A More Robust Procedure for Testing the Null Hypothesis in MANOVA

Chris N. Kladopoulos

Queens College of CUNY

And

Philip H. Ramsey

Queens College of CUNY

CHRIS N. KLADOPOULOS is Adjunct, Assistant Professor of Psychology at Queens College of the City University of New York, Flushing, NY 11367; Kladcuny@aol.com. His areas of specialization are learning theory, applied behavior analysis, applied statistics, and human psychophysics.

PHILIP H. RAMSEY is Professor of Psychology at Queens College of the City University of New York, Flushing, NY 11367; phramsey@qc1.qc.edu. His areas of specialization are applied statistics, psychometrics, time-series analysis, and computer simulation.

This work was supported in part by a grant from The City University of New York PSC-CUNY Research Award Program.

Abstract

Earlier research on Pillai's V statistic and Wilks' generalized correlation ratio statistic, U, suggest that a combination of these two procedures can be used to form a more robust test in MANOVA. The present investigation considers two methods of forming such a composite procedure. One of these procedures is shown to provide a robust alternative to existing methods for testing the full-null hypothesis for small samples while the other is shown to be superior for medium to large samples.

A More Robust Procedure for Testing the Null Hypothesis in MANOVA

The multivariate analysis of variance (MANVOA) has often been used in educational and behavioral research (Barling & Rosenbaum, 1986; Fassinger & Richie, 1994; Goldman & Harlow, 1993; Midgley, Feldlaufer, & Eccles, 1989;). The fixed-effects MANOVA in the one-way case has p response variables observed in k treatment groups of n experimental units per group. The full-null hypothesis asserts that the k population vectors of p means are equal. Numerous studies have investigated at least six procedures for testing the null hypothesis in MANOVA (O'Brien, Parente, & Schmitt, 1982; Olson, 1974; Hsu & Pillai, 1985). Perhaps the most extensive study is that of Olson (1974). He examined the robustness and power of six MANOVA procedures (stated below) under three values of k (3, 6, & 10), three values of n (5, 10, 50), and four values of p (2, 3, 6, & 10) for a total of 36 combinations of experimental conditions. Olson also considered three alpha levels (.10, .05, & .01) and numerous ways in which violations can occur.

If s = min (p, k - 1) is greater than one, no invariant test is uniformly most powerful. Let **H** (p x p) and **E** (p x p) be the sum-of-products matrices for hypothesis and error, respectively defined as

$$\mathbf{H} = \sum_{i=1}^{k} N_i (\overline{Y}_i - \overline{Y})(\overline{Y}_i - \overline{Y})', \quad (1)$$

$$\mathbf{E} = \sum_{i=1}^{k} \sum_{j=1}^{Ni} (\mathbf{Y}_{ij} - \overline{Y}_i)(\mathbf{Y}_{ij} - \overline{Y}_i)', \quad (2)$$

where $\mathbf{Y}_{ij}$ is the j^{th} of n_i observation vectors in group i, $\overline{Y}_i$ is the mean vector for the i^{th} group and $\overline{Y}$ is the grand mean vector. The product $\mathbf{HE}^{-1}$ has

order, p. The s nonzero eigenvalues (c_1, …, c_s) of $\mathbf{HE}^{-1}$ can be used to form a variety of tests of H_0. In some cases, the tests are expressed as the s nonzero eigenvalues (g_1, …, g_s) of $\mathbf{H(H + E)}^{-1}$. The present investigation is limited to the equal n case in which $n = n_1 = \ldots = n_k$. The population values of the (c_1, …, c_s) values are designated as (λ_1, …, λ_s).

Olson (1974) examined four common MANOVA tests of the full, null hypothesis. The Pillai-Bartlett trace, $V = \Sigma g_m$. Wilks' likelihood ratio, $W = \Pi(1 - g_m)$. The Lawley-Hotelling trace, $T = \Sigma c_m$. Roy's greatest, characteristic root, $R = g_1$. In addition, Olson (1974) investigated a less frequently considered procedure, $U = \Pi g_m$. The U statistic was proposed by Wilks (1932) and designated as a generalization of the correlation ratio. Olson (1974) attributed the U statistic to Gnanadesikan who briefly mentioned it in a paper in 1964 and an abstract in 1965.

It is customary to write the s values of λ_i in order with λ_1 the largest and λ_s the smallest. In the extreme case of a concentrated structure, λ_1 will be greater than zero and all over values will be zero. Previous research (Ito, 1962; Lee, 1971, Olson, 1974, Roy, 1966) indicates that the probability of rejecting a false, full-null hypothesis decreases as λ_i varies from a single nonzero λ_1 to the set of $\lambda_1 = \ldots = \lambda_s$ and the power of the test varies from $R \geq T \geq W \geq V \geq U$. In the reverse condition, when the order of the pattern of λ_i varies from all nonzero values, $\lambda_1 = \ldots = \lambda_s$, to a single nonzero λ_1, the power of the tests also reverses to $R \leq T \leq W \leq V \leq U$.

In comparison to R, T, W, and U, Olson (1974) reported V to provide, "protection against nonnormality and heterogeneity of covariance matrices" (p. 894). He found R to be particularly poor in such protection. However,

there were many cases reported by Olson (1974) which show rather high Type I error rates even for V. That is, Olson only found V to be relatively robust in comparison to the other five procedures he investigated.

Absolute cutoff values for acceptable Type I error rates were proposed by Bradley (1978) who suggested that no statistical test should be consider robust when applied under conditions in which the actual probability of a Type I error exceeds 1.5α. For $\alpha = .05$ that would be an upper limit of .075. One argument in support of Bradley's upper limit is that any value greater than .075 is closer to .10 than .05. Thus, a true rate above .075 is more accurately described as being approximately .10 rather than approximately .05.

The present investigation attempts to establish absolute robustness of MANOVA tests using Bradley's 1.5α criterion. Olson (1974) examined the relative robustness of MANOVA tests. The present investigation examines all experimental conditions considered by Olson (1974) and some additional conditions as well. It examines R, T, W, V and U considered by Olson and some additional procedures.

Significance Testing Procedures

The Pillai-Bartlett trace statistic, V, can be evaluated with an F test Pillai (1955) and Seber, 1984, p. 564) defined by

$$F = \frac{cV}{b(s - V)}, \tag{3}$$

where $s = \min(p, \underline{k} - 1)$,

$c = df_E - p + \underline{s}$,

and $b = \max(p, \underline{k} - 1)$.

To test at level α requires critical value, $CV = F_{1-\alpha}(sb, sc)$. The error df value is $df_E = \Sigma(n_i - 1) = k(n - 1)$. This is the method used by Olson (1974) and generally associated with the V statistic.

A more accurate F test for V is available due to Muller (1998). Two methods are presented. Method 1 is shown to be most effective for testing $\alpha > .01$ and Method 2 for $\alpha \leq .01$. In Method 1 we have

$$F = \frac{df_2}{df_1} \frac{V}{d - V}, \tag{4}$$

where $df_1 = \underline{p}(\underline{k} - 1)$,

$$df_2 = \frac{[p(k-1)+2]df_E(df_E + k - 1 - p)}{df_E(k+p) + (k+1)(k-2)},$$

and $\underline{d} = \frac{p(k-1) + df_2}{df_E + k - 1}$.

To test at level α requires $CV = F_{1-\alpha}(df_1, df_2)$. The value of df_E is $k(n - 1)$ as above. For Method 2 Muller (1998) the $\underline{F}$ test is

$$F = \frac{df_2}{df_1} \frac{V}{s - V}, \tag{5}$$

where $K = \frac{1}{s(df_E + k - 1)} \left[\frac{s(df_E + s - p)(df_E + k + 1)(df_E + k - 2)}{df_E(df_E + k - 1 - p)} - 2 \right]$

$df_1 = p(k - 1)K$,

$c = df_E - p + s$,

and $df_2 = scK$.

To test at level α requires $CV = F_{1-\alpha}(df_1, df_2)$. The value of df_E is again $k(n - 1)$. As noted above, applying Muller's Method 1 and Method 2 to the testing of the Pillai-Bartlett trace, V, at appropriate alpha levels is designated here as the Pillai-Bartlett-Muller procedure, M.

Wilks' likelihood ratio statistic, W, can also be applied to the MANOVA null hypothesis. One method of evaluating W for a group of k mean vectors is with an F test (Rao, 1951; Seber, 1984, p. 41) defined by

$$F = \frac{1-X}{X}\ \frac{df_2}{df_1}, \tag{6}$$

where $t = \sqrt{\frac{p^2(df_H)^2 - 4}{p^2 + (df_H)^2 - 5}}$

df_H

$f = df_E - (p - df_H + 1)/2,$

$g = \frac{p(df_H) - 2}{2},$

$df_1 = p(df_H),$

$df_2 = ft - g,$

and $X = W^{1/t}$.

To test at level α requires $CV = F_{1-\alpha}(df_1, df_2)$. The value of $df_E = k(n - 1)$ and $df_H = k - 1$. It can be shown that Eq. (6) provides an exact F test for p = 1, 2 or k = 2, 3 (Seber, 1984, pp. 40-41) provided the usual assumptions are satisfied. Applying Wilks' W is designated here as the Wilks' likelihood ratio procedure, W.

Wilks' (1932) generalization of the correlation ratio, $U = \Pi\, g_m$, presents a special case for significance testing. Olson (1974) generated critical values for U using Monte Carlo simulation with 1,000 replications. The F ratio for the W statistic used to test p dependent variables and parameters df_H and df_E is shown in Equation 6. Fujikoshi (1972) observed that U has the same distribution as W with the degrees of freedom values (df_E and df_H) reversed. Thus, the needed critical values of U come from the opposite tail of the

distribution. Equation 6 was used to produce estimated critical values for U which were found to be accurate provided $p \leq k - 1$. However, in other cases, this method presented a particular problem in generating CVs for some combinations of df_E and df_H. In those cases, it was necessary to use Monte Carlo simulation to produce accurate critical values. That simulation was based on 20,000 replications.

The Lawley-Hotelling trace, T, can be used to test the MANOVA hypothesis. One method of evaluating T for a group of k mean vectors is with an F test (McKeon, 1974; Seber, 1984, p. 39) defined by

$$F = \frac{T}{c} \tag{7}$$

where $a = p(k - 1)$,

$$B = \frac{(df_E + k - p - 2)(df_E - 1)}{(df_E - p - 3)(df_E - p)},$$

$$b = 4 + \frac{a + 2}{B - 1},$$

and $c = \frac{a(b - 2)}{b(df_E - p - 1)}$.

To test at level α requires $CV = F_{1-\alpha}(a, b)$. The value of $df_E = k(n - 1)$. Applying the Lawley-Hotelling trace procedure is designated here as T.

Roy's greatest characteristic root, g_1, can also be applied to the MANOVA null hypothesis. Tables of critical values for g_1 are available (Harris, 2001, pp. 518-531 & Sever, 1984, pp. 593-598). Computer routines for evaluating g_1 are also available (Harris, 2001). Testing g_1 is designated here as the greatest characteristic root procedure, R.

One MANOVA procedure has been developed that controls for Type 1 errors using Bonferroni testing in which a univariate procedure is applied to each dependent variable at level α/p. This procedure generally provides adequate control of Type I errors Bird (1975). This procedure allows researchers to determine the nature of the differences once H_0 is rejected. Pairwise testing applying Tukey's HSD to each dependent variable with the Bonferroni adjustment has been shown to be more powerful than some alternatives (Bird, 1975; Ramsey, 1982). The Tukey-Bonferroni combination can be used to test the full, null hypothesis. That is, if any pair of means is significantly different on any of the p dependent variables at level α/p by the HSD critical difference based on the studentized range distribution then the full, multivariate null can be rejected at level α. That procedure is designated in the present investigation as Q.

An alternative and possibly more powerful Bonferroni method could be obtained by applying the univariate ANOVA F test to each dependent variable at level α/p. That approach is designated here as P. A more robust method would be to test each dependent variable at α/p using the Welsh (1951) procedure which does not assume equal population variances. This method is designated here as B.

Although Olson (1974) generally found V to be the most robust of the MANOVA procedures, a careful examination of the results reported by Olson shows an interesting pattern. The V procedure is usually somewhat liberal producing rejection rates above the nominal level and occasionally conservative with rejection rates below the nominal level. The U procedure is generally the reverse in the same conditions. That is, U is usually liberal

where V is conservative and vice versa. This suggests that some combination of the U and V statistic might produce a more robust procedure.

Two methods are considered here for combining the U and V statistics. However, the V statistic is evaluated by the Pillai-Bartlett-Muller procedure Muller (1998) rather than the less accurate method used by Olson (1974) and others. The first method is designated as UM1 and is applied by testing the MANOVA null hypothesis with both U and M and rejecting that hypothesis if either test is significant at, α_1, a somewhat more stringent value of α.

Determination of, α_1, the alpha level for UM1 using a strict Bonferroni approach would suggest applying each test at .5α. However, empirical evaluation using Monte Carlo simulation showed such an approach to produce a very conservative procedure. Testing all the 36 experimental conditions considered by Olson (1974) in cases where all assumptions were satisfied but p, k, and n were varied, showed that the maximum rejection rate for M tested at α = .05 was .0645. Testing UM1 at α_1 = .8α (i.e., .04 for α = .05) produced a maximum rejection rate over the 36 conditions of .068. Thus, UM1 was taken as rejecting the null if either M or U was significant at α_1 = .8α. Similar results were found for α of .10 and .01.

The second method, UM2, was also based on testing both U and M at an adjusted α level, α_2. In UM2 the MANOVA null hypothesis would be rejected at level α if both U and M are significant at some level, α_2, above α. Applying UM2 at 1.7α (i.e., 1.7[.05] = .085) was found to produce a maximum rejection rate over the 36 conditions of .0682. Thus, UM2 was designated as rejecting the null if both U and M were significant at α_2 = 1.7α. Again similar results

were found for .10 and .01 levels. Critical values for UM1 and UM2 were similarly determined separately from U and M critical values.

Evaluation

A Monte Carlo simulation was conducted to evaluate the accuracy of all 11 procedures (V, W, T, R, Q, P, B, M, UM1, UM2, & U) for controlling Type I errors with and without assumption failure. Random normal deviates were generated by the method of Press, Teukolsky, Vetterling, and Flannery (1994). Assumption failure was examined by varying the degree of contamination in the groups, d, and how contamination varies across groups (concentration of contamination). Values of d included 1 (no contamination) and all the values used by Olson (1974): 4, 9, 16, and 36.

Multivariate data were generated in the same manner as was done by Olson (1974). Two normal distributions were combined from N(**0, I**) and N(**0, D**) where **D** could have two forms. For the low concentration of contamination condition **D** = d**I** with scalar d. In this case the contamination parameter, d, would be the same on all p variables. For the high concentration of contamination **D** = **C**(d) = diag (pd – p + 1, 1, 1, …). In this condition the contamination is exclusively on the first variable. In both cases the trace of **D** is pd. That is, the degree of contamination is the same and determined completely by the parameter, d.

Also following Olson (1974) the form of contamination was determined by a k-tuple (a_1, …, a_k) where each a_i ranges from 0 to 1 and represents the proportion of observations in group i which include contaminated data. Of course, (0, …, 0) represents the condition of no contamination across groups. Again following Olson (1974) the first form of contamination considered was

(1, 0, 0, …, 0). In this form all the contamination is in the first group. A second form of contamination is (.2, .2, …, .2) in which there is 20% contamination in all groups. A third form is (.2, 0, 0, …, 0). In this case there is 20% contamination but only in the first group.

All the conditions investigated by Olson (1974) were included plus some additional conditions. In particular, Olson did not include all 36 combinations of experimental conditions in all possible contamination conditions. All such combinations were included here. Each simulated experiment was replicated 10,000 times. Two methods were used to evaluate the statistical significance of the empirical rejection rates. For rejection rates between 0.0 and 1.0 the standard error (SE) depends on the value of the rate. If x is the proportion of replications exceeding a critical value, the SE is $[x(1 – x)/10000]^{1/2}$. For x = .5 the SE would be a maximum and have a value, SE = $\sqrt{.000025}$ = .005 so a 50% rejection rate would be included in a 2SE interval from .49 to .51 in approximately 95% of the simulations. An x of .05 would have SE = $\sqrt{.00000475}$ = .002179 and a 2SE interval from .045641 to .054358. Thus rates even as small as 5% will usually be estimated to differ from the correct value by less than .0044.

McNemar's (1947) test of correlated proportions was used to test the significance of the difference between proportions as rejection rates. Applying McNemar's test to all pairs of procedures is impractical and risks excessive testing. The present approach is to identify a specific reference procedure against which to test each of the 11 procedures. In particular, the procedures were considered in the order: V, W, T, R, Q, P, B, M, UM1, UM2, & U. Each method was tested against the following method (i.e. the method

to the right in the list) as a reference procedure with the final method, U, tested against the first, V. Thus, each method was tested for significance against the method on each side of it.

Results

Table 1 presents the Type I error rates for all 11 procedures for the no contamination condition (d = 1). All rates should not exceed .05 except by sampling error. That is, based on sampling error, 97.5% of the simulations should be less than .0544 when the true rate is .05. Many significant differences are shown in Table 1 when comparing one procedure to another using McNemar's test. For example, in the first row with n = 5, k = 3, and p = 2 the rate for U is .0493 which is significantly greater than the V rate of .0366. That rate for V is also below the 95% lower bound of .045641. Clearly, V is conservative in this condition. In fact, V is often conservative and has a mean rate in Table 1 of .04411. Averaging 36 rates produces a mean rate based on 36,0000 samples. The lower bound on a 95% interval of a nominal .05 rate would be .04927. The mean rate for V is well below that lower bound. That is, the mean rate for V in Table 1 shows that it is very conservative and, at least on the average, conservative over all 36 conditions.

Other procedures also have a mean rejection rate below .05. However, the Q, P, and U procedures all have mean rates above the .04927 lower bound. These three mean rates can be considered within sampling error of the nominal .05 level. The procedures B, UM1, and UM2 all have mean rates below the .04927 lower bound. These procedures appear to be slightly conservative.

Table 1.

Type I Errors for All Eleven Procedures Tested at $\alpha = .05$ with All Assumptions Are Satisfied

n	k	p	V	W	T	R	Q	P	B	M	UM1	UM2	U
5	3	2	.0366†	.0504†	.0482	.0476	.0462	.0474†	.0351†	.0495†	.0425†	.0317†	.0493†
		3	.0334†	.0478	.0467*	.0492	.0494	.0488†	.0348†	.0451†	.0535†	.0383†	.0639†
		6	.0370†	.0494*	.0465	.0480	.0514	.0507†	.0342†	.0534†	.0388†	.0489	.0486†
		10	.0444†	.0547	.0541†	.0658†	.0482	.0483†	.0312†	.0698†	.0402†	.0567†	.0496†
	6	2	.0410†	.0477	.0484	.0485	.0458	.0468†	.0346†	.0482†	.0432†	.0542†	.0506†
		3	.0404†	.0490	.0474	.0487	.0474	.0475†	.0323†	.0504†	.0422	.0397†	.0466†
		6	.0343†	.0536†	.0498	.0479	.0492	.0512†	.0320†	.0531*	.0481†	.0253†	.0544†
		10	.0351†	.0520†	.0472	.0475	.0483	.0488†	.0268†	.0517†	.0419†	.0370†	.0501†
	10	2	.0478†	.0517	.0504	.0509	.0498	.0516†	.0387†	.0517†	.0468†	.0655†	.0520†
		3	.0417†	.0465	.0463	.0487	.0461	.0443†	.0298†	.0480†	.0439†	.0518†	.0469†
		6	.0371†	.0502	.0498	.0484	.0478	.0477†	.0304†	.0501†	.0422†	.0334†	.0438†
		10	.0424†	.0490	.0482†	.0547*	.0493*	.0536†	.0469	.0485	.0467†	.0369*	.0406
10	3	2	.0345†	.0510	.0502	.0488	.0465	.0471†	.0271†	.0538†	.0328†	.0185†	.0302*
		3	.0467†	.0519	.0520	.0537	.0527	.0517†	.0490	.0523	.0542†	.0405†	.0486
		6	.0460†	.0529	.0531	.0535*	.0486	.0497†	.0460†	.0532*	.0568†	.0522†	.0624†
		10	.0436†	.0495	.0498	.0485	.0473	.0472†	.0431†	.0498†	.0432†	.0574†	.0483†
	6	2	.0423†	.0508	.0494	.0513	.0486	.0482†	.0422*	.0485†	.0397†	.0623†	.0485†
		3	.0474†	.0502	.0508	.0515	.0478	.0504†	.047	.0497	.0472†	.0565†	.0498
		6	.0459†	.0498†	.0516	.0514	.0498	.0513†	.0467	.0505	.0503†	.0455	.0481
		10	.0371†	.0461	.0470	.0498	.049	.0496†	.0446	.0463	.0442†	.0275†	.0403
	10	2	.0418†	.0504	.0510	.0511	.0467*	.0506†	.0432*	.0494†	.0442	.0424	.0450
		3	.0477†	.0500	.0502	.0512	.049	.0512†	.0476	.0494	.0475†	.0654†	.0486
		6	.0492†	.0533	.0540	.0509	.0549	.0527†	.0480	.0524	.0525†	.0598†	.0535*
		10	.0431†	.0480	.0478	.0490	.0487	.0489†	.0421*	.0477	.0451†	.0357†	.0408
50	3	2	.0463†	.0545	.0542	.0536*	.0468	.0486†	.0413†	.0545†	.0617†	.0315†	.0558†

(Table 1 continued)

	3		.0486	.0493	.0493	.0483	.0486	.0473	.0473	.0491†	.0578†	.0425†	.0487
	6		.0533*	.0541	.0541	.0538	.0538	.0533	.0532	.0538†	.0670†	.0568†	.0659†
	10		.0500*	.0508	.0504	.0521	.0513	.0494	.0494	.0510†	.0582†	.0651†	.0609†
6	2		.0514†	.0536	.0546	.0526	.0496	.0484	.0483*	.0534†	.0498†	.0687†	.0518
	3		.0489	.0493	.0494	.0511	.0517	.0489	.0487	.0493	.0507†	.0561†	.0465
	6		.0504*	.0510	.0511	.0496	.0506	.0523	.0522	.0511†	.0555†	.0502	.0492
	10		.0505†	.0524	.0526	.0523	.0491	.0523	.0522	.0526†	.0615†	.0372†	.0440†
10	2		.0478†	.0491	.0492	.0513	.0512	.0507	.0506	.0490†	.0566†	.0508	.0526*
	3		.0477	.0480	.0480*	.0518	.0518	.0519	.0519	.0481	.0467†	.0643†	.0483
	6		.0503	.0507	.0511	.0509	.0518	.0535	.0533	.0505†	.0540†	.0601†	.0511
	10		.0462†	.0471	.0475	.0507	.0503	.0524	.0524*	.0472†	.0525†	.0405	.0388†
Mean			.04411	.05044	.05004	.05096	.04931	.04984	.04262	.05089	.04888	.04741	.04928
SD			.00555	.00224	.00240	.00322	.00218	.00224	.00825	.00397	.00748	.01304	.00701
MAX			.0533	.0547	.0546	.0658	.0549	.0536	.0533	.0698	.0670	.0687	.0659
MIN		.0334	.0461	.0463	.0475	.0458	.0443	.0268	.0451	.0328	.0185	.0302	

*Rate is significantly different from the rate to the right at the .05 level by McNemar' test.
†Rate is significantly different from the rate to the right at the .01 level by McNemar' test.
Note: The * or † beside U indicates it is significantly different from V.

Only B has a mean rate below that of V. Only R and M have a mean rejection rate above the upper limit, .050726, of the 95% interval. This might suggest that R and M are slightly liberal rejecting slightly too often.

In addition to advocating a maximum upper limit of 1.5α for robustness, Bradley (1978) also suggested that rates in the interval of $\pm 1.1\,\alpha$ (i.e., .045 to .055 for $\alpha = .05$) represent negligible non-robustness. In Table 1 only V and B have mean Type I rejection rates outside that interval. The B procedure is based on a method that approximates rates when variances are unequal and could be expected to be somewhat different from the nominal level.

Both UM1 and UM2 have mean rejection rates below .05. The maximum rate for both procedures is less than the upper limit of alpha (.0544) and less than that for M. Thus, UM1 and UM2 seem to provide adequate control of Type I errors.

Table 1 also provides maximum and minimum rates for each of the 11 procedures taken over the 36 conditions. UM1 and UM2 both have maximum rates slightly lower than that of M and slightly above that of U. The lowest maximum is found for V and B. Thus, UM1 and UM2 have Type I error control that is comparable to that of the other procedures. Indeed, UM1 and UM2 have better control than V and B.

Table 2 gives the maximum rejection rates for all 11 procedures at various values of d when 100% contamination is in the first group for all variables. The maximum rates at each n in Table 1 are shown in Table 2 in the lines with d = 1. Corresponding maximum rates for other values of d are also shown in Table 2. For n = 5, V exceeds Bradley's robustness limit of .075 for d between 4 and 5. Procedures W, T, R, Q, P, B, M, and UM2 all exceed .075

for d between 1 and 4. M tests the same test statistic as V but uses a more accurate critical value. As shown in Table 1, M has a Type I error rate that is closer to the nominal value than does V. While these results agree with those of Olson (1974), they also show that part of the "robustness" of V was due to its conservative Type I error rate. UM1 limits the rate to .075 for $1 \leq d < 7$ without having a conservative rate at d = 1.

For $n \geq 10$ Table 2 shows UM2 to maintain the Type I error rate to or below .075 for all values of d. For a true rate of .075 the SE for 10,000 replications would be .002634. The upper limit of a 95% interval would be .08027. The rejection rate of .0761 for UM2 at n = 10 and d = 5 is below .08027 and therefore approximately equal to .075. Nevertheless, the true rejection rate for UM2 may slightly exceed the .075 upper limit for values of d near 4 or 5. U provides similar control but becomes unduly conservative for large values of d. For n = 50, UM2 limits all rates to .075 but becomes conservative at d = 36.

In general, the results in Table 2 suggest that for small n (i.e., $\underline{n} < 10$) and moderate values of d (i.e. $\underline{d} < 7$) UM1 is a robust procedure. For $n \geq 10$ and d not too large (i.e., $d < 36$) UM2 is robust by the .075 criterion or very nearly so. Both UM1 and UM2 are more robust than V while not being as conservative.

Table 3 gives the maximum rejection rates for all 11 procedures when 20% contamination is in all variables and in all groups and for various values of d. As in Table 2, the maximum rates at each n in Table 1 are shown in Table 3 in the lines with d = 1. UM1 is somewhat more robust in Table 3 than in Table2 with rates below .075 for n = 5 and $d \leq 9$.

Table 2.

Maximum Type I Errors Over all 36 Experimental Conditions for All Eleven Procedures Tested at $\alpha = .05$ with Violations of the Form (1, 0,...,0) and Concentration $d\mathbf{I}$ for Various Values of d.

n	d	V	W	T	R	Q	P	B	M	UM1	UM2	U
5	1	.0478	.0547	.0541	.0658	.0514	.0536	.0469	.0698	.0535	.0655	.0639
	4	.0708	.0993	.1511	.1989	.1788	.1359	.1081	.0960	.0641	.0782	.0682
	5	.0783	.1224	.2041	.2639	.2237	.1731	.1340	.1050	.0693	.0867	.0761
	6	.0814	.1492	.2560	.3299	.2659	.2103	.1568	.1121	.0695	.0937	.0839
	7	.0964	.1773	.3170	.4099	.3020	.2377	.1775	.1204	.0812	.1011	.0887
	9	.1096	.2354	.4321	.5267	.3578	.3007	.2073	.1280	.0926	.1092	.0961
	16	.1388	.4097	.6867	.7782	.4724	.4153	.2601	.1574	.1208	.1362	.1224
	36	.1882	.7324	.9191	.9550	.5833	.5336	.2871	.1934	.1660	.1594	.1398
10	1	.0492	.0533	.0540	.0537	.0549	.0527	.0490	.0538	.0568	.0654	.0624
	4	.0743	.1085	.1414	.2686	.1945	.1512	.1221	.0763	.0670	.0754	.0560
	5	.0858	.1352	.1946	.3573	.2602	.1897	.1491	.0875	.0764	.0761	.0551
	6	.0952	.1569	.2586	.4359	.3204	.2336	.1670	.0970	.0843	.0714	.0495
	7	.1026	.1868	.3131	.5038	.3712	.2771	.1837	.1041	.0919	.0756	.0516
	9	.1146	.2428	.4305	.6281	.4485	.3515	.2087	.1189	.0980	.0628	.0424
	16	.1514	.4068	.6933	.8541	.6043	.5136	.2894	.1561	.1358	.0541	.0354
	36	.1957	.7242	.9230	.9741	.7220	.6600	.3145	.2002	.1776	.0391	.0246
50	1	.0533	.0545	.0546	.0538	.0538	.0535	.0533	.0545	.0670	.0687	.0659
	4	.0779	.0941	.1278	.2901	.2099	.1538	.1445	.0798	.0700	.0679	.0488
	5	.0903	.1156	.1665	.3934	.2732	.1975	.1803	.0924	.0801	.0682	.0490
	6	.1004	.1343	.2059	.4900	.3392	.2448	.2040	.1005	.0909	.0707	.0508
	7	.1121	.1595	.2501	.5581	.3734	.2795	.2305	.1138	.0969	.0665	.0464
	9	.1359	.2035	.3248	.6793	.454	.3488	.2705	.1380	.1188	.0668	.0445

(Table 2 continues)

16	.1660	.3307	.5234	.8557	.5932	.5006	.3318	.1676	.1489	.0448	.0281
36	.2009	.5437	.7469	.9562	.6875	.6215	.3900	.2024	.1867	.0111	.0061

Table 3.

Maximum Type I Errors Over all 36 Experimental Conditions for All Eleven Procedures Tested at $\alpha = .05$ with Violations of the Form (.2, .2, …,.2) and Concentration d_I for Various Values of $\underline{d}$.

n	d	V	W	T	R	Q	P	B	M	UM1	UM2	U
5	1	0.0478	0.0547	0.0541	0.0658	0.0514	0.0536	0.0469	0.0698	0.0535	0.0655	0.0639
	4	0.0401	0.0470	0.0485	0.0587	0.0441	0.0423	0.0329	0.0598	0.0577	0.0561	0.0682
	9	0.0345	0.0404	0.0427	0.0541	0.0268	0.0271	0.0187	0.0539	0.0718	0.0447	0.0868
	16	0.0297	0.0403	0.0439	0.0541	0.0196	0.0157	0.0099	0.0484	0.0982	0.0401	0.1214
	36	0.0304	0.0419	0.0439	0.0526	0.0076	0.0051	0.0030	0.0514	0.1601	0.0414	0.1996
10	1	0.0492	0.0533	0.0540	0.0537	0.0549	0.0527	0.0490	0.0538	0.0568	0.0654	0.0624
	4	0.0439	0.0450	0.0449	0.0442	0.0467	0.0450	0.0415	0.0464	0.0523	0.0632	0.0575
	9	0.0354	0.0372	0.0355	0.0322	0.0367	0.0356	0.0313	0.0384	0.0803	0.0585	0.0932
	16	0.0271	0.0274	0.0255	0.0245	0.0278	0.0253	0.0210	0.0291	0.1484	0.0455	0.1758
	36	0.0177	0.0178	0.0174	0.0183	0.0147	0.0126	0.0096	0.0201	0.3285	0.0387	0.3693
50	1	0.0533	0.0545	0.0546	0.0538	0.0538	0.0535	0.0533	0.0545	0.0670	0.0687	0.0659
	4	0.0516	0.0517	0.0518	0.0517	0.0533	0.0513	0.0511	0.0517	0.0624	0.0697	0.0640
	9	0.0484	0.0491	0.0492	0.0487	0.0493	0.0486	0.0482	0.0493	0.0780	0.0647	0.0858
	16	0.0456	0.0454	0.0452	0.0446	0.0472	0.047	0.0468	0.0461	0.1028	0.062	0.1198
	36	0.046	0.0467	0.0468	0.0416	0.0476	0.0433	0.0426	0.0468	0.1471	0.0642	0.1760

UM2 is also more robust in Table 3 than in Table 2 with rates no higher than .0697 for all values of n and d. Table 4 gives the maximum rejection rates for all 11 procedures when contamination is only in the first group and in all variables (20% contamination) and when d = 36. V, M, and UM1 are all robust for n of 5 and 10 but nonrobust for n = 50. UM2 is robust for all values of n.

Table 5 gives the maximum rejection rates for all 11 procedures when contamination is also only in the first group, but only in the first variable (100% contamination) and for various values of d. The maximum rates at each n in Table 1 are shown in Table 5 in the lines with d = 1. For n = 5, UM1 is robust only for $d \leq 4$ rather than the less restrictive limit of $d < 7$ shown in Table 2. Unlike Table 2 results, UM2 is not robust for any values of n and d in Table 5. However, UM1 and UM2 are almost always more robust than V in Table 5. As noted in Table 1, UM1 and UM2 are less conservative than V and thus the relative robustness of each is almost certainly better than that of V in Table 5.

Table 6 gives the maximum rejection rates for all 11 procedures when contamination is in all groups, but only in the first variable (20% contamination), and when d = 36. All procedures are robust in this condition.

Table 4.

Maximum Type I Errors Over all 36 Experimental Conditions for All Eleven Procedures Tested at $\alpha = .05$ with Violations of the Form (.2, .0, …,.0) and Concentration d**I** for d = 36.

n	V	W	T	R	Q	P	B	M	UM1	UM2	U
5	.0386	.0482	.0508	.0683	.1346	.0687	.0354	.0585	.0451	.0557	.0560
10	.0510	.0549	.0577	.0713	.1333	.0708	.0404	.0528	.0473	.0626	.0452
50	.1051	.1066	.1078	.3936	.3021	.2218	.1980	.1054	.0939	.0659	.0453

Table 5.

Maximum Type I Errors Over all 36 Experimental Conditions for All Eleven Procedures Tested at $\alpha = .05$ with Violations of the Form (1, 0,…,0) and Concentration **C**(d) for Various Values of d.

n	d	V	W	T	R	Q	P	B	M	UM1	UM2	U
5	1	.0478	.0547	.0541	.0658	.0514	.0536	.0469	.0698	.0535	.0655	.0639
	4	.0793	.0891	.0958	.1045	.1472	.1261	.0897	.0845	.0741	.0883	.0782
	9	.1124	.1214	.1270	.1311	.1744	.1563	.0884	.1177	.1060	.1123	.0941
	16	.1371	.1472	.1513	.1591	.1988	.1729	.0904	.1421	.1297	.1338	.1095
	36	.1539	.1653	.1682	.1740	.2070	.1869	.0816	.1582	.1432	.1444	.1211
10	1	.0492	.0533	.0540	.0537	.0549	.0527	.0490	.0538	.0568	.0654	.0624
	4	.0816	.0865	.0908	.0978	.1564	.1384	.0846	.0834	.0784	.0898	.0703
	9	.1090	.1165	.1184	.1254	.1724	.1616	.0986	.1127	.1044	.1081	.0861
	16	.1338	.1374	.1376	.1473	.1870	.1645	.1050	.1354	.1277	.1284	.1071
	36	.1506	.1542	.1561	.1613	.1934	.1782	.1029	.1523	.1447	.1418	.1182
50	1	.0533	.0545	.0546	.0538	.0538	.0535	.0533	.0545	.0670	.0687	.0659
	4	.0841	.0851	.0870	.0991	.1438	.1281	.1061	.0842	.0818	.0860	.0683
	9	.1077	.1083	.1091	.1219	.1635	.1521	.1262	.1083	.1043	.1077	.0856
	16	.1227	.1238	.1242	.1326	.1720	.1514	.1325	.1232	.1189	.1141	.0973
	36	.1374	.1379	.1382	.1498	.1777	.1610	.1433	.1376	.1326	.1328	.1116

Table 6.

Maximum Type I Errors Over all 36 Experimental Conditions for All Eleven Procedures Tested at $\alpha = .05$ with Violations of the Form (.2, .2,…,.2) and Concentration $\mathbf{C}(d)$ for $d = 36$.

n	V	W	T	R	Q	P	B	M	UM1	UM2	U
5	.0388	.0494	.0514	.0658	.0490	.0486	.0335	.0634	.0444	.0498	.0522
10	.0437	.0510	.0500	.0496	.0478	.0466	.0419	.0505	.0458	.0617	.0553
50	.0494	.0512	.0516	.0528	.0500	.0498	.0496	.0502	.0585	.0660	.0570

Discussion

The present results were generally consistent with the findings of Olson (1974) for the same MANOVA procedures, and revealed no unexpected findings under the values of k, p ad n not tested by Olson. In all the above results except Table 5, UM1 and UM2 offer clear advantages over V or M. As noted previously, much of the apparent "robustness" of V is due to its conservative rejection rate when assumptions are satisfied. The M procedure provides a more realistic evaluation of the robustness of the Pillai-Bartlett statistic.

In most cases UM1 is a robust procedure for small n and UM2 for moderate to large n. Even in Table 5 that pattern holds except that neither UM1 nor UM2 meets the 1.5 α criterion for values of d much different from 1.0. However, even in Table 5, UM1 has better relative robustness than M for n = 5. That is, UM1 has lower rates than M for all values of d with n = 5. Also UM2 usually has lower rates than M for n of 10 and 50.

Table 5 requires special attention because it appears to contradict the results reported by Olson (1974). In particular, Table 5 shows the maximum Type I error rate for d = 36 and n of 5, 10, and 50 to be .1539, .1506, and .1374, respectively. For a nominal level of α = .05 those three rates are all approximately 3α. In the lower half of Figure E, Olson (1974) presents results for 16 of the possible 36 conditions in which V appears to have rejection rates consistently between α and 2 α. Olson's results are for α = .10 but he reported no differences due to alpha level.

Examination of the original raw data in Olson (1973) shows the actual 16 Type I error rates to range from .117 to .192 for a maximum of 1.92 α. In

fact, eight of the 16 rates were below .150 satisfying Bradley's 1.5α limit. However, in the same 16 conditions for α = .05 Olson (1973) reports the maximum rate to be .156 or 3.12 α. Also at α = .01 Olson (1973) reports the maximum rate to be .093 or 9.3 α. In this part of Olson's study α did make a difference with V being less robust for more stringent alpha levels.

It can also be seen both in Figure E (Olson, 1974) and the raw data (Olson 1973) that U is only moderately more conservative than V. Thus, neither UM1 nor UM2 is much better in Table 5 than V or M.

The procedures, R, Q, P, and B, are particularly attractive because they offer natural ways of following a rejection of the multivariate null hypothesis with additional tests of specific differences. However, their poor robustness raises doubts about their accuracy in such follow up testing.

Although the present investigation did not examine power rates, some observations are relevant. Olson (1974) reported power rates as the rejection under various nonnull conditions. However, just as V appears to be more robust due to its conservative rejection rate other procedures may appear to be more powerful due to their inflated rejection rates. For example, Olson (1974) reported W and T to be more powerful than V for certain conditions with d > 1 but did not control for the excessively high Type I error rates of W and T under those condition. A proper study of power requires comparing the rejection rates of procedures that have at least comparable control of Type I errors.

Stevens (1979) suggested that Olson (1974) had overstated the advantage of V over other procedures such as W and T by examining conditions rarely found in actual research. Stevens particularly questioned the likelihood of

finding d values as extreme as 36 to 1. He reported some data sets and a survey of several journals to support his accusation.

Olson (1979) responded to Stevens and argued that many of the conditions were quite common. However, Olson only identified d values of 2.2 and 1.9. If one accepts the arguments of Stevens (1979) that extreme results are not very common then UM1 and UM2 may well be robust by the 1.5α criterion in most research settings. Even in more extreme conditions UM1 and UM2 are likely to provide better control of Type I errors that the alternatives considered here.

Finn (1974, p. 12 & p. 333) reports a study in which four groups were given words to remember. Each group was given a list of words with the lists differing in word-meaning categories and natural hierarchical ordering. Each group included n = 12 subjects. The effectiveness of the four different treatments was evaluated by two dependent variables, number of words recalled and number of categories reconstructed.

The example is particularly interesting because the four groups are not significantly different when a univariate F test is applied separately to each dependent variable. For the number of words recalled $F(3,44) = 2.13$, $p = .110$ and for the number of categories reconstructed $F(3,44) = 1.36$, $p = .267$. That is, neither univariate test is significant at $\alpha = .10$. To apply procedures such as Q, P, or B at $\alpha = .05$ would require at least one of the two dependent variables to be significant at .025 (= .05/2).

Applying the usual MANOVA tests, V, W, T, and R, to the data from Finn (1974) would result in respective p values of .00778, .00584, .00454, .00039. This is the result a researcher would find in performing MANOVA on most

statistical packages. It would appear to be safe to consider the results significant at $\alpha = .01$. Applying U would result in $p = .02331$. Present results suggest that UM2 should be used for $n \geq 10$. Applying UM2 with Muller's Method 1 or Method 2 would produce $p = .013711$. This is not significant at $\alpha = .01$ but would be significant at $\alpha = .05$.

References

Barling, J. and Rosenbaum, A. (1986). Work stressors and wife abuse. Journal of Applied Psychology. 71, 346-348.

Bird, K. D. (1975). Simultaneous contrast testing procedures for multivariate experiments, Multivariate behavioral research, 10, 343-351.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 141-152.

Finn, J. D. (1974). A general model for multivariate analysis. New York: Holt, Rinehart, and Winston.

Fassinger, R. E, and Richie, B. S. (1994). Being the best: Preliminary results from a national study of the achievement of prominent Black and White women. Journal of Counseling Psychology. 41(2), 191-204.

Fujikoshi, Y. (1972). Asymptotic formulas for the distributions of the determinant and the trace of a noncentral beta matrix. Journal of Multivariate Analysis, 2, 208-218.

Goldman, J. A. and Harlow, L. L. (1993). Self-perception variables that mediate AIDS-preventive behavior in college students. Health Psychology. 12(6), 489-498.

Harris, R. J. (2001. A primer of multivariate statistics (3rd ed.), pp. 222-233, Mahway, NJ : Lawrence Erlbaum.

Hsu, Yu-S. and Pillai, K. C. S (1985). Asymptotic formulae for the c.d.f. of Hotelling's trace and robustness studies for three tests of hypotheses. Sankhya. The Indian Journal of Statistics. Series B, 47, 1-17.

Ito, K. (1962). A comparison of the powers of two multivariate analysis of variance tests, Biometrika, 49, 455-462.

Lee, Yoong-Sun (1971). Asymptotic formulae for the distribution of a multivariate test statistic: Power comparisons of certain multivariate tests, Biometrika, 58, 647-651.

McKeon, J. J. (1974). F approximations to the distribution of Hotelling's T_0^2. Biometrika, 61, 381-383.

Midgley, C., Feldlaufer, H., & Eccles, J. S. (1989). Change in teacher efficacy and student self- and task-related beliefs in mathematics during the transition to junior high school. Journal of Educational Psychology. 81, 247-258.

Muller, K. E. (1998). A new F approximation of the Pillai-Bartlett trace under H_0, Journal of Computational and Graphical Statistics, 7, 131-12.

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. Psychometrika, 12, 153-

Nijenhuis, A. and Wilf, H. S. (1978). Combinatorial algorithms for computers and calculators (2^{nd} ed.). New York: Academic Press.

O'Brien, P. N., Parente, F. J. and Schmitt,, C. J, (1982). A Monte Carlo study on the robustness of four MANOVA criterion tests. Journal of Statistical Computation and Simulation, 15, 183-192.

Olson, C. L. (1974). Comparative robustness of six tests of multivariate analysis of variance, Journal of the American Statistical Association, 69, 894-908.

Olson, C. L. (1973). A Monte Carlo investigation of the robustness of multivariate analysis of variance. (Doctoral dissertation , University of

Toronto). Dissertation Abstracts International , 1975, 35, 6106B. (Microfilm, National Library of Canada, Ottawa)

Olson, C. L. (1979). Practical considerations in choosing a MANOVA test statistic: A rejoinder to Stevens. Psychological Bulletin, 86,1350-1352.

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis, Annals of Mathematical Statistics, 26, 117-121.

Peritz, E. (1970) A note on multiple comparisons, unpublished paper, Hebrew University.

Press, W. H. Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1994). Numerical recipes in FORTRAN. New York: Cambridge University Press.

Ramsey, P. H. (1982). Empirical power of procedures for comparing two groups on p variables, Journal of Educational Statistics, 7, 139-156.

Ramsey, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. Psychological Methods, 7, 504-523.

Rao, C. R. (1951). An asymptotic expansion of the distribution of Wilks' criterion, Bulletin of the Institute of Intternational Statistics, 33, 177-180.

Roy, S. N. (1966). Sensitivity comparisons among tests of the general linear hypotheses, Journal of the American Statistical Association, 61, 415-435.

Seber, G. A. F. (1984). Multivariate observations. Wiley, New York.

Stevens, J. (1979). Comment on Olson: Choosing a statistic in multivariate analysis of variance. Psychological Bulletin, 86, 355-360.

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

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. Biometrika, 24, 471-494.