THE EFFECTS OF NONNORMALITY ON PARAMETRIC, NONPARAMETRIC, AND MODEL COMPARISON APPROACHES TO PAIRWISE COMPARISONS

ROBERT A. CRIBBIE
York University

H. J. KESELMAN
University of Manitoba

Researchers in the behavioral sciences are often interested in comparing the means of several treatment conditions on a specific dependent measure. When scores on the dependent measure are not normally distributed, researchers must make important decisions regarding the multiple comparison strategy that is implemented. Although researchers commonly rely on the potential robustness of traditional parametric test statistics (e.g., *t* and *F*), these test statistics may not be robust under all nonnormal data conditions. This article compared strategies for performing multiple comparisons with nonnormal data under various data conditions, including simultaneous violations of the assumptions of normality and variance homogeneity. The results confirmed that when variances are unequal, use of the traditional two-sample *t* test can result in severely biased Type I and/or Type II error rates. However, the use of Welch's two-sample test statistic with the REGWQ procedure, with either the usual means and variances or with trimmed means and Winsorized variances, resulted in good control of Type I error rates. The Kruskal-Wallis nonparametric statistic provided good Type I error control and power when variances were equal, although Type I error rates became severely inflated when variances were unequal. Furthermore, for researchers interested in eliminating intransitive decisions or comparing potential mean configuration models, a protected model-testing procedure suggested by Dayton provided good overall results.

Keywords: *multiple comparison procedures; pairwise comparisons*

Educational and Psychological Measurement, Vol. 63 No. 4, August 2003 615-635
DOI: 10.1177/0013164403251283

Pairwise Multiple Comparisons for Nonnormal Data

An underlying assumption of the test statistics used with many multiple comparison procedures (MCPs) is that the populations from which the data are sampled are normal in shape. Although it may be convenient (practically and statistically) for researchers to assume that their samples are obtained from normal populations, this assumption may rarely be accurate (Micceri, 1989; Wilcox, 1990). Researchers falsely assuming normally distributed data (and adopting methods of analysis designed for normally distributed data) risk obtaining biased Type I and/or Type II error rates for many patterns of nonnormality, especially when other assumptions (e.g., variance homogeneity) do not hold.

A common recommendation that has been put forth for conducting multiple comparisons with nonnormally distributed data is to simply apply MCPs with traditional parametric test statistics. Several researchers have demonstrated that under many conditions of nonnormality the usual *t* and *F* statistics are robust with respect to Type I error rates and power (e.g., Boneau, 1960; Sawilowsky & Blair, 1992). The robustness of the *t* and *F* statistics does not necessarily hold, however, for all degrees of nonnormality, when nonnormality is paired with unequal variances and sample sizes or when the distribution shapes are not identical (e.g., Keselman, Lix, & Kowalchuk, 1998; Wilcox, 1990).

Beyond recommendations supporting the use of traditional parametric test statistics, recent publications have discussed novel testing strategies that purport to provide specific advantages for researchers performing multiple comparisons tests with nonnormal data. Sprent (1993) proposes the use of nonparametric test statistics that place no restrictions on the shape of the underlying distributions and can be much more powerful than parametric tests with nonnormally distributed data (e.g., Penfield, 1994). Wilcox (1990) and others have recommended that researchers substitute trimmed means and Winsorized variances for the least squares estimators; Westfall, Tobias, Rom, Wolfinger, and Hochberg (1999) propose that researchers adopt bootstrapping methods; and Dayton (1998) has recommended the use of model-testing procedures. These new strategies for performing multiple comparisons with nonnormal data provide interesting alternatives to applied researchers, although it is important that researchers understand the properties of the proposed procedures, as well as how the proposed procedures perform relative to each other and existing procedures. Therefore, the objectives of this article are (a) to summarize the recently proposed multiple comparison strategies for nonnormal data, and (b) to report the results of a Monte Carlo study that compares Type I error and power rates of the different multiple comparison approaches under various distribution shapes.

Design, Notation, and Test Statistics

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

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij},$$

where Y_{ij} is the score of the ith subject (i = 1, . . ., n_j) in the jth group (j = 1, . . ., J), μ is the population grand mean, α_j is the fixed treatment effect associated with the jth group ($\mu_j - \mu$), and ε_{ij} is the random error for the ith subject in the jth group. In the typical application of the model, it is assumed that the ε_{ij}s are normally and independently distributed and that the group variances (σ_j^2) are equal. An omnibus test of $\mathrm{H_o}: \mu_1 = \mu_2 = \mu_3$ can be conducted using an appropriate omnibus test statistic. The $C = (J^2 - J)/2$ pairwise multiple comparisons ($\mathrm{H_o}: \mu_j = \mu_{j'}; j \neq j'$) can be conducted with an appropriate MCP, as discussed below.

Nonparametric Approach

A popular alternative for analyzing data from nonnormal populations is to select a nonparametric, or distribution free, test statistic. Nonparametric tests require that ranks be substituted in place of the original scores when testing for mean differences. The Kruskal-Wallis t (Sprent, 1993) begins by ranking the observations in the combined sample (N). Let the rank of the ith observation in the jth group be represented by r_{ij} and the sum of the ranks for the jth group be represented by $s_j = \Sigma_i r_{ij}$. The null hypothesis $\mathrm{H_o}: \lambda_j = \lambda_{j'}$ (where λ represents the population mean only under the assumption that the population shapes are identical, see below) is rejected if $t_{KW} \geq t(\alpha, df_W)$, where

$$t_{KW} = \left|m_j - m_{j'}\right| / \{[S_r - B)(n_j + n_{j'}) / [n_j n_{j'}(N - J)(N - 1)]\}^{1/2},$$

where $m_j = s_j/n_j$, $S_r = \Sigma_{ij} r_{ij}^2$, $B = [N(N+1)^2]/4$, and KW is the statistic from the omnibus Kruskal-Wallis test (which is available in most introductory statistics texts and is not reproduced here).

Two caveats are necessary regarding the use of nonparametric tests. First, the general nonparametric null hypothesis relates to the equality of population distributions, which considers differences in the scale, shape, and location of the distributions, not just location. This hypothesis relates specifically to location differences only when it can be assumed that the distributions are identical in scale and shape. Second, nonparametric tests are themselves not necessarily robust to violations of the variance homogeneity assumption (Gibbons & Chakraborti, 1991; Penfield, 1994; Zimmerman, 1996). The Kruskal-Wallis tests can become extremely liberal or conservative depending on the degree of variance heterogeneity and the pattern of variance and sample size heterogeneity (see also Zhou, Gao, & Hui, 1997).

Trimmed Means Approach

When researchers feel that they are dealing with populations that are nonnormal in form (Tukey [1960] suggested that outliers should be a common occurrence in distributions and others [e.g., Miller, 1988; Zumbo & Coulombe, 1997] have indicated that skewed distributions frequently depict psychological [e.g., reaction time] data) and thus subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least squares parameters, then procedures based on robust estimators should be adopted. Wilcox (1990, 1995, 1997) and others have discussed the use of robust estimators such as the trimmed mean and Winsorized variance with nonnormal data.

Trimmed means are computed by removing a percentage of observations from each of the tails of a distribution. Let $g_j = [\gamma_s n_j]$, where γ_s represents the proportion of observations to be trimmed from each tail of the distribution and $[x]$ is the largest integer less than or equal to x. Furthermore, let h_j represent the remaining (effective) sample size following removal of the trimmed observations. Recommendations have been made in the literature for 15% symmetric trimming (Mudholkar, Mudholkar, & Srivastava, 1991) and 20% symmetric trimming (Wilcox, 1995). The trimmed mean is represented as

$$\bar{X}_{ij} = (1/h_j)\sum_{i=g_j+1}^{n_j-g_j} X_{ij}$$

and the jth sample Winsorized mean as

$$\bar{X}_{wj} = (1/n_j)\sum_{i=1}^{n_j} Y_{ij},$$

where

$$\begin{aligned} Y_{ij} &= X_{(g_j+1)j} \text{ if } X_{ij} \le X_{(g_j+1)j} \\ &= X_{ij} \text{ if } X_{(g_j+1)j} < X_{ij} < X_{(n_j-g_j)j} \\ &= X_{(n_j-g_j)j} \text{ if } X_{ij} \ge X_{(n_j-g_j)j}. \end{aligned}$$

An associated Winsorized variance is computed by replacing the censored observations from each tail with the lowest uncensored observation (lower tail) or highest uncensored observation (upper tail). The Winsorized variance equals:

$$s_{wj}^2 = 1/(h_j - 1)\sum_{i=1}^{n_j}(Y_{ij} - \overline{X}_{ij})^2.$$

The trimmed sample means and Winsorized sample variances can then be substituted into Welch's (1938) test statistic to yield the following statistic (Yuen, 1974):

$$t_t = (\overline{X}_{t1} - \overline{X}_{t2})/[(s_{w1}^2/h_1) + (s_{w2}^2/h_2)]^{1/2},$$

with error degrees of freedom,

$$\nu_t = \frac{[(s_{w1}^2/h_1) + (s_{w2}^2/h_2)^2}{\{s_{w1}^4/(h_1^2(h_1 - 1))\} + \{s_{w2}^4/(h_2^2(h_2 - 1))\}}.$$

When trimmed means are being compared, the null hypothesis relates to the equality of population trimmed means, instead of population means. Therefore, instead of testing $\mathrm{H_o}: \mu_j = \mu_{j'}$, a researcher would test the null hypothesis, $\mathrm{H_o}: \mu_{tj} = \mu_{tj'}$, where μ_t represents the population trimmed mean.

One important application of trimmed means and Winsorized variances is in analyzing data that violate the assumptions of normality and variance homogeneity by using these robust estimators with the heteroscedastic Welch (1938) statistic. Although Penfield (1994) advocated using Welch's test with nonnormal heterogeneous data, Welch's test is not always robust for all degrees and patterns of assumption violations (Cressie & Whitford, 1986; Keselman, Lix, et al., 1998). Wilcox (1997) compared the power and Type I error rates of the REGWQ (Einot & Gabriel, 1975; Ryan, 1960; Welsch, 1977), Hayter (1986), and Dunnett (1980) MCPs, and found that substituting trimmed means and Winsorized variances for the usual least squares means and variances resulted in better Type I error control and power when the data were nonnormal and variances were heterogeneous. Keselman, Lix, et al. (1998) compared the power and Type I error rates of several pairwise MCPs with the usual least squares estimators or robust estimators and found that MCPs using trimmed means and Winsorized variances provided better Type I error control for some skewed distributions with unequal sample sizes and variances.

Bootstrap Tests

Researchers conducting multiple comparisons with nonnormal data also have the option of adopting a bootstrap analysis. With bootstrapped tests, the researcher generates an empirical distribution from the sample residuals ($\hat{\in}_{ij} = Y_{ij} - \mu_j$), instead of assuming that the underlying distribution is normal in shape. The empirical distribution is generated by resampling, with replacement, from the distribution of $\hat{\in}_{ij}$. p values are computed for each of the C pairwise comparisons ($p_1, \ldots, p_C$, $c = 1, \ldots, C$) for numerous (e.g., $S = 10{,}000$) bootstrap-generated samples ($s = 1, \ldots, S$). Bootstrap-adjusted p values (p_{c+}) for the cth comparison are the proportion of S for which $p_{cs} \le p_c$ (where p_{cs} is the p value for the cth comparison in bootstrap sample s). Multiplicity adjustment can be applied using several methods, although a step-down procedure suggested by Westfall et al. (1999), dubbed the “Stepboot” procedure, is especially appealing because it is more powerful than many simultaneous methods of multiplicity control and it is available through SAS’s MULTTEST program (see Westfall et al., 1999).

Dayton’s Model-Testing Procedure

Dayton (1998) proposed an innovative strategy for performing multiple comparisons that eliminates intransitive decisions by comparing all possible transitive population models (i.e., mean configurations). One of the distinct advantages of this procedure is that it can be applied with normally or nonnormally distributed data. To summarize the logic of the model-testing procedure, assume that a researcher is interested in testing all pairwise comparisons in a $J = 3$ design. With the model-testing procedure, the researcher would compare (and select the best of) the $k = 2^{J-1} = 2^{3-1} = 4$ transitive population models, instead of testing if any or all of the $C = 3$ pairwise comparisons are significant (as with a traditional MCP). 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. In addition to eliminating intransitive decisions, Dayton’s approach takes a more “wholistic” 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.

The model-testing procedure is based on the information criteria due to Akaike (1974), or AIC. Mutually exclusive and transitive models are each evaluated using AIC and the model having the minimum AIC is retained, where

$$AIC = SS_w + \sum_{j=1}^{J} n_j(\bar{X}_j - \bar{X}_{kj})^2 + 2q,$$

SS_w is the within-group sums of squares from an appropriate omnibus test, $\bar{X}_j$ is the mean of the jth group, $\bar{X}_{kj}$ is the estimated sample mean for the jth group (given the hypothesized population mean configuration for the *k*th model), and *q* is the number of independent parameters estimated in fitting the model. In addition, Dayton (1998) has also shown 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,

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

where *N* is the total number of subjects in the experiment ($\%_j n_j$) and *S* is the biased variance for the *j*th group, substituting the estimated group mean (given the hypothesized mean configuration for the *k*th model) for the actual group mean in the calculation of the variance. The heterogeneous variance AIC statistic adopted in this article is referred to by Dayton (1998) as the unrestricted heterogeneous model (in contrast to the restricted heterogeneous model also presented by Dayton).

Dayton (1998) and Cribbie and Keselman (in press) report that the model-testing procedure is more likely to identify the true underlying population mean configuration than several traditional MCPs (e.g., Tukey's, 1953, Honestly Significant Difference [HSD]). However, one finding reported by Dayton is that the AIC has a slight bias for selecting more complicated models than the true model, and consequently it is recommended that an omnibus test be used to screen for the complete null. Cribbie and Keselman reported a significant improvement in the overall accuracy of the model-testing procedure to detect the correct underlying model when an omnibus test was utilized.

Multiple Comparison Procedures

Four multiple comparison procedures were selected for investigation with the two-sample test statistics described above. Tukey's (1953) HSD procedure (Tukey) was selected for its familiarity and frequent use by applied researchers (see Keselman, Huberty, et al., 1998), Hommel's (1988) procedure (Hommel) was selected as a powerful modified Bonferroni procedure, and Hayter's (1986) modified LSD procedure (Hayter) and Ryan (1960), Einot and Gabriel (1975), and Welsch's (1977) protected stepwise studentized range procedure (REGWQ) were selected because previous research has found them to provide an excellent balance between Type I error control and power (e.g., Kromrey & La Rocca, 1995; Seaman, Levin, & Serlin, 1991).

Tukey. Tukey (1953) proposed an MCP for testing all pairwise comparisons in what Toothaker (1991) described as possibly “the most frequently cited unpublished paper in the history of statistics” (p. 41). Tukey’s HSD uses a critical value obtained from the Studentized Range (q) distribution. The Tukey procedure accounts for dependencies among the pairwise comparisons (i.e., nonorthogonality and the use of a common error term) in deriving a simultaneous critical value. A pairwise null hypothesis is rejected with the Tukey procedure if

$$t \geq q(\alpha, J, \nu) \,/\, (2)^{1/2},$$

where ν represents the degrees of freedom.

Hommel. Hommel (1988) proposed a stagewise (step-up) modified Bonferroni procedure. The first phase of the Hommel procedure contains I steps ($i = 1, \ldots, I$), with each step (i) containing k stages ($k = 1, \ldots, i$), where I is the largest i such that all $p_{(C-i+k)} > k\,\alpha/i$. The second phase of the Hommel procedure rejects all pairwise null hypotheses where $p \leq \alpha \,/\, II$.

REGWQ. Ryan (1960) proposed a modification to the popular Newman-Keuls procedure that ensures that the familywise error rate is maintained at α, even in the presence of multiple partial null hypotheses. Instead of controlling the Type I error rate at α for each stretch size, $p = J, J-1, \ldots, 2$, the Type I error rate is controlled at α_p, where $\alpha_p = p(\alpha)/J$. Ryan’s original procedure became known as the REGWQ after modifications to the procedure by Einot and Gabriel (1975) and Welsch (1977). Einot and Gabriel proposed that the Type I error rate should be controlled at $\alpha_\pi = 1-(1-\alpha)^{p/J}$ to increase power slightly and Welsch proposed that α_p be controlled at α for $p = J$ and $J-1$, given that the original Newman-Keuls procedure is only liberal for $p < J-1$. The REGWQ procedure sequentially tests all ordered mean differences for stretch sizes $p = J, J-1, \ldots, 2$, and rejects a pairwise null hypothesis if

$$t \leq q(\alpha_p, p, \nu) \,/\, (2)^{1/2},$$

where $\alpha_p = \alpha$ for $p = J, J-1$, and $\alpha_p = 1-(1-\alpha)^{p/J}$, for $p = J-2, \ldots, 2$. If any H_cs are retained for $p = p'$ then all H_cs contained in that stretch are retained and not tested at later stages (i.e., $p < p'$). If all H_cs are retained for $p = p'$ then all H_cs with p æ p’ are retained. A final modification proposed by Shaffer (1979) is that an omnibus ANOVA F test first be performed on the J means. If the omnibus test is significant, means separated by $p = J$ steps are tested using the critical value for $p = J-1$, and if the omnibus test is not significant all H_cs are retained.

Hayter. Hayter (1986) proposed a modification to Fisher's LSD that would provide strong control (i.e., control under conditions where all treatment group means are equal, as well as when a subset of the treatment group means are equal) of the familywise error rate. Like the LSD procedure, no comparisons are tested unless the omnibus test is significant. If the omnibus test is significant, then H_c is rejected if

$$t \geq q(\alpha, J - 1, \nu) / (2)^{1/2}.$$

Method

A Monte Carlo study was used to compare the Type I error and power rates of the traditional and recently proposed multiple comparison strategies for nonnormal data.

Eight variables were manipulated in this study: (a) number of levels of the independent variable, (b) total sample size, (c) degree of sample size imbalance, (d) degree of variance inequality, (e) pairings of unequal group sizes and variances, (f) configuration of population means, (g) population distribution shape, and (h) two-sample (and omnibus if necessary) test statistic applied.

To evaluate the effect of the number of pairwise comparisons computed (an important consideration for techniques involved in controlling for the effect of multiplicity of testing) on Type I error control and power, the number of levels of the independent variable was set at $J = 4$ and $J = 7$, resulting in 6 and 21 comparisons, respectively.

To investigate the effects of sample size, the total sample size (N) was manipulated by setting the average $n_j = 10$, 15, and 25, resulting in $N = 40$, 60, and 100 for $J = 4$, and $N = 70$, 105, and 175 for $J = 7$. For the nonnull mean configurations used in this study, the group sizes 10, 15, and 25 result in a priori omnibus (F statistic) power estimates of approximately .5, .7, and .9, respectively (assuming equal group sizes and variances).

Sample size balance or imbalance was also manipulated. Keselman, Huberty, et al. (1998) reported that unbalanced designs were more common than balanced designs in a review of studies published in educational and psychological journals. In addition, the effects of variance heterogeneity can be exacerbated when paired with heterogeneous sample sizes. Therefore, three sample size conditions were examined (equal n_j, moderately unequal n_j, and extremely unequal n_j). The specific sample sizes used in this study are enumerated in Table 1.

Degree of variance heterogeneity was also manipulated. According to Keselman, Huberty, et al. (1998), ratios of largest to smallest variances of 8:1 are not uncommon in educational and psychological studies and can have deleterious effects on the performance of many MCPs, especially when

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	

paired with unequal sample sizes. Therefore, three levels of variance equality/inequality were utilized in this study: (a) equal variances, (b) largest to smallest variance ratio of 4:1, and (c) largest to smallest variance ratio of 8:1. See Table 1 for specific group variances for $J = 4$ and $J = 7$.

The specific pairings of unequal variances and sample sizes can have differing effects on the Type I error and power rates of many test statistics. Specifically, when variances and sample sizes are directly (positively) paired, Type I error estimates can be conservative (with correspondingly deflated power). On the other hand, when variances and sample sizes are inversely (negatively) paired, Type I error estimates can be liberal (with correspondingly inflated power). Therefore, both positive and negative pairings were evaluated.

Several configurations of nonnull population means were investigated in this study, in addition to the complete null case. Following Toothaker's (1991) definitions of mean configuration, equally spaced, minimum variability, and maximum variability configurations were utilized (see Table 2 for a listing of the mean configurations used in this study).

Another factor examined in this study was population distribution shape. The three distribution shapes investigated were (a) normally distributed data, (b) moderately skewed data from a χ^2_3 distribution (skewness = 1.63, kurtosis = 4.00), and (c) substantially skewed data from the g and h distribution (Hoaglin, 1985), where $g = 1$ and $h = 0$ (skewness = 6.20, kurtosis = 114). The moderately skewed case is representative of values reported by Micceri

Table 2
Population Mean Configurations Used in the Simulation Study

	Population Means						
J	1	2	3	4	5	6	7
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			
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

(1989), and the substantially skewed case is intended to give one a picture of what could occur with respect to Type I error and power in an arguably worst-case scenario, with the premise being that if a method performs well in an extreme case, it is also likely to perform well for other cases not investigated.

Last, three two-sample test statistics (with corresponding omnibus statistics when required) were evaluated and compared in this study. These statistics include (a) Welch's (1938) *t* (omnibus Welch, 1951); (b) the trimmed Welch *t* (omnibus trimmed Welch), based on 20% symmetric trimming; and (c) Kruskal-Wallis *t* (omnibus Kruskal-Wallis).

Familywise error rates were recorded for all procedures, except the model-testing procedure, which is not designed to control a specific error rate. In this article the robustness of a procedure, with respect to Type I error control, will be determined using Bradley's (1978) liberal criterion. That is, a procedure is deemed robust with respect to Type I errors if the empirical rate of Type I error falls within the range +/-.5 α. Three conceptualizations of sensitivity were used in this study: (a) all-pairs power, (b) average per-pair power, and (c) the true model rate. All-pairs power is defined as the probability of rejecting all false pairwise null hypotheses. Average per-pair is defined as the average probability of rejecting any false pairwise null hypothesis. The true model rate is defined as the probability of selecting the correct model for the Dayton procedures, and as the probability of rejecting all false pairwise null hypotheses and not rejecting any true null hypotheses for the remaining MCPs.

The simulation program was written in SAS/IML (SAS Institute, 1999a). Pseudorandom 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 χ_3^2 data, three standard normal variates were squared and summed. The χ^2 variates were standardized and transformed to variates with mean μ_j and variance σ_j^2. To generate data from the *g* and *h* distributions, standard unit normal variables were converted to the random variable

$$X_{ij} = \{[\exp(gZ_{ij}) - 1] / g\} \{\exp(hZ_{ij}^2 / 2)\}.$$

To obtain a distribution with standard deviation &$_j$, we multiplied each X_{ij} by a value of σ_j. With $g = 1$ and $h = 0$, the *g* and *h* distribution population mean is .6487. Thus, .6487 was subtracted from X_{ij} before being multiplied by σ_j. When working with trimmed means, the population trimmed mean for the *j*th group ($\mu_{ij} = 0.111$) was also subtracted from the variate before multiplying by σ_j. Five thousand replications were performed for each condition, using a nominal significance level of .05.

Results

The pattern of Type I error and power results were consistent across unequal sample size conditions (moderately unequal n_j, very unequal n_j), unequal variance conditions (moderately unequal σ_j^2, very unequal σ_j^2), and nonnull mean configurations, and were therefore averaged across these conditions. Similar results were also found over the number of levels of the independent variable ($J = 4, 7$) and the sample size conditions (average $n_j = 10, 15$, and 25); thus, only results for $J = 7$ and average $n_j = 25$ are reported. The results were also consistent across the two skewed distributions (χ_3^2 and $g = 1$, $h = 0$) and were thus averaged over these distributions, unless otherwise noted. Last, partial null familywise error rates were controlled within Bradley’s limits in all cases where complete null Type I error rates were controlled within Bradley’s limits, and therefore are not reported (the complete set of results can be obtained from the first author).

Type I Error Control

Normal distribution. Complete null familywise rates for $J = 7$ and normally distributed data are presented in Table 3. The Stepboot procedure maintained complete null rates within Bradley’s liberal bounds when variances were equal or sample sizes and variances were positively paired, but not when sample sizes and variances were negatively paired (13.52%). The remaining procedures (Hommel, Tukey, Hayter, and REGWQ) maintained complete null rates within Bradley’s liberal bounds when Welch’s statistic was used with either the usual sample means and variances or with trimmed

Table 3
*Complete Null Familywise Error Percentages for J = 7 (*N *= 175)*

	Welch *t*			Welch Trimmed *t*			Kruskal-Wallis *t*		
	$= \sigma_j^2$	PP	NP	$= \sigma_j^2$	PP	NP	$= \sigma_j^2$	PP	NP
Normal distribution									
Hommel	4.08	4.25	4.00	4.32	4.31	4.57	4.25	3.46	7.18
Tukey	5.54	5.75	5.37	6.28	6.32	6.68	5.61	4.38	*8.95*
Hayter(P)	4.37	4.76	4.70	5.11	5.24	5.68	3.36	2.95	6.98
REGWQ(P)	3.52	3.96	2.97	3.76	3.84	3.45	3.22	2.87	6.63
Stepboot[a]	5.81	4.15	*13.52*	5.81	4.15	*13.52*	5.81	4.15	*13.52*
Skewed distributions[b]									
Hommel	3.25	3.39	*9.10*	2.72	2.97	5.28	4.48	*19.39*	*32.91*
Tukey	4.25	4.67	*11.08*	3.87	4.33	7.20	5.79	*22.62*	*36.51*
Hayter(P)	4.58	4.87	*10.77*	4.01	4.21	6.63	4.62	*23.07*	*35.76*
REGWQ(P)	2.56	2.53	4.11	2.74	2.65	3.14	3.95	*21.45*	*34.00*
Stepboot[a]	5.61	5.19	*17.92*	5.61	5.19	*17.92*	5.61	5.19	*17.92*

Note: $= \sigma_j^2$ = equal population variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively; (P) = protected test.

a. The Stepboot procedure does not use any of the above two-sample test statistics and is included for comparison only

b. The results for the skewed distribution are averaged over the χ_3^2 and $g = 1$, $h = 0$ distributions; values that are in italics are liberal according to Bradley's (1978) limits of +.5.

means and Winsorized variances. When the Kruskal-Wallis *t* was used, only the Tukey procedure (with negatively paired sample sizes and variances) had complete null rates (8.95%) that exceeded Bradley's criterion.

Skewed distributions. Complete null familywise error rates for *J* = 7 and skewed data are presented in Table 3. The Stepboot procedure, as with normal data, was not able to maintain complete null rates below Bradley's upper bound when sample sizes and variances were negatively paired (17.92%).

The REGWQ procedure maintained complete null rates within Bradley's liberal bounds under all conditions when Welch's statistic was used with either the usual sample means and variances or with trimmed means and Winsorized variances. The remaining procedures (Hommel, Tukey, and Hayter) were unable to maintain rates below .075 with the Welch statistic (usual means/variances) when sample sizes and variances were negatively paired; however, these procedures were able to maintain rates below .075 under all conditions when the Welch statistic was applied with trimmed means and Winsorized variances. When the Kruskal-Wallis *t* was applied, all the procedures maintained complete null rates within Bradley's liberal criterion when variances were equal, although none of the procedures was able to maintain rates below .075 when sample sizes and variances were unequal.

Power

Normal distribution. Per-pair and all-pairs power rates for J = 7 and normally distributed data are presented in Tables 4 and 5, respectively. Per-pair and all-pairs power rates were similar across the familywise error controlling procedures. With respect to test statistics, both per-pair and all-pairs power rates were lower when the Welch procedure was applied with trimmed means and Winsorized variances than when either the Welch procedure was applied with the usual means and variances or the Kruskal-Wallis statistic was utilized, although the differences were not substantial.

Skewed distributions. Per-pair and all-pairs power rates for J = 7 and skewed data are presented in Tables 4 and 5, respectively. The Stepboot procedure had per-pair and all-pairs power rates (approximately 25% and 2%, respectively, with equal variances) that were comparable to the remaining procedures when the Welch statistic was used, although rates for the Stepboot procedure were consistently greater than the rates for the other procedures with trimmed means and Winsorized variances, and considerably less than rates for the other procedures with the Kruskal-Wallis statistic. Differences in all-pairs power were minimized by floor effects.

Per-pair and all-pairs power rates for the Hayter and REGWQ procedures were greater than the rates for any of the remaining procedures under all conditions and across all test statistics, although the differences were in most cases minimal.

With respect to test statistics, both per-pair and all-pairs power rates were significantly lower when the Welch procedure was applied (usual means/variances or trimmed means/Winsorized variances) than when the Kruskal-Wallis statistic was utilized. For example, per-pair and all-pairs power rates with the REGWQ and equal variances were 16.94% and 2.31% when the Welch t was applied with the usual means and variances, 13.67% and 0.69% when the Welch t was applied with the trimmed means and Winsorized variances, and 41.49% and 13.64% when the Kruskal-Wallis statistic was applied, respectively.

In comparing the Welch t with the usual means and variances and the Welch t with trimmed means and Winsorized variances, the degree of skewness had a significant effect on the power of the MCPs (nontabled values). When the distributions were moderately skewed (zero distribution), the per-pair and all-pairs power rates for the Welch test with the usual means and variances (e.g., 10.82% and 0.27%, respectively, with the REGWQ procedure and negatively paired sample size and variances) were substantially greater than the rates for the Welch test with trimmed means and Winsorized variances (e.g., 2.14% and 0.00%, respectively, with the REGWQ procedure and negatively paired sample size and variances). However, for the g = 1, h = 0 skewed distribution, the per-pair and all-pairs power rates for the Welch test

Table 4

Average Per-Pair Power Percentages for J = 7 (N = 175)

	Welch *t*			Welch Trimmed *t*			Kruskal-Wallis *t*		
	= σ_j^2	PP	NP	= σ_j^2	PP	NP	= σ_j^2	PP	NP
Normal distribution									
Hommel	31.21	6.12	9.24	26.42	4.79	7.50	33.22	4.88	11.25
Tukey	32.32	7.11	10.55	27.32	5.86	8.95	33.49	5.51	12.19
Hayter(P)	34.12	7.69	11.69	31.65	6.26	9.84	36.58	6.08	13.43
REGWQ(P)	32.33	7.12	6.51	27.24	5.40	5.17	37.54	5.75	12.75
Stepboot[a]	35.13	6.47	11.19	35.13	6.47	11.19	35.13	6.47	11.19
Skewed distributions[b]									
Hommel	19.57	3.28	11.27	13.16	1.52	5.79	38.62	3.35	34.52
Tukey	20.87	3.93	12.41	14.57	2.06	6.77	39.98	3.78	34.95
Hayter(P)	22.18	4.29	13.32	15.24	2.17	7.14	40.27	4.02	37.47
REGWQ(P)	16.94	3.77	7.48	13.67	1.74	3.60	41.49	3.70	37.84
Stepboot[a]	24.08	4.03	10.97	24.08	4.03	10.97	24.08	4.03	10.97

Note. = σ_j^2 = equal population variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively; (P) = protected test.

a. The Stepboot procedure does not use any of the above two-sample test statistics and is included for comparison only.

b. The results for the skewed distribution are averaged over the χ_3^2 and g = 1, h = 0 distributions.

Table 5

All-Pairs Power Percentages for J = 7 and (N = 175)

	Welch *t*			Welch Trimmed *t*			Kruskal-Wallis *t*		
	= σ_j^2	PP	NP	= σ_j^2	PP	NP	= σ_j^2	PP	NP
Normal distribution									
Hommel	3.48	0.31	0.06	2.56	0.19	0.02	6.74	0.15	0.45
Tukey	3.69	0.33	0.07	2.74	0.21	0.03	7.04	0.16	0.45
Hayter(P)	5.25	0.44	0.10	3.13	0.28	0.04	8.93	0.20	0.58
REGWQ(P)	7.67	0.82	0.19	4.86	0.56	0.10	11.63	0.47	1.29
Stepboot[a]	6.44	0.57	0.17	6.44	0.57	0.17	6.44	0.57	0.17
Skewed distributions[b]									
Hommel	1.03	0.06	0.05	0.20	0.01	0.01	7.10	0.03	3.59
Tukey	1.17	0.07	0.06	0.23	0.01	0.01	6.97	0.03	3.49
Hayter(P)	1.34	0.09	0.08	0.38	0.02	0.01	8.07	0.04	4.05
REGWQ(P)	2.31	0.21	0.15	0.69	0.04	0.02	13.64	0.09	7.17
Stepboot[a]	2.42	0.24	0.27	2.42	0.24	0.27	2.42	0.24	0.27

Note. = σ_j^2 = equal population variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively; (P) = protected test.

a. The Stepboot procedure does not use any of the above two-sample test statistics and is included for comparison only.

b. The results for the skewed distribution are averaged over the χ_3^2 and g = 1, h = 0 distributions.

with the usual means and variances (e.g., 4.14% and 0.02%, respectively, with the REGWQ procedure and negatively paired sample size and variances) were less than the rates for the Welch test with trimmed means and Winsorized variances (e.g., 5.87% and 0.05%, respectively, with the REGWQ procedure and negatively paired sample size and variances).

True Model Rates

Normal distribution. True model rates for $J = 7$ and normally distributed data are presented in Table 6. True model rates for the Protected Dayton procedure (approximately 22% with equal variances, and 16% with unequal sample sizes and variances) were similar to the rates of the familywise controlling procedures (approximately 18% to 21% with equal variances, and 15% to 17% with unequal sample sizes and variances).

True model rates for the REGWQ and Hayter procedures were consistently larger than the rates for the Hommel or Tukey procedures. In addition, no substantial differences existed between true model rates with either the Welch, Welch with trimmed means and Winsorized variances, or Kruskal-Wallis test statistics.

Skewed distributions. True model rates for $J = 7$ and skewed data are presented in Table 6. True model rates for the Protected Dayton procedure (approximately 21% with equal variances) were similar to the rates for the familywise error controlling procedures. True model rates for the REGWQ were consistently larger than the rates of the remaining familywise error controlling procedures, although the differences were small (less than 5%).

With respect to test statistics, true model rates for procedures using the Kruskal-Wallis test were approximately 3% to 5% larger than the same procedures using the Welch statistic with the usual means and variances, and approximately 4% to 7% larger than the same procedures using the Welch statistic with trimmed means and Winsorized variances.

Discussion

Researchers in the behavioral sciences are often confronted with the task of evaluating pairwise multiple comparisons with nonnormal data. This article discussed available multiple comparison strategies that each purport to offer the researcher distinct advantages relative to other available strategies.

The Welch two-sample test (Welch, 1938) with the usual means and variances (with the REGWQ MCP), as well as the Welch two-sample test with trimmed means and Winsorized variances (with the Hommel, Tukey, Hayter, or REGWQ MCPs) maintained Type I error rates below Bradley's upper liberal bound under all of the conditions investigated in this study. However,

Table 6
*True Model Rate Percentages for J = 7 and (*N *= 175)*

	Welch *t*			Welch Trimmed *t*			Kruskal-Wallis *t*		
	= σ_j^2	PP	NP	= σ_j^2	PP	NP	= σ_j^2	PP	NP
Normal distribution									
Hommel	18.53	16.22	16.05	17.55	16.10	15.92	18.43	16.22	15.83
Tukey	18.31	15.98	15.83	17.29	15.78	15.58	18.30	16.07	15.54
Hayter(P)	19.14	16.23	15.97	17.61	16.02	15.76	19.12	16.29	15.94
REGWQ(P)	20.89	16.68	16.33	18.87	16.47	16.17	20.99	16.56	16.59
Stepboot[a]	19.91	16.45	14.55	19.91	16.45	14.55	19.87	16.45	14.55
Dayton[a](P)	22.30	16.42	16.85	22.30	16.42	16.85	22.48	16.71	16.42
Skewed distributions[b]									
Hommel	17.42	16.15	15.20	16.31	16.18	15.80	20.33	13.46	13.70
Tukey	17.23	15.94	14.87	16.33	15.96	15.47	20.47	12.92	13.06
Hayter(P)	17.53	15.93	14.94	16.42	15.99	15.57	21.12	15.26	16.51
REGWQ(P)	18.54	16.42	16.19	17.11	16.27	16.16	24.49	15.39	18.72
Stepboot[a]	17.92	16.11	14.17	17.92	16.11	14.17	17.92	16.11	14.17
Dayton[a](P)	20.74	15.42	15.54	20.74	15.42	15.54	20.94	14.20	13.31

Note. = σ_j^2 = equal population variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively; (P) = protected test.

a. The Stepboot and Dayton procedures do not use any of the above two-sample test statistics and are included for comparison only.

b. The results for the skewed distribution are averaged over the χ_3^2 and $g = 1$, $h = 0$ distributions.

although both strategies provided strong control of the Type I error rates under the conditions investigated, the per-pair and all-pairs power of the tests with the usual means and variances were consistently larger than the rates of the tests with trimmed means and Winsorized variances. More specifically, however, with moderately skewed data, the per-pair and all-pairs power rates of tests based on the usual means and variances were substantially larger than the rates of the tests based on trimmed means and Winsorized variances, although with substantially skewed data the per-pair and all-pairs power rates of the tests based on trimmed means and Winsorized variances were larger than the rates of the tests based on the usual means and variances. These findings are consistent with those reported by Keselman, Lix, et al. (1998) and highlight the importance of the degree of nonnormality in deciding between the use of the Welch test with the usual means and variances or with trimmed means and Winsorized variances.

The use of the Kruskal-Wallis nonparametric test statistic (Sprent, 1993) was also investigated in this study. The Kruskal-Wallis *t*, although having liberal Type I error rates when sample sizes and variances were unequal, provided strong Type I error control with equal variances, and was often much more powerful than MCPs with the Welch test. For example, with seven levels of the independent variable and skewed data, per-pair and all-pairs power rates for the REGWQ procedure were 17% and 2%, respectively, with the Welch test (usual means/variances), and 42% and 14%, respectively, with the Kruskal-Wallis *t*.

Of the familywise error controlling procedures investigated, the REGWQ procedure performed well relative to the Hommel, Tukey, and Hayter procedures with respect to Type I error control and power. With respect to power, the Tukey procedure (one of the most widely adopted procedures in the behavioral sciences) performed poorly relative to the REGWQ, especially with respect to all-pairs power. In addition, the REGWQ provided more consistent Type I error control than the Tukey procedure. For example, when the Welch test was applied with skewed data, the Tukey procedure allowed familywise error rates to exceed Bradley's upper liberal bound with negatively paired sample sizes and variances, whereas the Protected REGWQ procedure maintained the familywise error rates within Bradley's liberal bounds.

A different strategy that has been proposed for dealing with nonnormal data is bootstrapping, in which an empirical distribution is generated by sampling repeatedly from the raw data residuals. Although the bootstrapping procedure evaluated in this study (Westfall et al., 1999) provided strong Type I error control when variances were equal, Type I error rates became extremely liberal when sample sizes and variances were negatively paired. For example, familywise error rates with seven levels of the independent variable reached more than 13% with normally distributed and chi-square distributed data. In addition, although the Stepboot procedure provided strong Type I error control with equal variances when the data was skewed, per-pair and all-pairs power rates never exceeded those of the REGWQ procedure with the Kruskal-Wallis statistic. For example, with $J = 7$, equal variances, and skewed data, per-pair power rates for the Stepboot procedure were approximately 15% less than for the REGWQ procedure.

Although the measurement of power and Type I error rates are often undertaken separately, the investigation of true model rates in this study allows for a unique evaluation of the performance of MCPs that considers Type I error control and power simultaneously. In particular, the true model rate evaluates the ability of a MCP to reject all false null hypotheses while not rejecting any true null hypotheses, a definite goal of researchers utilizing MCPs (or any other inferential statistic). With seven levels of the independent variable, although the Protected Dayton procedure had minimally larger true model rates than MCPs with the Welch test, true model rates for the REGWQ procedure with the Kruskal-Wallis test were equal to or larger than the rates for the Protected Dayton procedure.

True model rate results also verified previous recommendations based on separate Type I error and power evaluations. For example, true model rates for MCPs with the Kruskal-Wallis statistic were consistently larger with equal variances and skewed data than with any of the other test statistics evaluated, and provide further evidence of the optimal balance between Type I error control and power for the Kruskal-Wallis statistic with equal variances. True model rate results also support the recommendation of the REGWQ procedure over remaining familywise controlling procedures. Across all conditions and test statistics evaluated in this study, true model rates for the Hommel, Tukey or Hayter procedures never exceeded the rates of the REGWQ procedure.

It is important to acknowledge that the primary limitation of this study is that not all potential data conditions could be investigated, and therefore the conclusions of this study may not necessarily extend to other data conditions. With this cautionary note in mind, we nonetheless offer the following general guidelines for researchers performing pairwise multiple comparison tests with nonnormal data: (a) When variances are unequal, researchers are advised to use the REGWQ procedure with the Welch test statistic(s). For data that appear to be moderately nonnormal, the Welch statistic can be applied with the usual least squares estimators, although for substantial degrees of nonnormality the Welch statistic should be applied with robust estimators. (b) When variances are equal, researchers are advised to use the REGWQ procedure with the Kruskal-Wallis *t* (and Kruskal-Wallis omnibus statistic) for superior Type I error control and power. However, it is important to acknowledge that although the Kruskal-Wallis MCP is recommended when variances are equal, recent research has reported that variances are often not equal (e.g., Keselman, Huberty, et al., 1998). Furthermore, given the highly inflated Type I error rates for the Kruskal-Wallis MCP with unequal variances, researchers must have explicit knowledge that variances are equal (i.e., variance equality should not be assumed) before adopting the Kruskal-Wallis procedure. (c) Finally, when motivation is present for eliminating intransitive decisions or comparing possible models, the Protected Dayton procedure can be adopted and provides a good balance between Type I error control and power.

References

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

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, *57*, 49-64.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Cressie, N. A. C., & Whitford, H. J. (1986). How to use the two sample *t*-test. *Biometrical Journal*, *28*, 131-148.

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

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

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association*, *75*, 796-800.

Einot, I., & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, *70*, 574-583.

Gibbons, J. D., & Chakraborti, S. (1991). Comparisons of the Mann-Whitney, Student's *t*, and alternate *t* tests for means of normal distributions. *Journal of Experimental Education*, *59*, 259-267.

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. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring data tables, trends, and shapes* (pp. 461-513). New York: Wiley.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, *75*, 383-386.

Keselman, H. J., Huberty, C. J., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350-386.

Keselman, H. J., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods*, *3*, 123-141.

Kromrey, J. D., & La Rocca, M. A. (1995). Power and Type I error rates of new pairwise multiple comparison procedures under heterogeneous variances. *Journal of Experimental Education*, *63*, 343-362.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 539-543.

Mudholkar, A., Mudholkar, G. S., & Srivastava, D. K. (1991). A construction and appraisal of pooled trimmed-*t* statistics. *Communications in Statistics: Theory and Methods*, *20*, 1345-1359.

Penfield, D. A. (1994). Choosing a two-sample location test. *Journal of Experimental Education*, *62*, 343-360.

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

SAS Institute. (1999a). *SAS/IML user's guide, Version 8*. Cary, NC: Author.

SAS Institute. (1999b). *SAS/STAT user's guide, Version 7-1*. Cary, NC: Author.

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.

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.

Sprent, P. (1993). *Applied nonparametric statistical methods* (2nd ed.). London: Chapman & Hall.

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

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

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probabilities and statistics*. Stanford, CA: Stanford University Press.

Welch, B. L. (1938). The significance of the difference between two means when population variances are unequal. *Biometrika*, *38*, 330-336.

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

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

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (1999). *Multiple comparisons and multiple tests using the SAS system*. Cary, NC: SAS Institute.

Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrics Journal*, *32*, 771-780.

Wilcox, R. R. (1995). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology*, *48*, 99-114.

Wilcox, R. R. (1997). Three multiple comparison procedures for trimmed means. *Biometrical Journal*, *37*, 643-656.

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika*, *61*, 165-170.

Zhou, X., Gao, S., & Hui, S. L. (1997). Methods for comparing the means of two independent log-normal samples. *Biometrics*, *53*, 1129-1135.

Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *Journal of Experimental Education*, *64*, 351-362.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology*, *51*, 139-150.