The Journal of Experimental Education, 1997, 65(4), 355–366

Specialized Tests for Detecting Treatment Effects in the Two-Sample Problem

H. J. KESELMAN
ROBERT CRIBBIE
University of Manitoba

BRUNO D. ZUMBO
University of Northern British Columbia

ABSTRACT. Nonparametric and robust statistics (those using trimmed means and Winsorized variances) were compared for their ability to detect treatment effects in the 2-sample case. In particular, 2 specialized tests, tests designed to be sensitive to treatment effects when the distributions of the data are skewed to the right, were compared with 2 nonspecialized nonparametric (Wilcoxon–Mann–Whitney; Mann & Whitney, 1947; Wilcoxon, 1949) and trimmed (Yuen, 1974) tests for 6 nonnormal distributions that varied according to their measures of skewness and kurtosis. As expected, the specialized tests provided more power to detect treatment effects, particularly for the nonparametric comparison. However, when distributions were symmetric, the nonspecialized tests were more powerful; therefore, for all the distributions investigated, power differences did not favor the specialized tests. Consequently, the specialized tests are not recommended; researchers would have to know the shapes of the distributions that they work with in order to benefit from specialized tests. In addition, the nonparametric approach resulted in more power than the trimmed-means approach did.

IT IS WELL KNOWN that the usual group mean and variance, which are the basis for the traditional tests of significance for treatment-group equality, are greatly influenced by the presence of extreme observations in a distribution of scores. In particular, the standard error of the usual mean can become seriously inflated when the underlying distribution has heavy tails, a situation likely to be encountered with education and psychological data.

Two alternative strategies for assessing treatment effects are nonparametric methods (see Penfield, 1994) and the use of robust estimators (see Wilcox,

1995a, 1995b). In addition, for each of these strategies there are test statistics that are intended to increase one's ability to detect treatment effects when underlying distributions are skewed. For nonparametric methods, there are specialized tests (two-sample) that are appropriate for assessing treatment-group equality when data are skewed (see Berry, 1995; Hogg, Fisher, & Randles, 1975; Randles & Wolfe, 1979). With robust estimators, in particular, trimmed means and Winsorized variances, asymmetric trimming of the data is expected to provide increased sensitivity to detect treatment effects.

In the present study, we compared four test statistics appropriate for examining treatment-group location equality in the two-sample problem. In particular, we evaluated (a) the Wilcoxon–Mann–Whitney rank sum test (WILC; Wilcoxon, 1949; Mann & Whitney, 1947), (b) a nonparametric test specialized "to detect a location difference between two distributions that are otherwise identical and skewed right" (RSKEW; Berry, 1995, p. 41), and (c)/(d) two versions of Yuen's (1974) modification of the Welch (1938) test with trimmed means and Winsorized variances (YUEN; see Wilcox, 1995a, 1995b; Yuen, 1974). To date, no researchers have examined the extent to which the specialized tests increase one's ability to detect nonnull treatment effects.

Definition of the Test Statistics

Suppose n_j independent random observations $X_{i1}, X_{i2}, \ldots, X_{in_j}$ are sampled from population j ($j = 1, 2$). We assume that the X_{ij}s are obtained from a normal population with mean μ_j and unknown variance σ_j^2, with $\sigma_1^2 = \sigma_2^2$. Then, let $\bar{X}_j = \sum_i X_{ij}/n_j$ and $s_j^2 = \sum_i (X_{ij} - \bar{X}_j)^2/(n_j - 1)$, where $\bar{X}_j$ is the estimate of μ_j and s_j^2 is the usual unbiased estimate of the variance for population j.

The definition of the Wilcoxon–Mann–Whitney test has been presented in many sources and will not be repeated here. Readers can refer to Marascuilo and McSweeney (1977; see also Penfield, 1994). The RSKEW test is described in Berry (1995, p. 41; see also Hogg et al., 1975; Randles & Wolfe, 1979). Although the Welch (1938) test also has been defined many times in the literature, we repeat its definition here so that its use with trimmed means and Winsorized variances is explicitly delineated.

The statistic presented by Welch (1938) for testing the null hypothesis H_0: $\mu_1 = \mu_2$ in the presence of variance heterogeneity is

$$t_W = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{1}$$

where error degrees of freedom (ν) are obtained from

$$\nu_W = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}. \quad (2)$$

It is well known that the usual group mean and variance, which are the basis for the previously described procedure, are greatly influenced by the presence of extreme observations in a distribution of scores. In particular, the standard error of the usual mean can become seriously inflated when the underlying distribution has heavy tails. Accordingly, adopting a nonrobust measure "can give a distorted view of how the typical individual in one group compares to the typical individual in another, and about accurate probability coverage, controlling the probability of a Type I error, and achieving relatively high power" (Wilcox, 1995a, p. 66). By substituting robust measures of location and scale for the usual mean and variance, one should be able to obtain a test statistic that is insensitive to nonnormality.

Although a wide range of robust estimators have been proposed in the literature (see Gross, 1976; Lind & Zumbo, 1993), the trimmed mean and Winsorized variance are intuitively appealing because of their computational simplicity and good theoretical properties (Wilcox, 1995a). The standard error of the trimmed mean is less affected than the usual mean by departures from normality because extreme observations, that is, observations in the tails of a distribution, are censored or removed. Furthermore, as Gross (1976) noted, "The Winsorized variance is a consistent estimator of the variance of the corresponding trimmed mean" (p. 410). When the Winsorized variance is computed, the most extreme observations are replaced with less extreme values in the distribution of scores.

However, these measures should be adopted only if the researcher is interested in testing for treatment effects across groups by using a measure of location that more accurately reflects the typical score within a group when working with heavy-tailed distributions. The hypothesis tested when the usual mean is used as an estimate of location is not the same as that tested when the trimmed mean is used. When trimmed means are being compared, the null hypothesis pertains to the equality of population trimmed means, that is, the μ_ts. That is, H_{0j}: $\mu_{t1} = \mu_{t2}$, (H_{Aj}: $\mu_{t1} \neq \mu_{t2}$).

The prevalent method of trimming is to remove outliers from each tail of the distribution of scores. However, asymmetric trimming has been theorized to be potentially advantageous when the distributions are known to be skewed, a situation likely to be realized with behavioral science data (see De Wet & van Wyk, 1979; Micceri, 1989; Tiku, 1980, 1982; Wilcox, 1994, 1995b).

Let $X_{(1)j} \leq X_{(2)j} \leq \ldots \leq X_{(n_j)j}$ represent the ordered observations associated with

the jth group. When trimming asymmetrically, let $g_j = [\gamma_a n_j]$, where γ_a represents the proportion of trimming in the longer tail and $[x]$ is the greatest integer $\leq x$. The effective sample size for the jth group is $h_j = n_j - g_j$. Assuming that the distribution is positively skewed so that trimming takes place in the upper tail, the jth sample trimmed mean is

$$\bar{X}_{tj} = \frac{1}{h_j} \sum_{i=1}^{n_j - g_j} X_{(i)j}, \tag{3}$$

and the jth sample Winsorized mean is

$$X_{Wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}, \tag{4}$$

where

$$Y_{ij} = X_{ij} \text{ if } X_{ij} < X_{(n_j - g_j)j}$$
$$= X_{(n_j - g_j)j} \text{ if } X_{ij} \geq X_{(n_j - g_j)j}.$$

When trimming symmetrically, let $g_j = [\gamma_s n_j]$, where γ_s represents the proportion of observations that are to be trimmed in each tail of the distribution. The effective sample size for the jth group becomes $h_j = n_j - 2g_j$. Under symmetric trimming, the jth sample trimmed mean is

$$\bar{X}_{tj} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j - g_j} X_{(i)j}. \tag{5}$$

The statistic $\bar{X}_{Wj}$ is computed by using Equation 4, but here

$$Y_{ij} = X_{(g_j+1)j} \text{ if } X_{ij} \leq X_{(g_j+1)j}$$
$$= X_{ij} \text{ if } X_{(g_j+1)j} < X_{ij} < X_{(n_j - g_j)j}$$
$$= X_{(n_j - g_j)j} \text{ if } X_{ij} \geq X_{(n_j - g_j)j}.$$

For both asymmetric and symmetric trimming, the sample Winsorized variance is

$$s^2_{Wj} = \frac{1}{h_j - 1} \sum_{i=1}^{n_j} (Y_{ij} - \bar{X}_{Wj})^2. \tag{6}$$

Yuen (1974) suggested that trimmed means and Winsorized variances be used in conjunction with Welch's (1938) statistic. For heavy-tailed symmetric distributions, Yuen pointed out that the statistic based on trimmed means and Winsorized variances could adequately control the rate of Type I errors and resulted in greater power than a statistic based on the usual mean and variance. Under robust

estimation, the trimmed sample means and Winsorized sample variances were substituted into Equations 1 and 2. Accordingly, Yuen's statistic is

$$t_Y = \frac{\bar{X}_{t1} - \bar{X}_{t2} - (\mu_{t1} - \mu_{t2})}{\sqrt{\frac{s^2_{W1}}{h_1} + \frac{s^2_{W2}}{h_2}}}, \text{ and} \tag{7}$$

$$\nu_Y = \frac{\left(\frac{s^2_{W1}}{h_1} + \frac{s^2_{W2}}{h_2}\right)^2}{\frac{(s^2_{W1}/h_1)^2}{h_1 - 1} + \frac{(s^2_{W2}/h_2)^2}{h_2 - 1}}. \tag{8}$$

Throughout the remainder of this article, Yuen's (1974) statistic, which is based on asymmetric and symmetric trimming, will be notated as YUEN(A) and YUEN(S), respectively.

Method

Sample size was set at eight per group; that is $n_1 = n_2 = 8$. Thus, for symmetric trimming there were six observations per group, whereas for asymmetric trimming there were seven observations per group. Sample size had to be restricted to this number because the amount of mainframe computer time required to perform a simulation was prohibitive because of the permutation tests.

We manipulated three variables in the study: (a) population distribution, (b) magnitude of the nonnull treatment effects, and (c) type of trimming with skewed distributions. In the current investigation, devoted primarily to exploring power differences between the tests, population variances were equal.

With respect to the effects of distributional shape, we chose to investigate conditions in which the data were obtained from a wide variety of nonnormal distributions. In addition to generating data from χ_3^2 and χ_6^2 distributions, we also used the method described in Hoaglin (1985) to generate distributions with more extreme degrees of skewness and kurtosis. We selected these particular types of nonnormal distributions because education and psychological research data typically have skewed distributions (Micceri, 1989; Wilcox, 1994). Furthermore, Sawilowsky and Blair (1992) investigated the effects of eight nonnormal distributions identified by Micceri on the robustness of Student's *t* test and found that only distributions with the most extreme degree of skewness that were investigated (e.g., $\gamma_1 = 1.64$) affected the Type I error control of the independent sample *t* statistic. Thus, because the statistics we investigated had operating characteristics similar to those reported for the *t* statistic, we felt that our approach to modeling skewed data would adequately reflect conditions in which those statistics might perform optimally. For the χ_3^2 distribution, skewness and kurtosis values

were $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$, respectively. We included the χ_6^2 distribution in our investigation to examine the effects of sampling from a distribution with moderate skewness. For this distribution, $\gamma_1 = 1.16$ and $\gamma_2 = 2.00$. The other types of nonnormal distributions were generated from the *g*- and *h*-distributions (Hoaglin, 1985). Specifically, we investigated four nonnormal *g*- and *h*-distributions; Table 1 contains an enumeration of the investigated distributions as well as measures of skewness and kurtosis. To give meaning to the values contained in Table 1, we note that for the standard normal distribution, $g = h = 0$. Thus, when $g = 0$ a distribution is symmetric and the tails of a distribution become heavier as *h* increases in value. Finally, although the selected combinations of *g* and *h* result in extremely skewed distributions, according to Wilcox (1994), these values are representative of psychometric measures.

The second variable manipulated was the magnitude of nonnull treatment effects. In particular, we chose three cases. We selected mean values such that the a priori power to detect differences in the two samples (based on the noncentrality parameter for the two-sample *t* test) would be .40, .60, or .80. That is, we intended to explore whether the specialized tests would outperform the usual test statistics when there were small, medium, or large differences between the two samples.

The third variable manipulated in this investigation for the test (YUEN) that involved trimming was the type of trimming in asymmetric distributions. When the simulated data were skewed, we investigated both symmetric and asymmetric trimming. Asymmetric trimming has been recommended by Tiku (1980, 1982) and others (e.g., see De Wet & van Wyk, 1979) for nonsymmetric skewed distributions as a means of reducing the effects of deviant observations in the longer tail. When trimming asymmetrically, we trimmed 20% from the longer tail of each group's set of scores. For symmetric trimming, we removed 20% of the observations from each tail of a group's set of scores (see Rosenberger & Gasko, 1983; Wilcox, 1994, 1995a, 1995b). This rule is based in part on optimizing power for nonnormal as well as normal distributions (see Wilcox, 1994, 1995a, 1995b).

TABLE 1
Distributions Investigated and Their Properties

Distribution	Skewness	Kurtosis
$g = 0/h = 0$	0.00	0.00
$g = 0/h = .2$	0.00	36.00
Chi-square (6)	1.16	2.00
Chi-square (3)	1.63	4.00
$g = 1/h = 0$	6.20	114
$g = 0/h = .5$	0.00	—
$g = 1/h = .5$	—	—

Note. Cells notated with — indicate an undefined value.

To generate pseudo-random normal variates, we used the SAS generator RANNOR (SAS Institute, 1989). If Z_{ij} is a standard normal variate, then $X_{ij} = \mu_j + \sigma_j \times Z_{ij}$ is a normal variate with mean equal to μ_j and variance equal to σ_j^2.

To generate pseudo-random variates having a chi-square distribution with three (six) degrees of freedom, we squared and summed three (six) standard normal variates. The variates were standardized and then transformed to χ_3^2 or χ_6^2 variates having mean μ_j (when comparing the tests based on the least squares estimates) or μ_{tj} (when comparing the tests based on trimmed means) and variance σ_j^2 (see Hastings & Peacock, 1975, pp. 46–51, for further details on the generation of data from these distributions).

To generate data from *g*- and *h*-distributions, standard unit normal variables (Z) were converted to the random variable

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

according to the values of g and h selected for investigation. To obtain a distribution with standard deviation σ_j, we multiplied each X_{ij} ($j = 1, \ldots, J$) by a value of σ_j. This operation does not affect the value of the null hypothesis when $g = 0$ (see Wilcox, 1994, p. 297). However, when $g > 0$, the population mean for a *g*- and *h*-distributed variable is

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

(see Hoaglin, 1985, p. 503). Thus, for those conditions where $g > 0$, μ_{gh} was first subtracted from X_{ij} before being multiplied by σ_j. When we were working with trimmed means, we first subtracted μ_{tj} from each observation.

We performed 1,000 replications of each condition, using a .05 significance level. We performed this number of replications because the two nonparametric tests, which were based on permutations of the data, required 18 hr per condition to execute on a mainframe computer. Type I error and power rates were computed within each condition examined. Type I rates were collected when the null hypothesis of no treatment effects was true, whereas power rates were assessed when treatment effects existed between the two groups.

Results

Table 2 contains Type I error (a priori power = 0) and power rates (a priori power = .4, .6, and .8) for the four test statistics when sampling from the seven empirical distributions. We show the Type I error rates to assure the reader that power rates were not being compared under conditions in which the tests were unable to maintain Type I error control (see Penfield, 1994). Indeed, as expect-

TABLE 2
Empirical Type I Error and Power Rates ($n_1 = n_2 = 8$)

Distribution	Power	WILC	RSKEW	YUEN(A)	YUEN(S)
$g = 0/h = 0$	0	.042	.035		.039
	.4	37	28		34
	.6	57	44		54
	.8	77	62		72
	P-mean	57	45		53
$g = 0/h = .2$	0	.042	.035		.030
	.4	30	22		27
	.6	45	34		44
	.8	61	47		60
	P-mean	45	34		44
Chi-square (6)	0	.045	.041	.041	.044
	.4	47	47	8	6
	.6	68	71	10	8
	.8	78	86	14	13
	P-mean	64	68	11	9
Chi-square (3)	0	.055	.050	.045	.043
	.4	51	60	12	9
	.6	67	78	19	15
	.8	82	90	30	24
	P-mean	67	76	20	16
$g = 1/h = 0$	0	.042	.035	.029	.025
	.4	37	55	32	26
	.6	54	73	53	42
	.8	65	84	66	55
	P-mean	52	70	50	41
$g = 0/h = .5$	0	.042	.035		.024
	.4	23	18		20
	.6	35	27		32
	.8	47	36		47
	P-mean	35	27		33
$g = 1/h = .5$	0	.042	.035	.019	.019
	.4	26	29	17	17
	.6	35	43	29	27
	.8	47	55	40	37
	P-mean	36	42	29	27
	G-mean	51	52	28	32 (23)

Note. When power = 0, the rates are Type I error; otherwise, rates are power values (expressed without %). P-mean = mean for power rates. G-mean = grand mean. Yuen(S) G-mean value in parentheses was obtained by averaging over four P-mean values [Chi-square (6), Chi-square (3), $g = 1/h = 0$, and $g = 1/h = .5$]. WILC = Wilcoxon–Mann–Whitney rank sum test; RSKEW = specialized nonparametric test designed for positively skewed data; YUEN(A) and YUEN(S) = two versions of Yuen's (1974) modification of the Welch test with trimmed means and Winsorized variances.

ed, all Type I error rates were very close to the .05 level of significance that was adopted to assess statistical significance.

The power results in Table 2 indicate that (a) the specialized nonparametric test RSKEW was typically more sensitive in detecting treatment effects when the empirical data were skewed; (b) the specialized statistic based on asymmetric trimming [YUEN(A)], though more powerful than its symmetric trimming counterpart [YUEN(S)] for skewed distributions, was not as sensitive as the nonparametric procedures (either WILC or RSKEW); and (c) averaged over all nonnull distributions investigated (grand mean, or G-mean), the nonparametric tests resulted in similar power, and both were substantially more powerful than the tests based on trimmed means (see Table 2).

In reference to Point a, the specialized nonparametric test RSKEW, as expected, was also less powerful than the usual Wilcoxon–Mann–Whitney test when the underlying distributions were symmetric. This finding also was not unexpected given that RSKEW was designed to be sensitive to detecting differences when distributions are skewed to the right. With regard to Point b, power differences favoring asymmetric trimming when data were positively skewed were not as large as the differences favoring RSKEW over WILC. Indeed, averaged over all the nonsymmetric distributions (i.e., when $g \neq 0$), the rates were 28 and 23 for asymmetric versus symmetric trimming, respectively. Finally, in reference to Point c, when sample sizes were small, the nonparametric tests resulted in substantially more power to detect treatment effects than did the Yuen (1974) test, which incorporates trimmed means. Our comparison between the nonparametric and trimmed means tests did not involve exactly comparable sample sizes (sample sizes were equal before trimming); thus, the trimmed tests were likely to have resulted in lower power values, although the power differences were greater than what we would have predicted based on sample-size disparity. Nonetheless, it will be useful for researchers who are confronted with skewed distributions and are searching for alternatives to the classical procedure to know that our data indicate that for a fixed small sample size (e.g., eight observations per group), nonparametric procedures will likely result in substantially greater sensitivity to detect treatment effects than will tests involving trimmed means.

To determine whether these results were generalizable beyond our sample size case of $n_1 = n_2 = 8$, we replicated our study with $n_1 = n_2 = 24$. For this sample size, permutation tests (WILC and RSKEW) cannot be empirically investigated with simulation methods because SAS/IML (SAS Institute, 1989) does not allow parallel processing. On a nonparallel processing mainframe computer, the simulation cannot be completed. However, we compared the trimmed procedures with normal approximation versions of the permutation tests. That is, for large sample sizes, WILC and RSKEW can be computed with normal approximation statistics (see Bradley, 1968, p. 111, and Hogg et al., 1975). The methods of the simulation

TABLE 3
Empirical Type I Error and Power Rates ($n_1 = n_2 = 24$)

Distribution	Power	WILC	RSKEW	YUEN(A)	YUEN(S)
$g = 0/h = 0$	0	.052	.060		.054
	.4	41	34		39
	.6	57	50		59
	.8	77	70		77
	P-mean	58	51		58
$g = 0/h = .2$	0	.052	.060		.053
	.4	31	28		32
	.6	47	41		48
	.8	64	57		67
	P-mean	47	42		49
Chi-square (6)	0	.055	.047	.053	.050
	.4	46	56	8	7
	.6	64	75	12	9
	.8	84	91	13	11
	P-mean	65	74	11	9
Chi-square (3)	0	.043	.040	.042	.043
	.4	57	73	17	14
	.6	77	89	22	17
	.8	91	98	29	21
	P-mean	75	87	23	17
$g = 1/h = 0$	0	.052	.060	.042	.040
	.4	48	72	41	30
	.6	64	89	59	47
	.8	79	96	77	61
	P-mean	64	85	59	46
$g = 0/h = .5$	0	.052	.060		.045
	.4	26	24		27
	.6	37	32		41
	.8	52	47		54
	P-mean	38	34		41
$g = 1/h = .5$	0	.052	.060	.034	.031
	.4	29	39	22	21
	.6	41	57	34	33
	.8	55	75	44	44
	P-mean	42	57	33	33
	G-mean	56	61	32	36 (26)

Note. When power = 0, the rates are Type I error; otherwise, rates are power values (expressed without %). P-mean = mean for power rates. G-mean = grand mean. Yuen(S) G-mean value in parentheses was obtained by averaging over four P-mean values [Chi-square (6), Chi-square (3), $g = 1/h = 0$, and $g = 1/h = .5$]. WILC = Wilcoxon–Mann–Whitney rank sum test; RSKEW = specialized nonparametric test designed for positively skewed data; YUEN(A) and YUEN(S) = two versions of Yuen's (1974) modification of the Welch test with trimmed means and Winsorized variances.

remained the same except for sample sizes. (Means were rescaled to achieve the a priori values of .4, .6, and .8.) The results for $n_1 = n_2 = 24$ are presented in Table 3.

The findings reported for Table 2 do indeed hold for larger sample sizes (see Table 3). However, the G-mean WILC and RSKEW values were slightly more divergent, favoring RSKEW, and power differences favoring the nonparametric tests were slightly larger.

Discussion

Because nonnormality, in particular, skewness of the data, negatively affects the ability of Student's two-sample *t* test to detect treatment effects, researchers have recommended adopting an alternative test statistic (see Penfield, 1994). Popular alternatives include nonparametric methods (e.g., Wilcoxon–Mann–Whitney rank sum test) as well as methods that use robust estimators (e.g., trimmed means).

When data are known to be skewed, researchers can adopt tests that have been designed to be more powerful to detect treatment effects. For example, Randles and Wolfe (1979; see also Berry, 1995) presented a nonparametric test, RSKEW, that was designed for positively skewed data. On the other hand, tests using robust estimators, such as trimmed means and Winsorized variances, have also been developed to capitalize on data that take a particular form, that is, positively skewed.

In our investigation, we compared four test statistics in the two-sample problem; two tests, one nonparametric and one involving trimmed means and Winsorized variances, did not take the shape of the data distributions into account, whereas the remaining two tests, again one nonparametric and one involving trimmed means and Winsorized variances, did.

Our results indicated that, as expected, researchers can achieve greater power to detect treatment effects when using the specialized tests with data that are skewed to the right. However, readers should remember the caveat that if distributions are not skewed to the right, these specialized tests will have less power than the nonspecialized nonparametric and trimmed-means alternatives defined in this article. On balance, that is, averaged over all the nonnormal distributions investigated, there was either no difference or very little difference favoring the specialized tests. Accordingly, based on the conditions explored in this investigation, we do not recommend that researchers adopt either of the two specialized tests. What researchers should do, however, is seriously consider abandoning the classical methods in favor of some alternative that will provide both reliable and valid rates of Type I and Type II errors. With regard to power superiority, our results indicate that researchers can expect substantial gains when adopting the nonparametric approach as compared with tests based on trimmed means.

NOTE

This research was supported by a Natural Sciences and Engineering Research Council of Canada grant (OGP0015855) to the first author.

REFERENCES

Berry, J. J. (1995). Obtaining exact significance levels for various nonparametric two-independent samples problems. *Observations, Fourth Quarter,* 40–52.

Bradley, J. V. (1968). *Distribution-free statistical tests.* Englewood Cliffs, NJ: Prentice-Hall.

De Wet, T., & van Wyk, J. W. J. (1979). Efficiency and robustness of Hogg's adaptive trimmed means. *Communications in Statistics, Theory and Methods, A8*(2), 117–128.

Gross, A. M. (1976). Confidence interval robustness with long-tailed symmetric distributions. *Journal of the American Statistical Association, 71,* 409–416.

Hastings, N. A. J., & Peacock, J. B. (1975). *Statistical distributions: A handbook for students and practitioners.* New York: Wiley.

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.

Hogg, R. V., Fisher, D. M., & Randles, R. H. (1975). A two-sample adaptive distribution-free test. *Journal of the American Statistical Association, 70,* 656–661.

Lind, J. C., & Zumbo, B. D. (1993). The continuity principle in psychological research: An introduction to robust statistics. *Canadian Psychology, 34,* 407–414.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one or two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 1947, 18,* 50–60.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA: Brooks/Cole.

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

Penfield, D. A. (1994). Choosing a two-sample location test. *The Journal of Experimental Education, 62*(4), 343–360.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics.* New York: Wiley.

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians and trimean. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 297–336). New York: Wiley.

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

Tiku, M. L. (1980). Robustness of MML estimators based on censored samples and robust test statistics. *Journal of Statistical Planning and Inference, 4,* 123–143.

Tiku, M. L. (1982). Robust statistics for testing equality of means and variances. *Communications in Statistics, Theory and Methods, 11*(22), 2543–2558.

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

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

Wilcox, R. R. (1995a). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research, 65*(1), 51–77.

Wilcox, R. R. (1995b). 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.

Wilcoxon, F. (1949). *Some rapid approximate statistical procedures.* Stamford, CT: Stamford Research Laboratories, American Cyanamid Company.

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