# A more realistic look at the robustness and Type II error probabilities of the Test to departures from population normality

**Article** *in* Psychological Bulletin · March 1992

DOI: 10.1037/0033-2909.111.2.352

Shlomo Sawilowsky
Wayne State University

QUANTITATIVE METHODS IN PSYCHOLOGY

# A More Realistic Look at the Robustness and Type II Error Properties of the *t* Test to Departures From Population Normality

Shlomo S. Sawilowsky
Educational Evaluation and Research, College of Education
Wayne State University

R. Clifford Blair
Department of Pediatrics, College of Medicine
and Department of Epidemiology/Biostatistics
College of Public Health, University of South Florida

The Type I and II error properties of the *t* test were evaluated by means of a Monte Carlo study that sampled 8 real distribution shapes identified by Micceri (1986, 1989) as being representative of types encountered in psychology and education research. Results showed the independent-samples *t* tests to be reasonably robust to Type I error when (a) sample sizes are equal, (b) sample sizes are fairly large, and (c) tests are two-tailed rather than one-tailed. Nonrobust results were obtained primarily under distributions with extreme skew. The *t* test was robust to Type II error under these nonnormal distributions, but researchers should not overlook robust nonparametric competitors that are often more powerful than the *t* test when its underlying assumptions are violated.

Along with Pearson's chi-squared test, the independent-samples *t* test must be counted among the best-known statistical procedures in current use. Given its familiarity and utility, it is not surprising that over the years, this test has received an inordinate amount of attention from statistical researchers. Much of this attention has focused on the question of robustness (or lack thereof) of the *t* statistic to departures from the underlying assumption of population normality.

Although there is some disagreement on the subject (see Bradley, 1978), the prevailing view seems to be that the independent-samples *t* test is reasonably robust, insofar as Type I errors are concerned, to non-Gaussian population shape so long as (a) sample sizes are equal or nearly so, (b) sample sizes are fairly large (Boneau, 1960, mentions sample sizes of 25 to 30), and (c) tests are two-tailed rather than one-tailed. Note also that when these conditions are met and differences between nominal alpha and actual alpha do occur, discrepancies are usually of a conservative rather than of a liberal nature. (See, e.g., Efron, 1969; Gayen, 1949, 1950; Geary, 1936, 1947; Pearson & Please, 1975. For related reviews of the literature and general discussion on this topic, see Blair, 1981; Glass, Peckham, & Sanders, 1972; Ito, 1980; Sawilowsky, 1990.)

Bradley (1968, 1977, 1982) raised a number of objections to these conclusions reached by previous researchers in this area. Among his objections is that distributions encountered in real research contexts may be much more radically nonnormal than the relatively tame population shapes typically used in robustness studies. Recently, Micceri (1989), in one of the most comprehensive studies of its kind to appear in the social and behavioral science literature, echoed and elaborated on this concern.

In his study, Micceri (1989) canvased educational and psychological sources to provide 440 large data sets, which he then used to estimate population characteristics. Micceri's findings dramatically supported Bradley's (1977) position. Indeed, of the 440 distributions investigated, virtually none could be characterized as Gaussian. The distributions studied produced widely varying patterns of tail weight and skew and often demonstrated varying degrees of digit preference and multimodality. Micceri concluded that previous studies of the robustness of the *t* test (as well as other statistics) failed to consider distributions of the types encountered in education and psychology research practice. He stated that "prior robustness studies have generally limited themselves either to computational evaluations of asymptotic theory or to Monte Carlo investigations of interesting mathematical functions" (Micceri, 1989, p. 163). Commenting on the often cited study of Boneau (1960), Micceri noted that almost none of the comparisons made in that study occurs with real-world data. The conclusion to be drawn is that the findings of previous researchers who modeled population shapes with convenient mathematical functions cannot, necessarily, be applied in educational and psychological research settings.

## Purpose

Given the findings and discussion provided by Micceri (1989), it seems obvious that the literature on the robustness of the *t* test to departures from population normality, though extensive, is nevertheless incomplete. The primary purpose of this study, then, is to investigate the robustness properties of the independent-samples *t* test when sampling is from distributions of the types identified by Micceri (1986).[1]

Note in addition that Micceri's (1989) appeal for more realis-

Correspondence concerning this article should be addressed to Shlomo S. Sawilowsky, P.O. Box 48023, Oak Park, Michigan 48237.

[1] The dependent-samples *t* test was also included in the Type I error portion of this study. Results of the simulation are available from Shlomo S. Sawilowsky.

Psychological Bulletin, 1992, Vol. 111, No. 2, 352–360
Copyright 1992 by the American Psychological Association, Inc. 0033-2909/92/$3.00

Table 1
*Descriptive Information Pertaining to Eight Real-World Distributions*

| Distribution | Type of measure | $\mu$ | Median | $\sigma$ | Skew | Kurtosis |
|---|---|---|---|---|---|---|
| Discrete mass at zero with gap | Psychometric | 1.85 | 0.00 | 3.80 | 1.65 | 3.98 |
| Mass at zero | Achievement | 12.92 | 13.00 | 4.42 | −0.03 | 3.31 |
| Extreme asymmetry | Psychometric | 13.67 | 11.00 | 5.75 | 1.64 | 4.52 |
| Extreme asymmetry | Achievement | 24.50 | 27.00 | 5.79 | −1.33 | 4.11 |
| Extreme bimodality | Psychometric | 2.97 | 4.00 | 1.69 | −0.08 | 1.30 |
| Multimodality and lumpy | Achievement | 21.15 | 18.00 | 11.90 | 0.19 | 1.80 |
| Digit preference | Achievement | 536.95 | 535.00 | 37.64 | −0.07 | 2.76 |
| Smooth symmetric | Achievement | 13.19 | 13.00 | 4.91 | 0.01 | 2.66 |

tic assessments of Type I error characteristics of statistical tests applies equally to Type II error (or power). With researchers relying more on power analyses and sample size determinations than in the past (Cohen, 1988), it has become increasingly important that these test characteristics also be evaluated in more realistic contexts. Treatments often produce changes in means, as well as variance, skew, tail weight, and other population parameters. Thus, in the second portion of this study, the robustness of the independent-samples *t* test with respect to Type II error is investigated for each of the eight prevalent distributions identified by Micceri (1986).

## Method

Monte Carlo methods were used to sample eight different distributions characterized by Micceri (1986) as being representative of the types found in his study.[2] Observations were sampled independently and with replacement from each of the eight distributions using the PC version of International Mathematical and Statistical Libraries (IMSL, 1987) RNSET and RNUND subroutines. We also sampled observations from a Gaussian distribution generated by subroutine NORMB1 from RANGEN (Blair, 1987) to demonstrate the adequacy of the simulation.

The Type I error portion of the study proceeded as follows: Independent samples of sizes $(n_1, n_2) = (5, 15), (10, 10), (10, 30), (20, 20), (15, 45), (30, 30), (20, 60), (40, 40), (30, 90)$, and $(60, 60)$ were generated. The independent-samples *t* test was computed on each sample pair. Simulations were carried out on an Intel 80386-based PC with an 80387 numeric coprocessor by means of a Microsoft FORTRAN 5.0 program. We performed ten thousand (10,000) repetitions for each condition studied.

The robustness of the independent-samples *t* test with respect to Type II error was investigated as follows: Let $X_{1i}$ and $X_{2j}$ be observations in two random samples taken from a common population with mean $\mu_1$ and standard deviation $\sigma_1$. The transformed variables $X'_{1i}$ and $X'_{2j}$ were generated by

$$X'_{1i} = X_{1i} - \mu_1; \quad i = 1, \cdots, n_1$$

$$X'_{2j} = c(X_{2j} - \mu_1) + k\sigma_1; \quad i = 1, \cdots, n_2$$

where *c* and *k* are constants. Cohen (1988) defined effect size (*ES*) for the unequal variance situation as

$$ES = (\mu'_2 - \mu'_1)/[(\sigma'^2_1 + \sigma'^2_2)/2]^{1/2},$$

where $\mu'_2$, $\mu'_1$, $\sigma'^2_1$, and $\sigma'^2_2$ are the population means and variances of $X'_{2j}$ and $X_{1i}$. In this case, $\mu'_1 = 0$, $\sigma'^2_1 = \sigma_1{}^2$, $\mu'_2 = k\sigma_1$, and $\sigma'^2_2 = c^2\sigma_1{}^2$.

Hypotheses of shift in location parameters were investigated by making $c = 1$ and $k$ equal to a constant of $0.2\sigma$, $0.5\sigma$, $0.8\sigma$, and $1.2\sigma$, where $\sigma$ represents the standard deviation of the distribution sampled, to $X_{2j}$. Hypotheses of shift in mean plus increase in variance were investigated by using values of *c* equal to $\sqrt{2}$, $\sqrt{3}$, and 2, and solving for *k* to produce an effect size of 0.50 for the balanced layouts studied. We did not investigate unbalanced layouts for $c \neq 1$, because Cohen (1988) noted that for conditions of both unequal variances and unequal sample sizes, the tabled power values "may be greatly in error" (p. 44). Although treatments modeled by a shift in location or by a shift in location combined with a change in scale are familiar to readers of Monte Carlo investigations of the properties of statistics, note that they are only a subset of reasonable treatment alternatives.

Table 1 contains certain descriptive information concerning the eight distributions. Skew and kurtosis measures are based on standardized third and fourth moments respectively.[3] Perhaps even more revealing than Table 1 are the frequency distributions depicted in Figures 1–8. These figures certainly confirm Micceri's (1989) contentions alluded to above and bring to mind Blair's (1981) statement regarding the construction of histograms, "This time-honored but often neglected practice usually paints pictures of distributions that are unimagined by researchers who think of data in terms of the normal curve" (p. 504). These histograms illustrate that data sets from psychology and education can be more radically nonnormal than data sets from other disciplines of science (e.g., Pearson & Please, 1975, p. 225).

## Results

### *Robustness to Type I Error*

Under the Gaussian distribution, the minimum and maximum upper and lower tail rejections for the .10, .05, and .01 alpha levels were .048–.052; .023–.026; and .004–.005, respectively. This demonstrates the adequacy of the algorithms used in the simulation.

Results of the robustness to Type I error portion of the study are shown in Tables 2–9. Columns headed U050 and L050, for example, show the proportion of rejections falling in the upper

[2] The eight distributions have been compiled into a Microsoft FORTRAN 5.0 library ("lib" file) that permits sampling with replacement (Sawilowsky, Blair, & Micceri, 1990). For a free copy of the compiled program and the source code, send a DOS formatted 5¼″ or 3½″ disk with a self-addressed stamped disk mailer to Shlomo S. Sawilowsky.

[3] The computational formulas are found in the *SPSS: Statistical package for the social sciences* (SPSS, 1975), p. 185, except that kurtosis has been scaled here so that normal kurtosis is 3.0.

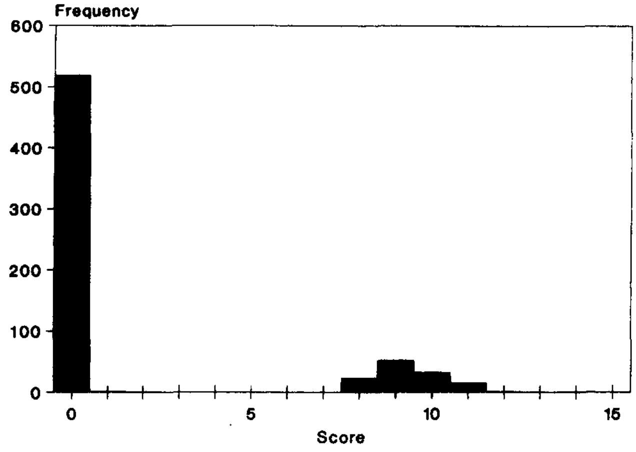

*Figure 1.* Discrete mass at zero with gap, psychometric measure.

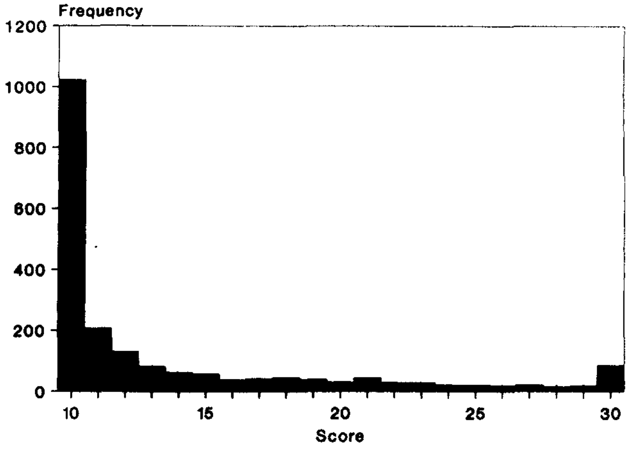

*Figure 3.* Extreme asymmetry, psychometric measure.

and lower 5% critical regions respectively. The Total columns simply sum the upper and lower columns, thereby providing results for two-tailed tests. Robustness results for each distribution are discussed below.

*Discrete mass at zero with gap (psychometric measure).* Table 2 shows that in the case $n_1 \neq n_2$, one-sided tests produced liberal or conservative results, with rejection rates near alpha occurring in a few instances. Two-sided tests were, with two exceptions, conservative in these instances with the degree of conservativeness being related to sample sizes and significance levels. With the exception of sample size (5, 15), two-sided tests at the .010 level produced results near alpha. For the larger alpha levels, two-tailed results were dramatically conservative for the smaller sample sizes but were somewhat ameliorated when samples of sizes (20, 60) were reached. Type I error rates near alpha were attained for samples of sizes (30, 90).

In the simulations in which $n_1 = n_2$, one-tailed independent samples tests were quite conservative when $n_1 = n_2 = 10$, but results were markedly improved when each sample consisted of 20 observations. When $n_1 = n_2 = 30$, the results were modestly conservative. The two larger sample sizes generated rates very near alpha. Patterns for the two-sided tests were similar to those for the one-sided tests when sample sizes were equal.

*Mass with gap (achievement measure).* As Table 3 indicates, when sampling was from this distribution, the *t* test tended to produce Type I error rates near nominal values for all conditions studied.

*Extreme asymmetry (psychometric measure).* As noted in Table 4, for unequal sample sizes, one-sided tests were generally conservative or liberal depending on the tail in which the test was conducted. Results were generally improved as sample sizes and alpha levels increased. Two-tailed results were usually near nominal levels except when $\alpha = .100$ or .050 and samples were of size (5, 15). In these cases, moderately conservative results were obtained. Results were improved when sample sizes were equal, with the test generating rates near alpha for all conditions studied. Some conservative exceptions occurred when small sample sizes were combined with small alpha levels.

*Extreme asymmetry (achievement measure).* Results depicted in Table 5 showed patterns that were somewhat similar to those found in Table 4. In this situation, however, Type I error rates were usually much closer to nominal levels than

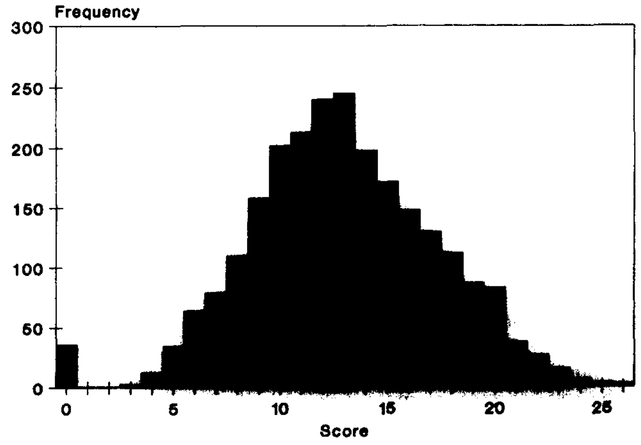

*Figure 2.* Mass at zero, achievement measure.

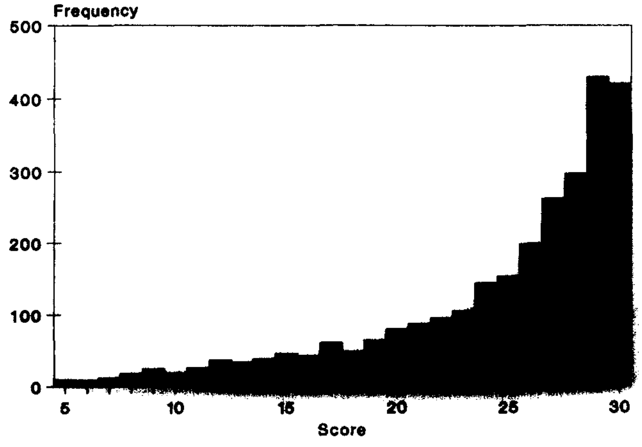

*Figure 4.* Extreme asymmetry, achievement measure.

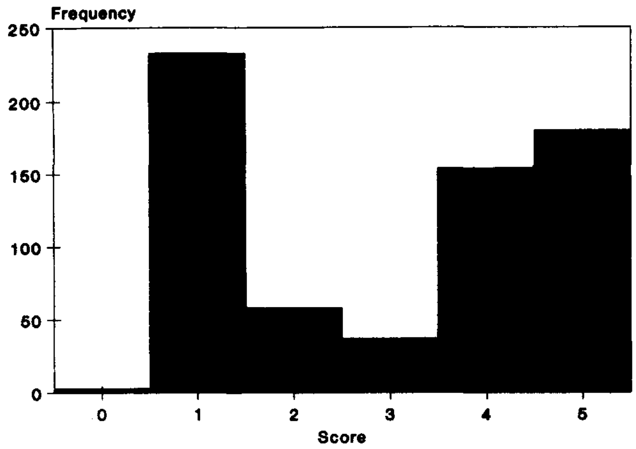

*Figure 5.* Extreme bimodality, psychometric measure.

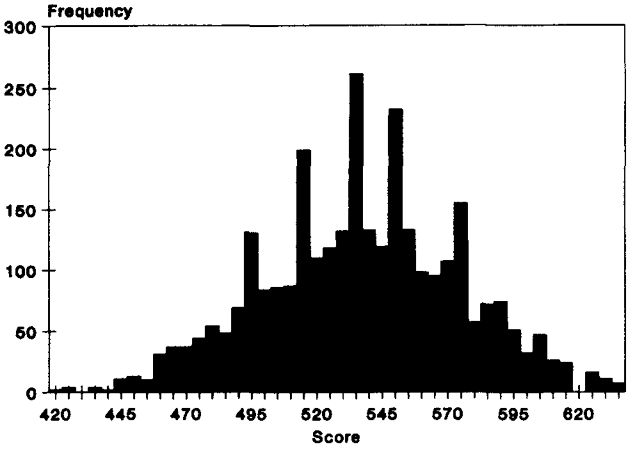

*Figure 7.* Digit preference, achievement measure.

were the rates shown in Table 4. The larger discrepancies between nominal and actual alpha were of a conservative nature.

*Remaining distributions.* Type I error results obtained from the remaining distributions are depicted in Tables 6–9. Results for the extreme bimodality (psychometric measure), multimodality and lumpiness (achievement measure), digit preference (achievement measure), and smooth symmetric (achievement measure) distributions, with few exceptions, were quite consistent with those expected under normal theory.

### *Robustness to Type II Error*

The Type II error properties of the independent-samples $t$ test, with $\alpha = .05$, for hypotheses-of-shift-in-location parameters for sample size (5, 15) are depicted in Figure 9 for the discrete mass at zero with gap, multimodal and lumpy, and digit preference distributions. For comparison, we provide the power curve for the Gaussian distribution. (The remaining real distributions produced power similar to the Gaussian distribution and are not presented here.) The power is not adjusted to the operational level (.05) to take into account the conservativeness of the test under some of the distributions but, rather, is accumulated against the actual level. The figure indicates the rejection rate when a shift in means of $0.2\sigma$, $0.5\sigma$, $0.8\sigma$, and $1.2\sigma$ was added to $X_{2j}$ and $c = 1$.

As projected by Cohen (1988), the independent-samples $t$ test on data sampled from real distributions produced power rates very similar to levels expected from normal curve theory. Skewed distributions produced slightly more power (about .03 to .05) than was obtained under the Gaussian distribution, with the remaining distributions yielding an average of about .03 power less than would an independent-samples $t$ test on Gaussian data with the same sample size. The power of the $t$ test under nonskewed real distributions converged with normal curve results for sample sizes of $(n_1 + n_2) > 60$. These power results indicate that for shift alternatives, the rejection rate of the independent-samples $t$ test was maintained at a fairly consistent level, regardless of the population shape, sample size, and effect size.

There was also a slight loss of power under hypotheses of shift in location plus change in scale. Table 10 contains power rates when a departure from normality is combined with a vio-

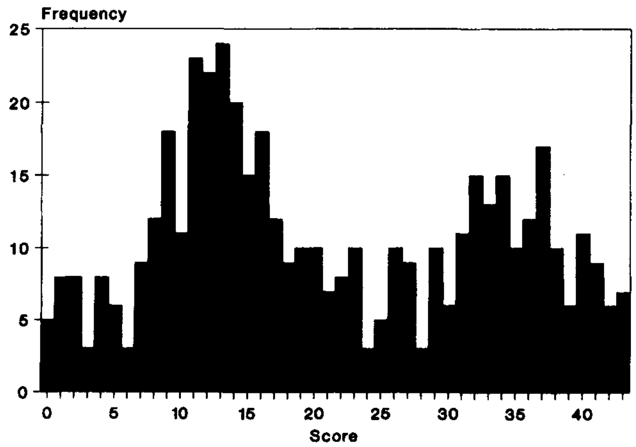

*Figure 6.* Multimodality and lumpiness, achievement measure.

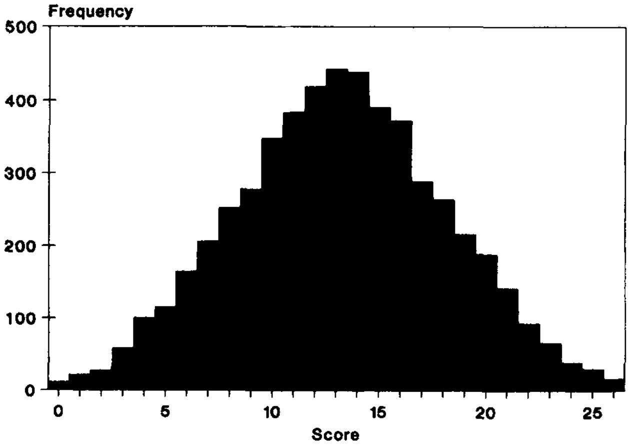

*Figure 8.* Smooth symmetric, achievement measure.

Table 2

*Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Discrete Mass at Zero With Gap (Psychometric) Distribution, 10,000 Repetitions*

| | $\alpha$ = .100 | | | $\alpha$ = .050 | | | $\alpha$ = .010 | | |
|---|---|---|---|---|---|---|---|---|---|
| Sample size | U050 | L050 | Total | U025 | L025 | Total | U005 | L005 | Total |
| 5, 15 | .008 | .000 | .008 | .003 | .000 | .003 | .001 | .000 | .001 |
| 10, 10 | .017 | .014 | .031 | .007 | .006 | .013 | .000 | .001 | .001 |
| 10, 30 | .063 | .005 | .068 | .036 | .000 | .036 | .011 | .000 | .011 |
| 20, 20 | .042 | .041 | .082 | .020 | .020 | .040 | .002 | .002 | .004 |
| 15, 45 | .056 | .019 | .075 | .030 | .004 | .034 | .010 | .000 | .010 |
| 30, 30 | .046 | .047 | .093 | .021 | .024 | .045 | .004 | .003 | .007 |
| 20, 60 | .057 | .029 | .086 | .030 | .009 | .039 | .009 | .000 | .009 |
| 40, 40 | .053 | .051 | .104 | .027 | .025 | .052 | .006 | .005 | .011 |
| 30, 90 | .054 | .041 | .096 | .031 | .018 | .049 | .007 | .002 | .009 |
| 60, 60 | .050 | .050 | .100 | .025 | .024 | .049 | .005 | .006 | .011 |

*Note.* Sample size = $n_1$, $n_2$. U050, U025, U005 = proportion of rejections in the upper tail. L050, L025, L005 = proportion of rejections in the lower tail.

Table 3

*Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Mass at Zero (Achievement) Distribution, 10,000 Repetitions*

| | $\alpha$ = .100 | | | $\alpha$ = .050 | | | $\alpha$ = .010 | | |
|---|---|---|---|---|---|---|---|---|---|
| Sample size | U050 | L050 | Total | U025 | L025 | Total | U005 | L005 | Total |
| 5, 15 | .049 | .047 | .097 | .026 | .024 | .050 | .005 | .003 | .008 |
| 10, 10 | .047 | .048 | .095 | .024 | .025 | .049 | .005 | .005 | .010 |
| 10, 30 | .047 | .054 | .101 | .022 | .027 | .049 | .005 | .006 | .011 |
| 20, 20 | .048 | .050 | .098 | .023 | .025 | .048 | .004 | .004 | .008 |
| 15, 45 | .049 | .049 | .098 | .022 | .024 | .046 | .006 | .004 | .010 |
| 30, 30 | .048 | .054 | .102 | .023 | .027 | .050 | .004 | .005 | .009 |
| 20, 60 | .051 | .052 | .103 | .024 | .026 | .050 | .004 | .005 | .009 |
| 40, 40 | .049 | .053 | .102 | .025 | .027 | .052 | .005 | .006 | .011 |
| 30, 90 | .052 | .052 | .104 | .026 | .026 | .052 | .005 | .004 | .009 |
| 60, 60 | .051 | .049 | .100 | .024 | .026 | .050 | .005 | .005 | .010 |

*Note.* Sample size = $n_1$, $n_2$. U050, U025, U005 = proportion of rejections in the upper tail. L050, L025, L005 = proportion of rejections in the lower tail.

Table 4

*Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From an Extreme Asymmetric (Psychometric) Distribution, 10,000 Repetitions*

| | $\alpha$ = .100 | | | $\alpha$ = .050 | | | $\alpha$ = .010 | | |
|---|---|---|---|---|---|---|---|---|---|
| Sample size | U050 | L050 | Total | U025 | L025 | Total | U005 | L005 | Total |
| 5, 15 | .065 | .011 | .076 | .039 | .002 | .041 | .012 | .000 | .012 |
| 10, 10 | .052 | .048 | .100 | .022 | .023 | .045 | .002 | .003 | .005 |
| 10, 30 | .062 | .034 | .096 | .037 | .011 | .048 | .011 | .000 | .011 |
| 20, 20 | .050 | .051 | .101 | .023 | .024 | .047 | .004 | .005 | .009 |
| 15, 45 | .057 | .040 | .097 | .031 | .015 | .046 | .009 | .001 | .010 |
| 30, 30 | .050 | .051 | .101 | .025 | .025 | .050 | .006 | .005 | .011 |
| 20, 60 | .058 | .040 | .098 | .033 | .015 | .048 | .009 | .001 | .010 |
| 40, 40 | .051 | .053 | .104 | .023 | .025 | .048 | .005 | .005 | .010 |
| 30, 90 | .053 | .046 | .099 | .027 | .019 | .046 | .006 | .002 | .008 |
| 60, 60 | .048 | .049 | .097 | .026 | .024 | .050 | .006 | .004 | .010 |

*Note.* Sample size = $n_1$, $n_2$. U050, U025, U005 = proportion of rejections in the upper tail. L050, L025, L005 = proportion of rejections in the lower tail.

Table 5

*Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From an Extreme Asymmetric (Psychometric) Distribution, 10,000 Repetitions*

| Sample | $\alpha$ = .100 | | | $\alpha$ = .050 | | | $\alpha$ = .010 | | |
|---|---|---|---|---|---|---|---|---|---|
| size | U050 | L050 | Total | U025 | L025 | Total | U005 | L005 | Total |
| 5, 15 | .030 | .059 | .089 | .011 | .033 | .044 | .000 | .010 | .010 |
| 10, 10 | .050 | .051 | .101 | .025 | .025 | .050 | .005 | .003 | .008 |
| 10, 30 | .041 | .058 | .099 | .015 | .032 | .047 | .001 | .007 | .008 |
| 20, 20 | .047 | .048 | .095 | .024 | .023 | .047 | .005 | .005 | .010 |
| 15, 45 | .048 | .055 | .103 | .021 | .030 | .051 | .001 | .008 | .009 |
| 30, 30 | .050 | .052 | .102 | .024 | .025 | .049 | .005 | .005 | .010 |
| 20, 60 | .047 | .054 | .101 | .020 | .029 | .049 | .003 | .008 | .011 |
| 40, 40 | .050 | .052 | .102 | .026 | .027 | .053 | .006 | .005 | .011 |
| 30, 90 | .045 | .051 | .096 | .019 | .026 | .045 | .004 | .006 | .010 |
| 60, 60 | .051 | .050 | .101 | .025 | .026 | .051 | .004 | .005 | .009 |

*Note.* Sample size = $n_1$, $n_2$. U050, U025, U005 = proportion of rejections in the upper tail. L050, L025, L005 = proportion of rejections in the lower tail.

Table 6

*Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From an Extreme Bimodal (Psychometric) Distribution, 10,000 Repetitions*

| Sample | $\alpha$ = .100 | | | $\alpha$ = .050 | | | $\alpha$ = .010 | | |
|---|---|---|---|---|---|---|---|---|---|
| size | U050 | L050 | Total | U025 | L025 | Total | U005 | L005 | Total |
| 5, 15 | .048 | .047 | .095 | .025 | .023 | .048 | .004 | .005 | .009 |
| 10, 10 | .053 | .050 | .103 | .028 | .027 | .055 | .008 | .008 | .016 |
| 10, 30 | .048 | .049 | .097 | .025 | .025 | .050 | .005 | .005 | .010 |
| 20, 20 | .049 | .052 | .101 | .025 | .027 | .052 | .006 | .006 | .012 |
| 15, 45 | .048 | .050 | .098 | .026 | .024 | .050 | .004 | .005 | .009 |
| 30, 30 | .050 | .048 | .098 | .023 | .025 | .048 | .005 | .005 | .010 |
| 20, 60 | .050 | .049 | .099 | .024 | .025 | .049 | .006 | .006 | .012 |
| 40, 40 | .051 | .047 | .098 | .027 | .024 | .051 | .005 | .005 | .010 |
| 30, 90 | .051 | .049 | .100 | .025 | .024 | .049 | .005 | .005 | .010 |
| 60, 60 | .052 | .049 | .101 | .025 | .026 | .051 | .005 | .005 | .010 |

*Note.* Sample size = $n_1$, $n_2$. U050, U025, U005 = proportion of rejections in the upper tail. L050, L025, L005 = proportion of rejections in the lower tail.

Table 7

*Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Multimodal and Lumpy (Achievement) Distribution, 10,000 Repetitions*

| Sample | $\alpha$ = .100 | | | $\alpha$ = .050 | | | $\alpha$ = .010 | | |
|---|---|---|---|---|---|---|---|---|---|
| size | U050 | L050 | Total | U025 | L025 | Total | U005 | L005 | Total |
| 5, 15 | .051 | .048 | .099 | .025 | .023 | .048 | .005 | .004 | .009 |
| 10, 10 | .048 | .050 | .098 | .026 | .026 | .052 | .006 | .005 | .011 |
| 10, 30 | .047 | .046 | .093 | .025 | .023 | .048 | .005 | .005 | .010 |
| 20, 20 | .051 | .050 | .101 | .025 | .027 | .052 | .006 | .006 | .012 |
| 15, 45 | .054 | .049 | .103 | .029 | .025 | .054 | .007 | .005 | .012 |
| 30, 30 | .053 | .051 | .104 | .025 | .026 | .051 | .005 | .005 | .010 |
| 20, 60 | .051 | .050 | .101 | .025 | .025 | .050 | .006 | .005 | .011 |
| 40, 40 | .048 | .050 | .098 | .025 | .024 | .049 | .006 | .006 | .012 |
| 30, 90 | .049 | .051 | .100 | .025 | .026 | .051 | .006 | .006 | .012 |
| 60, 60 | .054 | .054 | .108 | .029 | .026 | .055 | .005 | .006 | .010 |

*Note.* Sample size = $n_1$, $n_2$. U050, U025, U005 = proportion of rejections in the upper tail. L050, L025, L005 = proportion of rejections in the lower tail.

Table 8
*Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Digit Preference (Achievement) Distribution, 10,000 Repetitions*

| Sample | $\alpha$ = .100 | | | $\alpha$ = .050 | | | $\alpha$ = .010 | | |
|---|---|---|---|---|---|---|---|---|---|
| size | U050 | L050 | Total | U025 | L025 | Total | U005 | L005 | Total |
| 5, 15 | .047 | .053 | .100 | .023 | .028 | .051 | .004 | .007 | .011 |
| 10, 10 | .051 | .048 | .099 | .025 | .025 | .050 | .004 | .004 | .008 |
| 10, 30 | .051 | .049 | .100 | .026 | .024 | .050 | .006 | .006 | .012 |
| 20, 20 | .048 | .050 | .098 | .025 | .025 | .050 | .006 | .006 | .012 |
| 15, 45 | .044 | .047 | .091 | .021 | .024 | .045 | .004 | .005 | .009 |
| 30, 30 | .052 | .051 | .103 | .027 | .026 | .053 | .006 | .005 | .011 |
| 20, 60 | .047 | .051 | .098 | .024 | .026 | .050 | .005 | .005 | .010 |
| 40, 40 | .049 | .049 | .098 | .025 | .025 | .050 | .006 | .006 | .011 |
| 30, 90 | .049 | .050 | .099 | .025 | .025 | .050 | .004 | .005 | .009 |
| 60, 60 | .049 | .053 | .102 | .024 | .029 | .053 | .003 | .005 | .008 |

*Note.* Sample size = $n_1$, $n_2$. U050, U025, U005 = proportion of rejections in the upper tail. L050, L025, L005 = proportion of rejections in the lower tail.

lation of homogeneous variances. The *t* test was applied to a Gaussian distribution with mean of zero and standard deviation of one, at the .05 alpha level (results for the .01 alpha level are not presented here). With the level of heterogeneous variances at 4:1 (i.e., $c = 2$; results for $c = \sqrt{2}$ and $\sqrt{3}$ are not presented here), the power rate of about .49 was obtained for each of the balanced layouts studied. Under these conditions, there was a slight power loss (an average of about .04) for several of the real distributions. The extreme asymmetry (achievement measure) distribution, with a large negative skew, generated slightly more power than the Gaussian distribution.

## Discussion

The present study does not address the issue of heterogeneous variances insofar as their impact on Type I errors is concerned (the "Behrens-Fisher problem"). In keeping with Micceri's (1989) premise, this study was designed to produce results that are as realistic as possible. We have spent many years examining large data sets but have never encountered a treatment or other naturally occurring condition that produces heterogeneous variances while leaving population means exactly equal. While the impact of some treatments may be seen primarily in measures of scale, they always (in our experience) impact location as well. Thus, we see this issue as being more important when viewed from a theoretical rather than an applications point of view.

An issue of interest concerns the use to which real data distributions are put. Figure 1 depicts a distribution of scores from a psychometric measure having a discrete probability mass at zero with about 10% of the observations scoring above zero, some substantially so. Treating such as a single distribution is not always appropriate, however, because researchers using these measures often are interested only in those subjects scoring above zero. For example, in recent substance use literature two paradigms are emerging: user/abuser and nonuser/user/abuser. For studies using the former case, such as Hillman and Sawilowsky (1991a), the percentage of nonusers is reported, but hypotheses (such as comparing means) are tested only on the user and abuser categories. However, for the latter case, such as

Table 9
*Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Smooth Symmetric (Achievement) Distribution, 10,000 Repetitions*

| Sample | $\alpha$ = .100 | | | $\alpha$ = .050 | | | $\alpha$ = .010 | | |
|---|---|---|---|---|---|---|---|---|---|
| size | U050 | L050 | Total | U025 | L025 | Total | U005 | L005 | Total |
| 5, 15 | .049 | .050 | .099 | .025 | .024 | .049 | .004 | .005 | .009 |
| 10, 10 | .049 | .052 | .101 | .026 | .027 | .053 | .006 | .005 | .011 |
| 10, 30 | .049 | .049 | .098 | .026 | .023 | .049 | .006 | .005 | .011 |
| 20, 20 | .051 | .048 | .099 | .025 | .025 | .050 | .005 | .005 | .010 |
| 15, 45 | .050 | .050 | .100 | .025 | .025 | .050 | .005 | .005 | .010 |
| 30, 30 | .052 | .050 | .102 | .024 | .024 | .048 | .005 | .004 | .009 |
| 20, 60 | .049 | .055 | .104 | .026 | .029 | .055 | .004 | .006 | .010 |
| 40, 40 | .049 | .046 | .095 | .023 | .022 | .045 | .005 | .005 | .010 |
| 30, 90 | .049 | .049 | .098 | .023 | .023 | .046 | .004 | .005 | .009 |
| 60, 60 | .049 | .048 | .097 | .026 | .023 | .049 | .005 | .004 | .009 |

*Note.* Sample size = $n_1$, $n_2$. U050, U025, U005 = proportion of rejections in the upper tail. L050, L025, L005 = proportion of rejections in the lower tail.

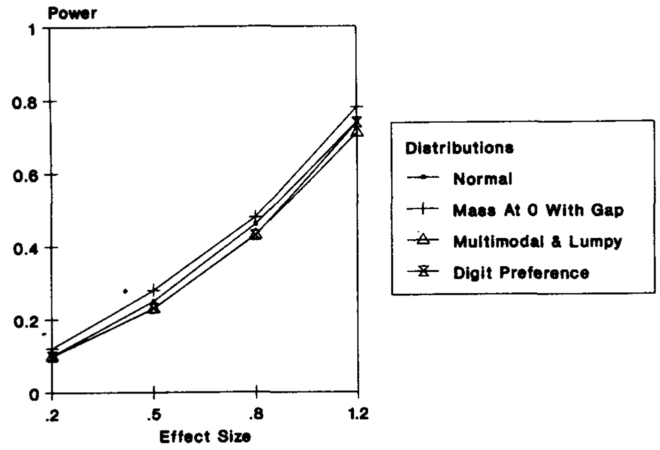

*Figure 9.* One-tailed *t*-test power curves for $(n_1, n_2) = (5, 15)$; $\alpha = .05$; 10,000 repetitions.

Hillman and Sawilowsky (1991b) and Shedler and Block (1990), who studied abstainers, experimenters, and frequent users, it is appropriate to compare means for all three groups, even though the nonuser scores are discrete zeros.

We must agree with Micceri (1989) that the distributions studied here provided a more realistic and stringent test of the *t* test's sensitivity to population shape than has been afforded by previous studies of this topic. These real distributions highlight situations in which the *t* test was, by any definition, nonrobust to Type I error. The degree of nonrobustness seen in these instances was at times more severe than has been previously reported. Having said this, however, we must note that the results obtained from these distributions do not change, in any fundamental fashion, the conclusions reached on the basis of studies that focused on populations modeled by well-known mathematical functions. That is to say, this study showed the *t* test to be reasonably robust under the conditions outlined in the introduction to this article: when sample sizes are equal or nearly so, sample sizes are fairly large (25 to 30), and tests are two-tailed rather than one-tailed. This study also showed that departures from nominal values were almost always of a conservative rather than a liberal nature for two-tailed tests. Also consistent with the prevailing literature is the fact that a dominant factor bringing about nonrobustness to Type I error was extreme skew. Furthermore, kurtosis appears to influence Type I error when combined with skew in some of these real data sets.

Despite the conservative nature with regard to Type I error of the *t* test for some of these real distributions, there was little effect on the power levels for the variety of treatment conditions and sample sizes studied. Researchers may easily compensate for the slight loss in power by selecting a slightly larger sample size than recommended by Cohen (1988). However, the robustness of the independent-samples *t* test to Type II error does not preclude the need for alternative tests, as Scheffé (1959) pointed out, "The question of whether *F* tests [and in this case *t* tests] preserve against nonnormal alternatives the power calculated under normal theory should not be confused with that of their efficiency against such alternatives relative to other kind of tests" (p. 351). Theoretical studies by Dixon (1954) and Hodges and Lehmann (1956) and empirical studies by Chernoff and Savage (1958) and Neave and Granger (1968) indicated that nonparametric alternatives to the *t* test are robust and often more powerful than the *t* under population nonnormality (see, e.g., Blair & Higgins, 1980a, 1980b, 1981, 1985; Hemelrijk, 1961; Randles & Wolfe, 1979).

As an example, an important competitor to the independent-samples *t* test, for alternatives of shift in location parameter, is the Wilcoxon rank-sum test, which is also known as the Mann-Whitney *U* test. A Monte Carlo comparison (10,000 repetitions) of the power for the *t* test and Wilcoxon test with a sample size of (5, 15) drawn from the extreme asymmetric (psychometric measure) distribution, depicted in Figure 3, indicates that at the .05 alpha level and *ES* of .2, the power of the Wilcoxon test is .395 as compared with only .139 for the *t* test. When the *ES* is increased to .5, the Wilcoxon test rejects at the rate of .723, but the *t* test only rejects at the rate of .495. This issue also applies to competitors to the *t* statistic for alternative hypotheses of change in means plus increase in variances.

In summary, researchers in psychology and education should note that real distributions in these areas of inquiry are sufficiently nonnormal to bring about nonrobust Type I results under certain circumstances. However, the two primary ele-

Table 10
*Comparative Power Rates for Independent-Samples t Test at the .05 Alpha Level When Effect Size Produces Power of Approximately .50 for the Gaussian Distribution*

| | | Sample size | | | | |
|---|---|---|---|---|---|---|
| Distribution | Type of measure | (10, 10) | (20, 20) | (30, 30) | (40, 40) | (60, 60) |
| Discrete mass at zero | Psychometric | .44 | .42 | .44 | .43 | .43 |
| Discrete mass with gap | Achievement | .48 | .48 | .47 | .48 | .48 |
| Extreme asymmetry | Psychometric | .44 | .46 | .44 | .46 | .45 |
| Extreme asymmetry | Achievement | .54 | .53 | .52 | .53 | .54 |
| Extreme bimodality | Achievement | .44 | .46 | .45 | .46 | .45 |
| Multimodality and lumpy | Achievement | .44 | .44 | .43 | .43 | .44 |
| Digit Preference | Achievement | .47 | .47 | .47 | .47 | .47 |
| Smooth Symmetric | Achievement | .47 | .47 | .47 | .47 | .46 |
| Gaussian | NA | .49 | .48 | .49 | .49 | .49 |

*Note.* Heteroscedasticity of 4:1 ($c = 2$); 10,000 repetitions. NA = not applicable.

ments identified by Micceri (1989) as occurring in real data but not considered in previous robustness studies, namely (a) multimodality and lumpiness and (b) digit preference, appeared to have little impact in terms of the Type I or Type II error properties of the parametric statistic studied.

## References

Blair, R. C. (1981). A reaction to "consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance.' *Review of Educational Research, 51,* 499–507.

Blair, R. C. (1987). *RANGEN.* Boca Raton, FL: IBM Corporation.

Blair, R. C., & Higgins, J. J. (1980a). A comparison of the power of the *t* test and the Wilcoxon statistics when samples are drawn from a certain mixed normal distribution. *Evaluation Review, 4,* 645–656.

Blair, R. C., & Higgins, J. J. (1980b). A comparison of the power of the Wilcoxon's rank-sum statistic to that of student's *t* statistic under various non-normal distributions. *Journal of Educational Statistics, 5,* 309–335.

Blair, R. C., & Higgins, J. J. (1981). A note on the asymptotic relative efficiency of the Wilcoxon rank-sum test relative to the independent means *t* test under mixtures of two normal populations. *British Journal of Mathematical and Statistical Psychology, 31,* 124–128.

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples *t* test to that of Wilcoxon's signed-ranks tests under various population shapes. *Psychological Bulletin, 97,* 119–128.

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

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

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shape. *American Statistician, 31,* 147–150.

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

Bradley, J. V. (1982). The insidious L-shaped distribution. *Bulletin of the Psychometrics Society, 20*(2), 85–88.

Chernoff, H., & Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematical Statistics, 29,* 972–999.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dixon, W. J. (1954). Power under normality of several nonparametric tests. *Annals of Mathematical Statistics, 25,* 610–614.

Efron, B. (1969). Student's *t*-test under symmetry conditions. *Journal of the American Statistical Association, 64,* 1278–1302.

Gayen, A. K. (1949). The distribution of 'Student' *t* in random samples of any size drawn from non-normal universes. *Biometrika, 36,* 353–369.

Gayen, A. K. (1950). The distribution of the variance ratio in random samples of any size drawn from non-normal universes. *Biometrika, 37,* 236–255.

Geary, R. C. (1936). The distribution of 'Student's' ratio from non-normal samples. *Journal of the Royal Statistical Society, 3,* 178–184.

Geary, R. C. (1947). Testing for normality. *Biometrika, 34,* 209–242.

Glass, G., Peckham, P., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42,* 237–288.

Hemelrijk, J. (1961). Experimental comparison of student's and Wilcoxon's two sample test. In H. de Jonge (Ed.), *Qualitative methods in psychology.* New York: Interscience.

Hillman, S. B., and Sawilowsky, S. S. (1991a). Multidimensional differences between adolescent substance abusers and users. *Psychological Reports, 68,* 115–122.

Hillman, S. B., and Sawilowsky, S. S. (1991b, August). *Profiles of adolescent substance abstainers, users, and abusers.* Paper presented at the 99th Annual Convention of the American Psychological Association, Division 12, Clinical Psychology, San Francisco.

Hodges, J. C., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the *t* test. *Annals of Mathematical Statistics, 27,* 324–335.

International Mathematical and Statistical Libraries. (1987). *IMSL library reference manual* (10th ed.). Houston, TX: Author.

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In P. R. Krishnaiah (Ed.), *Handbook of statistics,* (Vol. 1, pp. 199–236). Amsterdam: North-Holland.

Micceri, T. (1986, November). A futile search for that statistical chimera of normality. Paper presented at the 31st Annual Convention of the Florida Educational Research Association, Tampa.

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

Neave, H. R., & Granger, C. W. J. (1968). A Monte Carlo study comparing various two-sample tests for differences in mean. *Technometrics, 10,* 509–522.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 63,* 223–241.

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

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60,* 91–126.

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). A PC FORTRAN subroutine library of psychology and education data sets. *Psychometrika, 55,* 729.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Shedler, J., & Block, J. (1990). Adolescent drug use and psychological health. *American Psychologist, 45,* 612–630.

SPSS. (1975). *SPSS: Statistical package for the social sciences* (2nd ed.). New York: McGraw-Hill.

Received December 28, 1989
Revision received October 30, 1990
Accepted June 13, 1991 ■