10	1, 1, 1, 1
	9, 10, 10, 11	1, 2, 4, 4
	5, 7, 13, 15	1, 3, 5, 8
	15, 15, 15, 15	
	13, 15, 15, 17	
	7, 11, 19, 23	
	19, 19, 19, 19	
	17, 19, 19, 21	
	9, 14, 24, 29	
5	10, 10, 10, 10, 10	1, 1, 1, 1, 1
	9, 10, 10, 10, 11	1, 1, 2, 3, 4
	5, 6, 10, 14, 15	1, 2, 4, 6, 8
	15, 15, 15, 15, 15	
	13, 14, 15, 16, 17	
	7, 9, 15, 21, 23	
	19, 19, 19, 19, 19	
	17, 18, 19, 20, 21	
	9, 11, 19, 27, 29	

procedures, the true-model rate is defined as the probability of detecting all false pairwise hypotheses and not rejecting any true null hypotheses. For example, for the $J = 3$ population model $\{12, 3\}$, the stepwise procedures would be required to reject $H_0 : \mu_1 = \mu_3$ and $H_0 : \mu_2 = \mu_3$, but not to reject $H_0 : \mu_1 = \mu_2$. For the PSHR and HTR procedures the Welch (1951) omnibus test was used. Pairwise comparisons for the stepwise procedures were examined with Welch's (1938) two-sample statistic.

Seven variables were manipulated in this study: number of levels of the independent variable; total sample size; degree of sample-size imbalance; degree of variance inequality; pairings of group sizes and variances; configuration of population means; and population distribution shape.

To evaluate the effect of the number of pairwise comparisons on the true-model rate, the number of levels of the independent variable was set at $J = 3$, 4 and 5, resulting in $m = 3$, 6 and 10 pairwise comparisons, and $k = 4$, 8 and 16 transitive models, respectively.

In order to investigate the effects of sample size, the total sample size (N) was manipulated by setting the average $n_j = 10$, 15, and 19, resulting in $N = 30$, 45 and 57 for $J = 3$, $N = 40$, 60 and 76 for $J = 4$, and $N = 50$, 75 and 95 for $J = 5$. For the

Table 2. Population mean configurations used in the Monte Carlo study

Population means				
μ_1	μ_2	μ_3	μ_4	μ_5
		$J = 3$		
0.000	0.000	0.000		
0.000	0.000	1.021		
0.000	0.386	1.158		
0.000	0.590	1.179		
		$J = 4$		
0.000	0.000	0.000	0.000	
0.000	0.000	0.000	1.031	
0.000	0.000	0.893	0.893	
0.000	0.000	0.538	1.077	
0.000	0.410	0.410	1.229	
0.000	0.631	0.631	1.263	
0.000	0.399	0.799	1.198	
		$J = 5$		
0.000	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	1.048
0.000	0.000	0.000	0.856	0.856
0.000	0.000	0.383	0.383	1.148
0.000	0.000	0.469	0.938	0.938
0.000	0.000	0.360	0.719	1.079
0.000	0.663	0.663	0.663	1.326
0.000	0.560	0.560	1.121	1.121
0.000	0.411	0.411	0.822	1.234
0.000	0.226	0.452	0.904	1.131
0.000	0.297	0.593	0.890	1.186

non-null mean configurations used in this study, the group sizes of 10, 15 and 19 result in a prior omnibus (F statistic) power estimates of approximately .6, .8 and .9, respectively (assuming equal group sizes and variances).

Sample-size balance/imbalance was also manipulated in this study. Keselman *et al.* (1998) reported in a review of studies published in educational and psychological journals that unbalanced designs were more common than balanced designs. Three sample-size conditions were used (equal n_j, moderately unequal n_j and extremely unequal n_j). The sample sizes used in this study are presented in Table 1.

Degree of variance heterogeneity was also manipulated. According to Keselman *et al.* (1998) ratios of largest to smallest variances of 8:1 were not uncommon in educational and psychological studies and can have deleterious effects on the performance of many MCPs, especially when paired with unequal sample sizes. Therefore, three levels of variance equality/inequality were used: equal variances; largest to smallest variance ratio of 4:1; and largest to smallest variance ratio of 8:1. See Table 1 for specific group variances for $J = 3$, 4 and 5. The homogeneous and heterogeneous variance AIC models were adopted with equal and unequal variances, respectively.

Table 3. True-model rates (%) over all mean conditions

J	*N*	Condition	PSHR	HBG	HTR	MTP	WMTP
3	30	$= n_j$ or $= \sigma_j^2$	30.8	29.2	30.8	35.9	37.1
		PP	26.0	25.6	26.0	27.5	27.2
		NP	26.2	25.4	26.2	31.0	31.0
	45	$= n_j$ or $= \sigma_j^2$	35.6	33.7	35.6	41.3	43.6
		PP	28.2	27.2	28.2	31.0	29.8
		NP	28.0	26.8	28.0	34.5	34.7
	57	$= n_j$ or $= \sigma_j^2$	39.3	37.4	39.3	44.8	47.9
		PP	29.8	28.5	29.8	33.2	31.8
		NP	29.9	28.2	29.9	36.9	37.8
4	40	$= n_j$ or $= \sigma_j^2$	15.2	15.2	15.5	25.2	24.4
		PP	13.8	14.0	13.9	17.7	16.2
		NP	13.6	13.7	13.6	20.3	19.0
	60	$= n_j$ or $= \sigma_j^2$	17.8	17.6	18.2	30.3	31.1
		PP	14.3	14.3	14.4	20.8	18.7
		NP	14.0	14.1	14.1	23.8	22.5
	76	$= n_j$ or $= \sigma_j^2$	20.2	19.9	20.5	33.5	35.1
		PP	14.7	14.6	14.8	23.1	20.6
		NP	14.4	14.4	14.5	26.3	25.3
5	50	$= n_j$ or $= \sigma_j^2$	9.0	9.1	9.1	13.4	14.4
		PP	8.8	8.8	8.7	9.9	10.1
		NP	8.6	8.6	8.5	11.3	12.0
	75	$= n_j$ or $= \sigma_j^2$	9.8	9.8	9.9	16.6	18.4
		PP	8.9	8.9	8.9	12.0	11.6
		NP	8.7	8.7	8.6	13.8	14.7
	95	$= n_j$ or $= \sigma_j^2$	10.9	10.7	11.0	19.1	21.4
		PP	9.0	9.0	9.1	13.6	13.0
		NP	8.8	8.8	8.8	15.5	16.7

Note: *J* = number of groups; *N* = total sample sizes; PSHR = protected Shaffer, HBG = Hochberg, HTR = Hayter, MTP = model testing procedure, WMTP = model testing procedure with a Welch omnibus test; $= n_j$ or $= \sigma_j^2$ represents equal sample sizes or population variances; PP represents positively paired variances and sample sizes; NP represents negatively paired variances and sample sizes.

The specific pairings of unequal variances and sample sizes can have differing effects on test statistics. Therefore, both positive and negative pairings were evaluated for conditions with unequal variances and unequal sample sizes.

Several configurations of non-null population means were also investigated in this study. Following Ramsey's (1978) definitions of mean configuration, we examined equally spaced, minimum variability and maximum variability configurations (see Table 2).

Another factor examined was population distribution shape. In addition to normally distributed data, we also investigated cases where the data were obtained from a χ_3^2 distribution (skewness = 1.63, kurtosis = 4.00). We selected this non-normal distribution because Sawilowsky and Blair (1992) found that popular test statistics were adversely affected when distributions had similar values of skewness and kurtosis.

Table 4. True-model rates (%) for the complete null cases

J	N	Condition	PSHR	HBG	HTR	MTP	WMTP
3	30	$= n_j$ or $= \sigma_j^2$	94.6	95.7	94.6	65.4	94.6
		PP	94.7	95.9	94.7	75.4	94.8
		NP	93.0	93.8	93.0	66.4	93.0
	45	$= n_j$ or $= \sigma_j^2$	94.7	95.7	94.7	67.9	94.7
		PP	95.0	95.9	95.0	79.3	95.0
		NP	93.4	94.3	93.4	67.9	93.5
	57	$= n_j$ or $= \sigma_j^2$	94.3	95.1	94.3	68.1	94.3
		PP	94.6	95.5	94.6	80.1	94.6
		NP	93.5	94.3	93.5	69.2	93.6
4	40	$= n_j$ or $= \sigma_j^2$	94.8	96.0	94.5	51.5	94.1
		PP	95.0	96.0	94.7	65.0	94.2
		NP	92.9	94.1	92.6	59.1	92.5
	60	$= n_j$ or $= \sigma_j^2$	95.2	96.1	94.8	54.2	94.5
		PP	95.2	96.3	95.0	69.1	94.6
		NP	93.7	94.6	93.5	61.7	93.2
	76	$= n_j$ or $= \sigma_j^2$	94.6	95.6	94.4	55.0	94.1
		PP	94.9	95.9	94.7	70.8	94.3
		NP	93.8	94.8	93.6	63.5	93.3
5	50	$= n_j$ or $= \sigma_j^2$	95.5	96.3	94.9	39.6	94.0
		PP	95.7	96.5	95.2	53.5	94.1
		NP	93.8	94.5	93.1	49.9	92.5
	75	$= n_j$ or $= \sigma_j^2$	95.3	96.0	94.8	41.9	93.9
		PP	95.6	96.4	95.1	57.7	94.1
		NP	93.7	94.5	93.2	53.2	92.5
	95	$= n_j$ or $= \sigma_j^2$	95.3	96.1	94.8	42.5	94.0
		PP	95.5	96.2	95.1	60.2	94.2
		NP	94.0	94.8	93.6	54.9	93.0

Note: See Table 3.

The simulation program was written in SAS/IML (SAS Institute, 1989). Pseudo-random normal variates were generated with the SAS generator RANNOR (SAS Institute, 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_j^2. To generate data from a χ_3^2 distribution, three standard normal variates were squared and summed. The χ_3^2 variates were standardized and transformed to variates with mean μ_j and variance σ_j^2.

Five thousand replications were performed for each condition, with a nominal significance level of .05 used with the omnibus tests and stepwise MCPs.

5. Results

The true-model rates for the stepwise and model-testing procedures are presented in Tables 3, 4 and 5. Table 3 presents the true-model rates across all complete null and non-null conditions, Table 4 presents the true-model rates under only complete null conditions and Table 5 presents the true-model rates under only non-null conditions.

Table 5. True-model rates (%) for the non-null cases

J	*N*	Condition	PSHR	HBG	HTR	MTP	WMTP
3	30	$= n_j$ or $= \sigma_j^2$	9.5	7.1	9.5	26.0	17.9
		PP	3.1	2.2	3.1	11.5	4.7
		NP	3.9	2.5	3.9	19.2	10.3
	45	$= n_j$ or $= \sigma_j^2$	15.9	13.1	15.9	32.4	26.5
		PP	5.9	4.3	5.9	15.0	8.1
		NP	6.2	4.3	6.2	23.4	15.1
	57	$= n_j$ or $= \sigma_j^2$	21.0	18.2	21.0	37.1	32.5
		PP	8.2	6.2	8.2	17.5	10.9
		NP	8.7	6.2	8.7	26.1	19.2
4	40	$= n_j$ or $= \sigma_j^2$	1.9	1.7	2.3	20.8	13.0
		PP	0.3	0.3	0.4	9.8	3.2
		NP	0.3	0.3	0.4	13.8	6.8
	60	$= n_j$ or $= \sigma_j^2$	5.0	4.5	5.4	26.3	20.5
		PP	0.8	0.7	0.9	12.8	6.1
		NP	0.7	0.6	0.9	17.4	10.7
	76	$= n_j$ or $= \sigma_j^2$	7.8	7.3	8.2	29.9	25.3
		PP	1.3	1.1	1.5	15.1	8.3
		NP	1.2	1.0	1.4	20.1	13.9
5	50	$= n_j$ or $= \sigma_j^2$	0.4	0.3	0.5	10.8	6.5
		PP	0.1	0.1	0.1	5.5	1.7
		NP	0.0	0.0	0.1	7.5	4.0
	75	$= n_j$ or $= \sigma_j^2$	1.3	1.1	1.5	14.0	10.9
		PP	0.2	0.2	0.3	7.4	3.3
		NP	0.1	0.1	0.2	9.9	6.9
	95	$= n_j$ or $= \sigma_j^2$	2.4	2.2	2.6	16.7	14.1
		PP	0.4	0.3	0.4	9.0	4.9
		NP	0.3	0.2	0.3	11.6	9.2

Note: See Table 3.

The pattern of results for the normally distributed and the chi-square distributed data were very similar and accordingly the results have been averaged over these conditions. We will, however, discuss cases where the results differed between normally and non-normally distributed data.

5.1. Overall true-model rates

For $J = 3$, the PSHR and HTR stepwise MCPs (which are equivalent for $J = 3$) were uniformly (although not substantially) better at detecting the true model than the HBG procedure. In addition, the MTP and WMTP had uniformly higher true-model rates than any of the stepwise MCPs. For example, when sample sizes and variances were negatively paired, there was approximately a 10 percentage point advantage for the WMTP over any of the stepwise MCPs. There was little difference between the true-model rates of the MTP and WMTP approaches.

For $J = 4$, the true-model rates of all of the procedures were substantially less than the rates for $J = 3$. However, the fact that the true-model rates for $J = 4$ were lower than

the rates for $J = 3$ is somewhat expected given that there is an increase in the complexity of the designs (i.e., for the stepwise MCPs six correct decisions must be made instead of three, and for the MTP and WMTP there are eight models to select from instead of four). There was very little difference in the true-model rates of the stepwise MCPs. As was found for $J = 3$, the true-model rates for the MTP and WMTP were significantly larger than the rates for the stepwise MCPs, with advantages reaching more than 10%

For $J = 5$, there was a continued decrease in the true-model rates relative to the rates for $J = 3$ and $J = 4$. The true-model rates for the stepwise MCPs were again very similar, with true model rates for the MTP and WMTP consistently larger than the rates for the stepwise procedures.

5.2. True-model rates for the complete null case

The true-model rates for the PSHR, HBG, HTR and WMTP were large (0.95) and consistent across most conditions, indicating that one or more comparisons were falsely declared significant approximately 5% of the time. The true-model rates became moderately depressed when the variances and sample sizes were negatively paired, although a further examination of the results indicates that this was true primarily for the chi-square data. Specifically, for chi-square data with negatively paired sample sizes and variances, the true-model rates were on average .91 for the WMTP.

The true-model rates for the MTP were significantly lower than the rates for any of the stepwise procedures or the WMTP. For example, with $J = 5$ and equal variances or sample sizes, the MTP selected models other than the complete null model in more than 50% of the cases, regardless of sample size, whereas PSHR, HBG, HTR or WMTP never selected models other than the complete null model less than 92.5% of the time.

5.3. True-model rates for the non-null cases

For the non-null cases, the HTR procedure had consistently higher true-model rates than the HBG or PSHR procedures over all conditions, although the differences were not substantial under any condition. However, there was a significant difference in the true-model rates between the MTP/WMTP and the stepwise MCPs. For $J = 3$, the true-model rates for the MTP were, on average, three times as large as the rates for any of the stepwise procedures. These differences were even more pronounced for $J = 4$ and $J = 5$, even though floor effects influenced absolute true-model rate differences between the procedures. For $J = 4$ the true-model rates for the MTP were on average 15 times as large as the rates for any of the stepwise procedures, and the true-model rates for the WMTP were on average eight times as large as the rates for the stepwise procedures. For $J = 5$ the true-model rates for the MTP were on average 34 times as large as the rates for any of the stepwise procedures, and the true-model rates for the WMTP were on average 21 times as large as the rates for the stepwise procedures. To further explore these effects, we simulated several $J = 5$ conditions with larger treatment effects than those previously investigated, in an attempt to remove floor effects. As expected, the advantage of the MTP and WMTP over the stepwise procedures increased as floor effects were removed. For example, with $J = 5$, $N = 50$, extremely unequal n_j and largest to smallest variance ratio of 8:1, the average true model rate for the HTR procedure across ten non-null configurations was 0.70%, whereas the average true model rate for the WMTP across the ten non-null configurations was 37.48%, or 53 times that of the HTR procedure.

6. Addendum

Based on the comments of a reviewer, additional data were generated. In particular, this reviewer felt that the procedures should also be compared for more extreme cases of non-normality and variance heterogeneity. Accordingly, we examined two more cases of non-normality by generating data from two *g*-and-*h* distributions (Hoaglin, 1985) and by creating unequal variance such that the ratio of the largest to smallest variance was 16:1.

Specifically, we chose to investigate a $g = 0$ and $h = 0.5$ ($\gamma_1 = 0$ and γ_2 is undefined) and a $g = 0.5$ and $h = 0.5$ (γ_1 and γ_2 are undefined) distribution (see Wilcox, 1997, p. 73). To give meaning to these values it should be noted that for the standard normal distribution $g = h = 0$. Thus, when $g = 0$ a distribution is symmetric and the tails of a distribution will become heavier as h increases in value. Finally, it should be noted that though the selected combinations of g and h result in extremely heavy-tailed distributions, these values were investigated to indicate how well/poorly the tests will perform under *extreme* conditions.

To generate data from a *g*-and-*h* distribution, standard normal variables were transformed via

$$X_{ij} = \frac{\exp(gZ_{ij}) - 1}{g} \exp\left(\frac{hZ_{ij}^2}{2}\right),$$

according to the values of g and h selected for investigation ($g > 0$). To obtain a distribution with standard deviation σ_j, each X_{ij} was multiplied by a value of σ_j. It is important to note that this does not affect the value of the null hypothesis when $g = 0$ (see Wilcox, 1994, p. 297). However, when $g > 0$, the population mean for a *g*-and-*h*-distributed variable is

$$\mu_{gh} = \frac{\exp\{g^2/(2(1-h))\} - 1}{g(1-h)^{1/2}}$$

(see Hoaglin, 1985, p. 503). Thus, for those conditions where $g > 0$, μ_{gh} was first subtracted from X_{ij} before multiplying by σ_j.

Lastly, it should be noted that the standard deviation of a *g*-and-*h* distribution is not equal to one, and thus the standard deviation values reflect only the amount that each random variable is multiplied by and not the actual values of the standard deviations (see Wilcox, 1994, p. 298). As Wilcox noted, the values for the variances (standard deviations) more aptly reflect the ratio of the variances (standard deviations) between the groups.

The unequal variances were, as indicated, modified so that the disparity between the largest and smallest cases would now be in a 16:1 ratio. Accordingly, the $J = 3$, $J = 4$ and $J = 5$ values were (1, 8, 16), (1, 6, 10, 16) and (1, 4, 8, 12, 16), respectively.

For the three designs investigated, we selected the two most discrepant sample-size cases: 5, 10, 15 and 7, 15, 23 ($J = 3$), 5, 7, 13, 15 and 7, 11, 19, 23 ($J = 4$) and 5, 6, 10, 14, 15 and 7, 9, 15, 21, 23 ($J = 5$). These unequal sample sizes were both positively and negatively paired with the unequal variances (equal sample-size cases were not reinvestigated).

We generated data under the same mean configurations as enumerated in Table 2 (we also included some larger mean differences to eliminate some floor effects). As in the original investigation, 5000 replications for each combination of non-normal distribution and sample-size condition were generated. The true-model rates under all conditions, the null condition and the non-null conditions are presented in Table 6. A

Table 6. True-model rates (%) for the most heterogeneous sample-size conditions and a largest to smallest variance ratio of 16:1

J	Mean condition	Distribution	Pairing	PSHR	HBG	HTR	MTP	WMTP
3	All conditions	$g = 0, h = .5$	PP	19.2	18.5	19.2	24.7	23.5
			NP	17.8	16.8	17.8	28.2	26.5
		$g = .5, h = .5$	PP	13.5	13.6	13.5	19.1	16.5
			NP	19.6	18.7	19.6	32.3	31.5
	Complete null	$g = 0, h = .5$	PP	97.5	98.0	97.5	58.8	97.5
			NP	97.4	98.0	97.4	45.9	97.4
		$g = .5, h = .5$	PP	87.7	89.9	87.7	44.4	87.8
			NP	82.7	85.4	82.7	32.6	82.8
	Non-null	$g = 0, h = .5$	PP	6.1	5.3	6.1	19.1	11.2
			NP	4.5	3.3	4.5	25.3	14.7
		$g = .5, h = .5$	PP	1.1	0.9	1.1	14.9	4.6
			NP	9.0	7.5	9.0	32.3	22.9
4	All conditions	$g = 0, h = .5$	PP	8.8	8.8	8.6	17.8	15.7
			NP	8.3	8.3	8.2	19.7	18.1
		$g = .5, h = .5$	PP	7.1	7.3	7.0	13.4	10.5
			NP	8.9	9.9	8.8	19.4	19.4
	Complete null	$g = 0, h = .5$	PP	97.8	98.4	97.7	52.7	97.4
			NP	97.8	98.5	97.6	42.1	97.3
		$g = .5, h = .5$	PP	89.0	91.6	88.5	35.4	87.6
			NP	83.1	85.8	82.2	26.9	81.0
	Non-null	$g = 0, h = .5$	PP	1.4	1.2	1.3	14.8	8.9
			NP	0.8	0.8	0.7	17.8	11.5
		$g = .5, h = .5$	PP	0.2	0.2	0.2	11.6	4.1
			NP	1.5	1.5	1.5	18.6	13.2
5	All conditions	$g = 0, h = .5$	PP	4.9	4.9	4.8	9.8	9.2
			NP	4.8	4.8	4.7	10.5	10.8
		$g = .5, h = .5$	PP	4.3	4.4	4.3	7.3	6.3
			NP	4.4	4.4	4.3	10.9	11.8
	Complete null	$g = 0, h = .5$	PP	98.6	98.8	98.3	45.5	97.8
			NP	98.3	98.6	97.9	37.6	97.5
		$g = .5, h = .5$	PP	90.7	92.4	89.9	28.7	88.2
			NP	85.0	87.3	83.4	22.3	80.9
	Non-null	$g = 0, h = .5$	PP	0.2	0.2	0.2	8.4	4.8
			NP	0.1	0.1	0.1	9.2	6.5
		$g = .5, h = .5$	PP	0.0	0.0	0.0	6.2	2.2
			NP	0.4	0.4	0.3	10.3	8.3

Note: See Table 3.

comparison of the entries in this table with those presented in Tables 3, 4 and 5 reveals similar findings. Indeed, just about everything that we noted earlier holds here as well. That is, the true-model rates for the model testing methods (MTP and WMTP) were always larger than the stepwise MCPs for the all-conditions data and the model testing methods rates were always larger than the stepwise MCPs rates for the non-null mean configurations. In addition the unprotected model testing approach (MTP) was not as

successful as the protected version (WMTP) in correctly identifying the underlying true models, except in the non-null cases. One difference notable from Table 6 is that under the complete null case, all procedures were affected by extreme non-normality (very heavy-tailed skewed distribution $g = 0.5$ and $h = 0.5$); however, the MTP was most affected, and moreover, was also affected when data were obtained from the other extremely non-normal $g = 0$ and $h = 0.5$ (symmetric heavy-tailed) distribution.

7. Discussion

A number of MCPs have been proposed over the past few decades that purport to provide a better balance between Type I error control, power and ease of application, although each of the proposed procedures is plagued with a multiple comparison strategy that often results in intransitive decisions. Recently, Dayton (1998) proposed a model testing strategy for pairwise multiple comparisons testing that eliminates intransitive decisions and provides a more practical method of summarizing mean differences. The current study investigated how the MTP proposed by Dayton performed relative to three stepwise MCPs with respect to the proportion of times in which each procedure correctly selected the 'true' pattern of differences among a set of means (i.e., true-model rate). The true-model rates for Dayton's MTP were typically larger than the true-model rates for the protected Shaffer (1986), Hochberg (1988) or Hayter (1986) stepwise MCPs. In addition, a protected version of the MTP using Welch's (1951) omnibus test also had larger overall true-model rates relative to the stepwise procedures.

Under the complete null hypothesis the MTP performed poorly, which supports the findings of Dayton (1998) that the AIC statistic has a bias for selecting population models more complex than the true null model. However, the true-model rates for the WMTP were substantially larger than those for the MTP. This finding supports Dayton's recommendation that an omnibus test could be used with the MTP to screen for the complete null case, and in conditions where sample sizes and variances are unequal the use of the Welch (1951) omnibus test is generally recommended. Nonetheless, it could be argued that when all the population means are equal the model testing approach will too frequently not identify this configuration, based on the results we report in Tables 4 and 6. Accordingly, one might argue that the model testing approach has some difficulty identifying the so called 'null' case. Though this may be the case, we nonetheless believe, based on all the available data, that the approach merits serious consideration by data analysts. Specifically, the method is in general no more prone to problems associated with the 'null' case than the conventional piecemeal methods of analysis for pairwise comparisons. That is, it is probably safe to say that no one method of analysis will ever be discovered that works well under all data analysis scenarios. What we seek is a procedure that performs well in most situations.

With this caveat in mind, our results, along with those presented by Dayton (1998), clearly establish that the model testing approach much more frequently identifies the 'true' pattern of differences among a set of means than do classical methods of analysis. And moreover, researchers can get a holistic, rather than piecemeal, analysis of the data; that is, they can compare the competitiveness of various models through Dayton's approach to examining differences between treatment group means.

When all population means were not equal the MTP and WMTP had significantly higher true-model rates than any of the stepwise MCPs. Further, although the true-model

rates were larger for the MTP than for the WMTP (as expected), the true-model rates for the WMTP were consistently and significantly larger than the most powerful stepwise procedure (Hayter or Shaffer). Therefore, we are favourably impressed with the protected model testing approach to pairwise multiple comparison testing. However, what we have also discovered for the model testing approach is that for sample sizes that are representative of studies in psychology, the true-model rates for the model testing procedure, and even more so for the traditional MCPs, can be quite modest. One must remember, however, that identifying the 'true' pattern of mean differences is a very stringent criterion. For example, for $J = 5$ and the mean configuration [0.000 0.226 0.452 0.904 1.131] we found, through simulation, that one would need approximately 320 subjects per group to detect this pattern 80% of the time with the WMTP. However, the sample-size requirements would be even more demanding (460 per group) to achieve 80% all-pairs power (detecting all true pairwise differences, a standard identical to the true-model rate for this mean configuration) for the means just enumerated with a stepwise MCP, say Hochberg's (1988) approach.

Our results also suggest that the WMTP outperforms the conventional methods when variances are heterogeneous and/or if the data are non-normal in shape, except under the complete null case when data are extremely non-normal (i.e., under the two *g*-and-*h* distributions) when its performance is similar to the stepwise procedures. That is, for all procedures examined, the true-model rates under the null configuration were most deviant from .95 when data were very non-normal and variances were very heterogeneous. In this case it may be possible to use robust estimators (see Wilcox, 1997) with Dayton's (1998) model comparison approach.

The reader should also take note that this 'deficiency' of the model testing approach (as well as the stepwise MCPs) occurred under quite extreme cases of non-normality and variance heterogeneity, that is, under conditions that most likely do not typify the data obtained in most applied settings. These conditions were included, as correctly suggested by a reviewer, because they are intended to indicate the operating characteristics of procedures under extreme conditions, with the premise being that, if a procedure works under the most extreme of conditions, it is likely to work under most conditions likely to be encountered by researchers. What we have rediscovered is that there will always be (extreme) cases where a particular procedure will not perform well; alas, there still is no procedure that will work well in all conditions that may arise.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AG19*, 716–723.

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

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

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. *American Statistician*, *52*, 144–151.

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

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.

Hoaglin, D. C. (1985). Summarizing shape numerically: The *g*-and-*h* distributions. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds), *Exploring data tables, trends, and shapes* (pp. 461–513). New York: Wiley.

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

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

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350–386.

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.

Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, *78*, 479–485.

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

SAS Institute Inc. (1985). *SAS/STAT User's Guide, Version 6* (4th Ed.). Cary, NC: Author.

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

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

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error probabilities of the *t* test to departures from population normality. *Psychological Bulletin*, *111*, 352–360.

Scheffe, H. (1953). A method for judging all contrasts in analysis of variance. *Biometrika*, *40*, 87–104.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

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.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*, 751–754.

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

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.

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

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. New York: Academic Press.

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

Received 18 January 2001; revised version received 5 July 2002