Research Article

The New and Improved Two-Sample *t* Test

H.J. Keselman,[1] Abdul R. Othman,[2] Rand R. Wilcox,[3] and Katherine Fradette[1]

[1]University of Manitoba, Winnipeg, Manitoba, Canada; [2]Universiti Sains Malaysia, Penang, Malaysia; and [3]University of Southern California

ABSTRACT—*This article considers the problem of comparing two independent groups in terms of some measure of location. It is well known that with Student's two-independent-sample* t *test, the actual level of significance can be well above or below the nominal level, confidence intervals can have inaccurate probability coverage, and power can be low relative to other methods. A solution to deal with heterogeneity is Welch's (1938) test. Welch's test deals with heteroscedasticity but can have poor power under arbitrarily small departures from normality. Yuen (1974) generalized Welch's test to trimmed means; her method provides improved control over the probability of a Type I error, but problems remain. Transformations for skewness improve matters, but the probability of a Type I error remains unsatisfactory in some situations. We find that a transformation for skewness combined with a bootstrap method improves Type I error control and probability coverage even if sample sizes are small.*

An issue that has received considerable attention is whether any method for comparing measures of location, corresponding to two independent groups, can provide reasonably accurate control over the probability of a Type I error when distributions are nonnormal and there is heteroscedasticity. The literature pertaining to the effects of variance heterogeneity and nonnormality is extensive, beginning well over 50 years ago (see, e.g., Aspin, 1949; Behrens, 1929; Brown & Forsythe, 1974; Fisher, 1935; James, 1951; Satterthwaite, 1941; Tomarken & Serlin, 1986; Welch, 1938; Wilcox, 1990). It is well known that Student's two-independent-sample t test can be highly unsatisfactory in this regard and that it can have poor power under arbitrarily small departures from normality. In fact, under general conditions, it is not even asymptotically correct (Cressie & Whitford, 1986). Heteroscedastic methods for means improve the control over the probability of a Type I error and are asymptotically correct, but problems remain (see Wilcox, 1997), and any method based on means can have relatively low power (e.g., Staudte & Sheather, 1990). As Marazzi and Ruffieux (1999) noted, "the (usual) mean is a difficult parameter to estimate well: the sample mean, which is the natural estimate, is very nonrobust" (p. 79). Yuen (1974) derived a generalization of Welch's (1938) heteroscedastic method for means to trimmed means. Asymptotic results, plus simulations, indicate that this method improves control over the probability of a Type I error (Wilcox, 1997, 2003), but problems remain.

Another development in this area was to apply a transformation to a heteroscedastic statistic to eliminate the biasing effects of skewness. Indeed, Luh and Guo (1999) and Guo and Luh (2000) demonstrated that better Type I error control was possible when transformations (Hall's, 1992, or Johnson's, 1978, method) were applied to the Welch (1938) statistic with trimmed means.

In this article, we consider combining transformations with trimmed means and a bootstrap method to assess significance and illustrate that good control over the Type I error probability is possible with small sample sizes.

Address correspondence to H.J. Keselman, Department of Psychology, The University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2; e-mail: kesel@ms.umanitoba.ca.

THE TWO-SAMPLE TEST

To test H_0: $\mu_{t1} = \mu_{t2}$ (equality of population trimmed means), let $d_j = \frac{(n_j-1)\hat{\sigma}^2_{wj}}{h_j(h_j-1)}$, where $\hat{\sigma}^2_{wj}$ is the gamma-Winsorized variance and h_j is the effective sample size, that is, the size after trimming ($j = 1, 2$) (Appendix A defines the Winsorization process). Yuen's (1974) test is

$$t_y = \frac{\hat{\mu}_{t1} - \hat{\mu}_{t2}}{\sqrt{d_1 + d_2}},$$

where $\hat{\mu}_{tj}$ is the γ-trimmed mean for the jth group (see Appendix A) and the estimated degrees of freedom are

$$\nu_y = \frac{(d_1 + d_2)^2}{d_1^2/(h_1 - 1) + d_2^2/(h_2 - 1)}. \quad (1)$$

TRANSFORMATIONS FOR YUEN'S STATISTIC

Guo and Luh (2000) and Luh and Guo (1999) found that Johnson's (1978) and Hall's (1992) transformations improved the performance of several heteroscedastic test statistics when they were used with trimmed means in the presence of heavy-tailed and skewed distributions. Johnson (1978) used the Cornish-Fisher expansion (e.g.,

see Balkin & Mallows, 2001) to modify the one-sample t test in order to achieve robustness to skewness, whereas Hall (1992) provided a general transformation for removing skewness.

Let $Y_{ij} = (Y_{1j}, Y_{2j}, \ldots, Y_{n_jj})$ be a random sample from the jth distribution ($i = 1, \ldots, n_j$; $j = 1, 2$). Let $\hat{\mu}_{tj}$, $\hat{\mu}_{wj}$, and $\hat{\sigma}^2_{wj}$ be, respectively, the trimmed mean, Winsorized mean, and Winsorized variance of group j (see Appendix A). Define the Winsorized third central moment of group j as

$$\hat{\mu}_{3j} = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_{ij} - \hat{\mu}_{wj})^3,$$

where the X_{ij}s are the Winsorized scores (see Appendix A). Let

$$\tilde{\mu}_{wj} = \frac{n_j}{h_j}\hat{\mu}_{3j},$$

$$\tilde{\sigma}^2_w = d_1 + d_2,$$

and

$$\tilde{\mu}_w = \frac{\tilde{\mu}_{31}}{h_1^2} - \frac{\tilde{\mu}_{32}}{h_2^2}.$$

Guo and Luh (2000) generalized Yuen's (1974) test statistic via Johnson's (1978) transformation, yielding

$$t_{y(\text{Johnson})} = \frac{(\hat{\mu}_{t1} - \hat{\mu}_{t2}) + \frac{\tilde{\mu}_w}{6\tilde{\sigma}^2_w} + \frac{\tilde{\mu}_w}{3\tilde{\sigma}^4_w}(\hat{\mu}_{t1} - \hat{\mu}_{t2})^2}{\tilde{\sigma}_w},$$

whereas Yuen's statistic with Hall's (1992) transformation would be

$$t_{y(\text{Hall})} = \frac{(\hat{\mu}_{t1} - \hat{\mu}_{t2}) + \frac{\tilde{\mu}_w}{6\tilde{\sigma}^2_w} + \frac{\tilde{\mu}_w}{3\tilde{\sigma}^4_w}(\hat{\mu}_{t1} - \hat{\mu}_{t2})^2 + \frac{\tilde{\mu}^2_w}{27\tilde{\sigma}^8_w}(\hat{\mu}_{t1} - \hat{\mu}_{t2})^3}{\tilde{\sigma}_w}.$$

Each of these statistics is distributed approximately as a t variable with degrees of freedom given in Equation 1.

BOOTSTRAPPING

Now we consider how extensions of the method just outlined might be improved. Investigations have indicated that when using a γ-trimmed mean, the most successful method, in terms of probability coverage and controlling the probability of a Type I error, is some type of bootstrap method.

Following Westfall and Young (1993), and as enumerated by Wilcox (1997), let $C_{ij} = Y_{ij} - \hat{\mu}_{tj}$; thus, the C_{ij} values are the empirical distribution of the jth group, centered so that the sample trimmed mean is zero. That is, the empirical distributions are shifted so that the null hypothesis of equal trimmed means is true in the sample. The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value. For each j, obtain a bootstrap sample by randomly sampling with replacement n_j observations from the C_{ij} values, yielding $Y^*_1, \ldots, Y^*_{n_j}$. Let, for example, t^*_y be the value of Yuen's (1974) test based on the bootstrap sample. Now we randomly sample (with replacement) B bootstrap samples from the shifted distributions, each time calculating the statistic t^*_y. The B values of t^*_y are put in ascending order, that is, $t^*_{y(1)} \leq \cdots \leq t^*_{y(B)}$. If we set $l = \alpha B/2$, rounding to the nearest integer, and $u = B - l$, then we would reject the null hypothesis of location equality (i.e., H_0: $\mu_{t1} = \mu_{t2}$) when $t_y \leq t^*_{y(l)}$ or when $t_y \geq t^*_{y(u)}$, where t_y is the value of Yuen's statistic based on the original nonbootstrapped data.

BOOTSTRAP INTERVALS FOR $\mu_{t1} - \mu_{t2}$

The $100(1-\alpha)\%$ bootstrap percentile interval for $\mu_{t1} - \mu_{t2}$ based on the Yuen procedure would be

$$\left[(\hat{\mu}_{t1} - \hat{\mu}_{t2}) - t^*_{y(u)}\tilde{\sigma}_w,\ (\hat{\mu}_{t1} - \hat{\mu}_{t2}) - t^*_{y(l+1)}\tilde{\sigma}_w\right].$$

If Johnson's (1978) transformation is applied to the Yuen approach, then according to Johnson (p. 538), a $100(1-\alpha)\%$ bootstrap percentile interval would be

$$\left[\left\{(\hat{\mu}_{t1} - \hat{\mu}_{t2}) + \frac{\tilde{\mu}_w}{6\tilde{\sigma}^2_w}\right\} - t^*_{y(\text{Johnson})(u)}\tilde{\sigma}_w,\right.$$

$$\left.\left\{(\hat{\mu}_{t1} - \hat{\mu}_{t2}) + \frac{\tilde{\mu}_w}{6\tilde{\sigma}^2_w}\right\} - t^*_{y(\text{Johnson})(l+1)}\tilde{\sigma}_w\right].$$

Finally, based on Guo and Luh (2000), a $100(1-\alpha)\%$ bootstrap percentile interval for Hall's (1992) approach would be

$$\left[(\hat{\mu}_{t1} - \hat{\mu}_{t2}) - \tilde{\sigma}_w B^*_{(u)},\ (\hat{\mu}_{t1} - \hat{\mu}_{t2}) - \tilde{\sigma}_w B^*_{(l+1)}\right],$$

where

$$B^*_{(l+1)} = \frac{3}{\hat{\nu}}\left[1 + \hat{\nu}\left(t^*_{y(\text{Hall})(l+1)}\right) - \frac{\hat{\nu}}{6}\right]^{1/3} - \frac{3}{\hat{\nu}},$$

and

$$B^*_{(u)} = \frac{3}{\hat{\nu}}\left[1 + \hat{\nu}(t^*_{y(\text{Hall})(u)}) - \frac{\hat{\nu}}{6}\right]^{1/3} - \frac{3}{\hat{\nu}},$$

where $\hat{\nu} = \tilde{\mu}_w/\tilde{\sigma}^3_w$.

THE SIMULATION

We examined Yuen's (1974) procedure with and without a transformation for skewness and with and without bootstrapping. In particular, the procedures were compared for (a) three nonnormal distributions (a chi-squared distribution with 3 degrees of freedom and two g-and-h distributions; see Hoaglin, 1985); (b) unequal group variances that were in a 36:1 ratio; (c) variances and group sizes that were both positively and negatively paired for two cases of sample size, $N = 30$ (10, 20) and $N = 40$ (15, 25); and (d) three percentages of trimming (20%, 15%, and 10%). The recommended amount of symmetric trimming varies in the literature. Rosenberger and Gasko (1983) recommended 25% trimming when sample sizes are small, though they indicated that generally 20% suffices. Wilcox (1997) also recommended 20%, whereas Mudholkar, Mudholkar, and Srivastava (1991) suggested 15%. Ten percent has been suggested by Hill and Dixon (1982), Huber (1977), Stigler (1977), and Staudte and Sheather (1990). Details of the simulation are presented in Appendix B.

RESULTS

Table 1 contains summary statistics for the 18 versions of t_y that were examined. In particular, the table contains the range of empirical Type

TABLE 1

Summary Statistics

Result	Procedure					
	t_y	t_yJ	t_yH	t_yB	t_yJB	t_yHB
	20% symmetric trimming					
Range of Type I errors	.042–.075	.043–.071	.043–.071	.044–.061	.042–.06	.045–.059
No. of values not in the interval	9	8	8	2	3	2
Average Type I error	.058	.056	.056	.052	.053	.053
	15% symmetric trimming					
Range of Type I errors	.038–.07	.044–.063	.044–.064	.038–.058	.045–.059	.044–.056
No. of values not in the interval	7	5	5	2	2	0
Average Type I error	.054	.055	.055	.05	.053	.052
	10% symmetric trimming					
Range of Type I errors	.042–.077	.051–.068	.051–.069	.043–.06	.05–.06	.049–.061
No. of values not in the interval	8	5	6	4	3	1
Average Type I error	.056	.056	.057	.051	.054	.054

Note. Nonrobust values are those not contained in the interval .044–056. t_y = Yuen's two-sample test; J = Johnson's (1978) transformation; H = Hall's (1992) transformation; B = bootstrapping.

I error values for the 12 conditions examined (3 distributions × 2 sample-size cases × 2 pairings of variances and sample sizes); the number of values, out of 12, that were not contained in the interval .044–.056 ($\pm 2\sigma_\alpha$ for $\alpha = .05$); and the average error rate over the conditions examined. We consider empirical values not contained in the interval .044–.056 to be nonrobust.

As we indicated in our introduction, applying a transformation to a heteroscedastic statistic and assessing significance through a bootstrap method improves rates of Type I error over the rates obtained with methods that do not use these modifications. Also noteworthy is that the "best" procedure is one that relies on a moderate amount of trimming (i.e., 15% from each tail). That is, Yuen's (1974) method with Hall's (1992) transformation with bootstrapping methodology resulted in no empirical values outside the .044–.056 interval. It is important to note that this level of accuracy in Type I error control has not been reported in the literature for any other method related to the problem we investigated. Moreover, the range of empirical values for this test (range = .012) was the lowest among the procedures that resulted in no more than two deviant values. The same procedure with 10% trimming might be regarded as best, however, even though it resulted in one deviant value (range = .012), if discarding the least amount of the data is primary to an applied researcher.

SUMMARY AND CONCLUSIONS

It is well known to statisticians that variance heterogeneity can distort rates of Type I error for Student's two-independent-sample t test. Hence, numerous solutions to the problem of assessing mean equality in the presence of variance heterogeneity have appeared over the decades since this problem was first identified. Also fairly well known is that nonnormality, typically in the form of heavy tailedness, can depress the power to detect effects when the usual mean and variance are used to assess treatment-group equality with Student's two-independent-sample t test or with Welch's (1938) test. Thus, Yuen (1974) suggested that in order to counter the effects of nonnormality and variance heterogeneity, applied researchers should apply trimmed means and Winsorized variances with Welch's two-sample test. Indeed, Yuen found that one can achieve better Type I error control with her procedure, and many other statisticians have indicated that the power to detect effects is better with her procedure than it would be using either Student's t test or Welch's test based on least squares means and variances (e.g., see Wilcox, 1997).

Recent research has indicated that applied researchers can achieve even better Type I error control if Yuen's (1974) test is modified by applying to her statistic a transformation that eliminates the effects of skewness and if statistical significance is assessed with a bootstrap method. Accordingly, we have outlined the mechanics of this approach and presented the results from a Monte Carlo investigation that demonstrated its effectiveness. We have also demonstrated that very tight Type I error control can be achieved with only modest amounts of trimming—namely, 15% or 10% from each tail of the distribution. This last finding we consider noteworthy because most applied researchers are generally uncomfortable about discarding data.

We want to remind the reader that we examined 18 test statistics under conditions of extreme heterogeneity and nonnormality. Thus, we believe we have identified procedures that maintain the actual α level in cases of heterogeneity and nonnormality likely to be encountered by applied researchers, and therefore we are very comfortable with our recommendation.

To conclude, we wish to reiterate that though the Yuen (1974) statistic tests a null hypothesis stipulating that the population trimmed means are equal, we believe this is a reasonable hypothesis to examine because in distributions that either contain outliers or are skewed, trimmed means provide better estimates of the typical individual than the usual (least squares) means do. That is, when distributions are skewed, trimmed means do not estimate μ but rather some value (i.e., μ_t) that is typically closer to the bulk of the observations. (Another way of conceptualizing the unknown parameter μ_t is that it is simply the population counterpart of $\hat{\mu}_t$; see Huber, 1972, and Hogg, 1974.) Finally, as Zhou, Gao, and Hui (1997) pointed out, distributions are typically skewed, and our results indicate that a

heteroscedastic statistic utilizing trimmed means and Winsorized variances will provide very good control over the Type I error probability for a broader range of situations when applied with a transformation to eliminate skewness and a bootstrap method.

Acknowledgments—Work on this project was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Aspin, A.A. (1949). Tables for use in comparisons whose accuracy involves two variances separately estimated. *Biometrika*, *36*, 290–293.

Balkin, S.D., & Mallows, C.L. (2001). An adjusted, asymmetric two-sample t test. *The American Statistician*, *55*(3), 203–206.

Behrens, W.V. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtsch Jahrbucher*, *68*, 807–837.

Brown, M.B., & Forsythe, A.B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, *16*, 129–132.

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

Fisher, R.A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, *6*, 391–398.

Guo, J.H., & Luh, W.M. (2000). An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. *Statistics & Probability Letters*, *49*, 1–7.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, *14*, 1431–1452.

Hall, P. (1992). On the removal of skewness by transformation. *Journal of the Royal Statistical Society, Series B*, *54*, 221–228.

Hill, M., & Dixon, W.J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics*, *38*, 377–396.

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. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association*, *69*, 909–927.

Huber, P.J. (1972). Robust statistics: A review. *Annals of Mathematical Statistics*, *43*, 1041–1067.

Huber, P.J. (1977). Discussion. *The Annals of Statistics*, *5*, 1090–1091.

James, G.S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, *38*, 324–329.

Johnson, N.J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, *73*, 536–544.

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

Luh, W., & Guo, J. (1999). A powerful transformation trimmed mean method for one-way fixed effects ANOVA model under non-normality and inequality of variances. *British Journal of Mathematical and Statistical Psychology*, *52*, 303–320.

Marazzi, A., & Ruffieux, C. (1999). The truncated mean of an asymmetric distribution. *Computational Statistics & Data Analysis*, *32*, 79–100.

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.

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.

Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika*, *6*, 309–316.

Staudte, R.G., & Sheather, S.J. (1990). *Robust estimation and testing*. New York: Wiley.

Stigler, S.M. (1977). Do robust estimators work with real data? *The Annals of Statistics*, *5*, 1055–1098.

Tomarken, A.J., & Serlin, R.C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*, 90–99.

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

Westfall, P.H., & Young, S.S. (1993). *Resampling-based multiple testing*. New York: Wiley.

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

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

Wilcox, R.R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R.R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.

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.

(Received 10/4/02; Revision accepted 2/13/03)

APPENDIX A: TRIMMING AND ESTIMATORS

Let $Y_{(1)j} \leq Y_{(2)j} \leq \cdots \leq Y_{(n_j)j}$ represent the ordered observations associated with the jth group. Let $g_j = [\gamma n_j]$ indicate that γn_j is rounded down to the nearest integer and γ 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 $h_j = n_j - 2g_j$. The jth sample trimmed mean is

$$\hat{\mu}_{tj} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j-g_j} Y_{(i)j}.$$

The sample Winsorized mean is necessary to compute the Winsorized variance and is computed as

$$\hat{\mu}_{wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij},$$

where

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

The sample Winsorized variance, which is required to get a theoretically valid estimate of the standard error of a trimmed mean, is then given by

$$\hat{\sigma}^2_{wj} = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \hat{\mu}_{wj})^2.$$

The standard error of the trimmed mean is estimated with

$$\sqrt{(n_j - 1)\hat{\sigma}^2_{wj}/[h_j(h_j - 1)]}.$$

APPENDIX B: THE SIMULATION STUDY

Eighteen tests for treatment-group equality were compared for their rates of Type I error under conditions of nonnormality and variance heterogeneity in an independent-groups design with two treatments. The procedures we investigated were Yuen's (1974) two-sample test, Yuen's two-sample test with bootstrapping, Yuen's two-sample test with Johnson's (1978) transformation, Yuen's two-sample test with Johnson's transformation and bootstrapping, Yuen's two-sample test with Hall's (1992) transformation, and Yuen's two-sample test with Hall's transformation and bootstrapping. Each of these procedures was implemented with 10%, 15%, and 20% trimming.

We examined (a) the percentage of symmetric trimming (10%, 15%, or 20%), (b) the utility of transforming the Yuen (1974) statistic with either Johnson's (1978) or Hall's (1992) transformation, and (c) the utility of bootstrapping the data. Three additional variables were manipulated in the study: (a) sample size, (b) pairing of variances and group sizes, and (c) population distribution.

The two cases of total sample size and the group sizes were $N = 30$ (10, 20) and $N = 40$ (15, 25). We selected these values, in part, because other researchers have found them to be generally sufficient to provide reasonably effective Type I error control (e.g., see Wilcox, 1994). The unequal variances were in a 36:1 ratio. Though a ratio of 36:1 may seem extreme, similar ratios, and larger ones, have been reported in the literature. For example, Keselman et al. (1998), after reviewing articles published in prominent education and psychology journals, noted that they found ratios as large as 24:1 in one-way completely randomized designs. Wilcox (2003) cited data sets in which the ratio was 17,977:1! Variances and group sizes were both positively and negatively paired. In positive pairs, the largest n_j was associated with the population having the largest variance; in negative pairs, the largest n_j was associated with the population having the smallest variance. These conditions were chosen because they typically produce conservative and liberal results, respectively.

With respect to the effects of distributional shape on Type I error, we chose to investigate nonnormal distributions in which the data were obtained from a variety of skewed distributions. In addition to generating data from a χ^2_3 distribution, we used the method described in Hoaglin (1985) to generate distributions with more extreme degrees of skewness and kurtosis. For the χ^2_3 distribution, skewness and kurtosis values are $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$, respectively. The other nonnormal distributions were generated from the g-and-h distribution (Hoaglin, 1985). Specifically, we chose to investigate two g-and-h distributions: (a) $g = .5$ and $h = 0$ and (b) $g = .5$ and $h = .5$, where g and h are parameters that determine the third and fourth moments of a distribution. When $g = 0$, a distribution is symmetric, and the tails of a distribution will become heavier as h increases in value. It should be noted that for the standard normal distribution, $g = h = 0$. Values of skewness and kurtosis corresponding to the investigated values of g and h are (a) $\gamma_1 = 1.75$ and $\gamma_2 = 8.9$, respectively, and (b) $\gamma_1 = \gamma_2 =$ undefined. These values of skewness and kurtosis are theoretical values; Wilcox (1997, p. 73) reported computer-generated values, based on 100,000 observations, for these statistics—$\hat{\gamma}_1 = 1.81$ and $\hat{\gamma}_2 = 9.7$ for $g = .5$ and $h = 0$, and $\hat{\gamma}_1 = 120.10$ and $\hat{\gamma}_2 = 18{,}393.6$ for $g = .5$ and $h = .5$. Thus, the conditions we chose to investigate could be described as extreme. That is, they were intended to indicate the operating characteristics of the procedures under substantial departures from homogeneity and normality, with the premise being that if a procedure works under the most extreme conditions, it will probably work under most conditions likely to be encountered by researchers.

To obtain pseudorandom normal variates, we used the SAS generator RANNOR (SAS Institute, 1989). If Z_{ij} is a standard-unit normal variate, then $Y_{ij} = \mu_j + \sigma_j \times Z_{ij}$ is a normal variate with mean equal to μ_j and variance equal to σ_j^2. Pseudorandom variates having a chi-squared distribution with three degrees of freedom were generated by squaring and summing three standard normal variates.

To generate data from a g-and-h distribution, we converted standard-unit normal variables to random variables via

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

according to the values of g and h selected for investigation. We then subtracted μ_{ij} from the generated variates under every generated distribution. In particular, we generated one million observations from each of the distributions investigated and applied each possible trimming strategy (10%, 15%, or 20%), calculating the mean of the remaining values. We then used these mean values to standardize the data so that the null hypothesis of trimmed-mean equality was true in every null case investigated.

To obtain a distribution with standard deviation σ_j, we then multiplied each Y_{ij} by a value of σ_j. It should be noted that the standard deviation of a g-and-h distribution is not equal to 1, and thus the values reflect only the amount that each random variable is multiplied by and not the actual values of the standard deviations (see Wilcox, 1994, p. 298). As Wilcox noted, the values for the variances (standard deviations) more aptly reflect the ratio of the variances (standard deviations) between the groups.

Five thousand replications of each condition were performed using a .05 statistical significance level. B was set at 599 because the results of Wilcox (1997) and Hall (1986) suggest that it may be advantageous to choose B such that $1 - \alpha$ is a multiple of $(B+1)^{-1}$.