DOCUMENT RESUME

ED 443 850 TM 031 495

AUTHOR Delaney, Harold D.; Vargha, Andras
TITLE The Effect of Nonnormality on Student's Two-Sample T Test.
PUB DATE 2000-04-00
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (81st, New Orleans, LA, April 24-28, 2000). Sponsored by Hungarian OTKA, Hungarian FKFP, and the Open Society Support Foundation.
CONTRACT T018353; 0194/2000; T032157; 584/1998
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Monte Carlo Methods; Sampling; *Statistical Distributions
IDENTIFIERS *T Test; Violation of Assumptions

ABSTRACT

While violation of the homogeneity of variance assumption has received considerable attention, violation of the assumption of normally distributed data has not received as much attention. As a result, researchers may have the mistaken impression that as long as the assumptions of independence of observations and homogeneity of variance are satisfied, violations of the distributional assumption gave inconsequential effects. This paper reviews some of the relevant literature and reports the results of a new Monte Carlo study indicating that this is not the case. The simulation investigated the effects of varying skewness and kurtosis levels, while maintaining equal population variances, on the two-sample "t" test and Welch's robust "t" test. Sample sizes were either small or moderate, and equal or unequal. Results indicate that, with skewed distributions, the validity of both the "t" test and the Welch test clearly depends on the two distributions being skewed in the same direction. When the two parent distributions are skewed in the same direction, both tests have quite acceptable Type I error rates, even with relatively small samples. However, when the two parent distributions are skewed in opposite directions, then the true Type I error rates can deviate markedly from the nominal level even though population variances are equal. The actual Type I error rate of the "t" test performed at a 0.05 nominal level with homogeneous variances can be higher than 0.08 with a two-tailed test, and can be higher than 0.11 with a one-tailed test. (Contains 16 figures and 23 references.) (SLD)

THE EFFECT OF NONNORMALITY ON STUDENT'S TWO-SAMPLE T TEST

Harold D. Delaney

Department of Psychology, UNM, USA

András Vargha

Institute of Psychology, ELTE, Hungary

Please address all correspondence concerning this manuscript to:
Harold Delaney
Department of Psychology
University of New Mexico
Albuquerque, NM 87131
E-mail: hdelaney@unm.edu

Author Note

Much of the work reported herein resulted from the collaborative efforts of Drs. Vargha and Delaney while Vargha was a Széchenyi Professor Scholar and supported by Hungarian OTKA, grant Nos.: T018353 and T032157. Vargha's work was also supported by the Hungarian FKFP grant No. 0194/2000 and the Research Support Scheme of the Open Society Support Foundation, grant No.: 584/1998.

Abstract

A decade ago, Micceri (1989) presented a convincing case that data sets in education and psychology are very often not normally distributed. In a follow-up Monte Carlo study, Sawilowsky and Blair (1992) demonstrated that the familiar two-group *t* test is nonetheless reasonably robust when used in situations where both groups were sampled from one of the non-normal forms that had been documented by Micceri as relatively common with real data.

These results serve to confirm the typical textbook advice of the robustness of such parametric procedures as the *t* test to violations of the normality assumption. While violation of the homogeneity of variance assumption has received considerable attention, violation of the assumption of normally distributed data has not received as much attention. As a result, researchers may have the mistaken impression that as long as the assumptions of independence of observations and homogeneity of variance are satisfied, violations of the distributional assumption have inconsequential effects. The current paper will review some of the relevant literature and report the results of a new Monte Carlo study indicating that this is not the case.

The simulation investigated the effects of varying skewness and kurtosis levels, while maintaining equal population variances, on the two-sample *t* test and Welch's robust *t* test. Sample sizes were either small or moderate, and equal or unequal. Results indicated that, with skewed distributions, the validity of both the *t* test and the Welch test clearly depends on the two distributions being skewed in the same direction. When the two parent distributions are skewed in the same direction, both tests have quite acceptable Type I error rates, even with relatively small samples (say, m = n = 9). However, when the two parent distributions are skewed in opposite directions, then the true Type I error rates can deviate markedly from the nominal level even though population variances are equal. Specifically, the actual Type I error rate of the *t* test performed at a .05 nominal level with homogeneous variances can be higher than 0.08 with a two-tailed test, and can be higher than 0.11 with a one-tailed test.

Key Words: group comparison, *t* test, skewness, kurtosis, Welch's test.

1. INTRODUCTION

Statistics texts for the behavioral sciences tend to emphasize the robustness of parametric procedures, such as Student's two-group *t* test, to violations of the distributional assumption of normality. While consequences of violation of the homogeneity of variance assumption are widely recognized and many texts introduce Welch-type tests for use in comparing means when variances are heterogeneous, violation of the assumption of normally distributed data has not received as much attention. As a result, researchers may have the mistaken impression that as long as the assumptions of independence of observations and homogeneity of variance are satisfied, violations of the distributional assumption have inconsequential effects. The current paper will review some of the relevant literature and report the results of a new Monte Carlo study indicating that this is not the case. Practical guidance about the effects of skewness or non-normal kurtosis will be offered.

Suppose that one has a quantitative variable X (e.g., WAIS IQ, Reaction time in a perception study, or the Tolerance scale of the California Personality Inventory), and wants to decide whether the mean of X is the same in two different populations (e.g., in males and females; in neurotics and psychopaths; or in low and high education groups). If one has two independent samples taken from the two populations, the most common way to address this question is to use the familiar Student's two-sample *t* test (see, e.g., Wilcox, 1996, p. 126). To perform this test one calculates the *t* statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_b\sqrt{\frac{1}{m} + \frac{1}{n}}} \qquad (1)$$

In this formula $\bar{x}_1$ and $\bar{x}_2$ are the averages of the two independent samples, *m* and *n* are the corresponding sample sizes, and s_b is the square-root of the pooled within sample variance, which can be determined by computing the weighted average of the two sample variances, $(s_1)^2$ and $(s_2)^2$, by the following formula:

$$s_b^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} \quad (2)$$

(see.Wilcox, 1996, p. 128). If we test the null hypothesis

$$H_0: \mu_1 = \mu_2 \quad (3)$$

against the alternative hypothesis

$$H_1: \mu_1 \neq \mu_2 \quad (4)$$

at significance level α, H_0 will be retained if t falls into the

$$T = (-t_{\alpha/2}, t_{\alpha/2})$$

region of acceptance, where $t_{\alpha/2}$ is the $1\text{-}t_{\alpha/2}$ percentile value of the t distribution with df = m+n-2 degrees of freedom. If t falls on or outside of the border of this region, H_0 will be rejected.

In the one-tailed case if we test H_0 against H_1: $\mu_x < \mu_0$ or H_2: $\mu_x > \mu_0$, the corresponding regions of acceptance are $T_1 = (-t_{0.05}, \infty)$ and $T_2 = (-\infty, t_{0.05})$.

Besides the independence of the two data samples, the validity of the t test rests on two other assumptions as well: the distribution of X must be normal and the variances identical in the two populations. It is well known that the variance homogeneity assumption of the two-sample t-test must be taken seriously. Admittedly, the general finding is that the distorting effect of variance heterogeneity can be somewhat diminished by using equal sample sizes. However, if the sample sizes vary and the larger sample size is paired with the larger variance, the t test will be unduly conservative (yielding an unwanted drop of power), and if the larger sample size is paired with the smaller variance, the t test becomes too liberal (yielding a marked increase in the Type I error rate). As an example, if m = 15, n = 5, and $\sigma_1 = 2\sigma_2$, then the Type I error of the t test at $\alpha = .05$ level drops to .038, and if m = 15, n = 5, and $\sigma_1 = .5\sigma_2$, then the Type I error of the t test at $\alpha = .05$ level increases to .072 (see Scheffé, 1959, p. 353, Table 10.4.1).

If the variance homogeneity condition of the *t* test is not satisfied, the best alternative for testing the null hypothesis in (3) seems to be the Welch test. The test statistic of this test is

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{a+b}}, \tag{5}$$

where $a = (s_1)^2/m$, and $b = (s_2)^2/n$. If the H_0 null hypothesis (cf. (3)) is true, the t' statistic can be approximated with a *t* distribution having degrees of freedom

$$df = \frac{(a+b)^2}{\frac{a^2}{m-1} + \frac{b^2}{n-1}} \tag{6}$$

(see, e.g., Wilcox, 1996, p. 133).

The normality assumption of the *t* test is always mentioned in the textbooks written for the social and behavioral sciences, but the effects of its violation are rarely explained in detail. For example, Maxwell and Delaney (1990) comment about the omnibus test in analysis of variance (ANOVA), the multi-group generalization of the *t* test, as follows: "ANOVA is generally robust to violations of the normality assumption, in that even when data are nonnormal, the actual Type I error rate is usually close to the nominal (i.e., desired) value" (p. 109). Although Maxwell and Delaney acknowledge that exceptions to this robustness can occur, others, particularly Wilcox (1996, p. 131), have argued that the robustness is assured only in the case of equal sample sizes, and even then the power can be unsatisfactory.

Increased interest in the effects of non-normality was prompted in part by Micceri (1989), who a decade ago presented a convincing case that data sets in education and psychology are very often not normally distributed. In a follow-up Monte Carlo study, Sawilowsky and Blair (1992) demonstrated that the familiar two-group *t* test is nonetheless reasonably robust when used in situations where both groups were sampled from one of the non-normal forms that had been documented by Micceri as relatively common in real data. Their results thus seemed to confirm the typical textbook advice about the robustness of such parametric procedures as the *t* test to

violations of the normality assumption.

However, a limitation of the Sawilowsky and Blair study is that the two samples being compared were always sampled from the same population distribution. In practice it may well be the case that a given dependent variable will display a similar type of non-normality in all groups, e.g. an achievement measure that is easy for most students may show extreme negative skew because of a consistent ceiling effect seen in all groups. But both simulation and mathematical results suggest that such similarity of distributional forms across groups may be critical.

Sawilowsky and Blair (1992, p. 359) note that their study did not change the conclusions reached by studies that "focused on populations modeled by well-known mathematical functions". For example, one such previous study by Posten (1978) demonstrated that the *t* test is "an extremely robust test statistic" (Cressie & Whitford, 1986, p. 137) with non-normal distributions when one has equivalent skewness and kurtosis across the two populations being compared. Similarly, Monte Carlo studies reported by Pearson and Please (1975) and O'Gorman (1995) showed that if the distribution of X has the same shape in the two populations, then the *t* test performs adequately with respect to a Type I error.

Based on mathematical derivations, some authors conclude that if the distribution of X is not normal, then in order to maintain the validity of the *t* test one has to insure that the skewness level (including its sign) be the same in the two populations to be compared (see, e.g., Scheffé, 1959, pp. 346-347; Miller, 1975, p. 42). In a paper providing helpful commentary on such derivations, Cressie and Whitford (1986, p. 135-137) remark that "the effect of perhaps large (but equal) skewnesses in the two populations, cancel".

Cressie and Whitford's formulas (1986, Equations 2.4-2.5) show that to approximate the value of the *t* test statistic in non-normal populations one must take into consideration not only the sample sizes and variances but also the skewness and kurtosis in the two groups being compared. In those special cases where sample sizes are equal and variances are equal, the *t* statistic is still affected by the difference in the skewness parameters for the populations being compared. The current simulation study was conducted to determine how large a difference in skewness levels the *t* test could tolerate without being affected by the varying distribution shapes across groups.

Simulation analyses investigating performance of the the robustness of the *t* test with non-

normal distributions have been less common than investigations of heterogeneity of variance. Some of the relevant literature is helpfully summarized by Cliff (1996, p. 144ff.). For example, based on their Monte Carlo analyses Pearson and Please (1975), and O'Gorman (1995) concluded that if the distribution of X has the same shape in the two populations, then the *t* test performs adequately with respect to a Type I error. However a simulation study of Algina, Oshima, and Lin (1994) revealed that the validity of the *t* test can be substantially lowered even with identical variances and distribution shapes in the situation where the sample sizes are markedly different. As an example, in the case of lognormal distribution (with skewness level, $\alpha_3 = 6.1$), m = 75, and n = 25 Algina et al. (1974) found that at the $\alpha = .05$ nominal level the Type I error rate was actually .065 for the lower one-tailed *t* test, and .036 for the upper one-tailed *t* test, whereas for the two-tailed *t* test it was .045 (see Algina et al., 1994, Table 3). It seems fair to say that performance of the *t* test when distributions have identical variances but different shapes is not well understood by researchers, and in particular there seems to be little understanding of how large a difference in skewness across the two distributions the *t* test can tolerate without an unacceptable decrease in its validity. Given this uncertainty, in practice one has essentially no information about the chance of a false decision with the *t* test if the distributions of X are skewed but the variances are not significantly different.

A logical idea is that the Welch test, which is offered as a substitute of the *t* test in the case of variance heterogeneity, may also be an appropriate alternative to *t* in the case of nonnormality. Unfortunately, the results of the simulation study of Algina et al. (1994) clearly show that the Welch test does not always fulfill this expectation. For example, with lognormal distributions, $\sigma_1 = \sigma_2$, m = 75, and n = 25, the estimated Type I error rate of the Welch test at $\alpha = .05$ level turned out to be .021, .100, and .063 in the lower one-tailed, upper one-tailed, and two-tailed cases respectively (see Algina et al., 1994, Tables 2 and 5). When the two samples were generated from different distributions but having identical variances (the first from a normal, the second from an exponential), then in the two-tailed version the type I error rate of the Welch test seemed to be acceptable (varying between .048 and .059) with equal sample sizes, but increased markedly (varying between .062 and .068) with unequal sample sizes (see Algina et al., 1994, Table 4). The performance of the Welch test in the heterogeneous case can be even worse if the distribution of X is extremely skewed (see Algina et al., 1994, Tables 2 to 9).

In the present study the effects of a number of different combinations of distribution shapes were investigated by systematically varying the the skewness and kurtosis levels in the two populations. Using simulation methods to approximate Type I error rates, we were able to determine what types of distribution pairs would invalidate Student's *t* test and the Welch test.

2. DESCRIPTION OF THE MONTE CARLO EXPERIMENT

2.1 Type of parent distributions

Random variates were generated from the generalized lambda family of distributions, which offers a variety of different shapes (Ramberg, Tadikamalla, Dudewicz, & Mykytka, 1979). These distributions are given in standardized form and can be described in terms of skewness ($\alpha_3 = \mu_3/\sigma^3$) and kurtosis ($\alpha_4 = \mu_4/\sigma^4$), where μ_3 and μ_4 are the third and fourth central moments. The generalized lambda family covers a wide range of values of skewness and kurtosis so that for any given value of skewness, several values of kurtosis can be specified (see Table 4 in Ramberg et al., 1979). For the present study three levels of skewness were applied, and for each level of skewness three levels of kurtosis were used (see Table 1). The lowest and highest levels of kurtosis always represent the most extreme levels available in Table 4 of Ramberg et al. (1979). The middle levels correspond to a medium level of kurtosis, which for a symmetric distribution gives a generalized lambda distribution having the first four moments equal to those of the standard normal. Note that the range of the possible kurtosis values depends heavily on the skewness level. At a higher skewness level both the minimal and maximal kurtosis values are higher than at a lower skewness level.

Table 1

Skewness (α_3) and kurtosis (α_4) values of the lambda-distributions used in the simulation.

	Kurtosis		
Skewness	Low	Moderate	High
symmetric	$\alpha_3 = 0$, $\alpha_4 = 1.8$	$\alpha 3 = 0, \alpha_4 = 3.0$	$\alpha_3 = 0$, $\alpha_4 = 9.0$
moderately asymmetric	$\alpha_3 = 1$, $\alpha_4 = 3.4$	$\alpha_3 = 1, \alpha_4 = 4.6$	$\alpha_3 = 1$, $\alpha_4 = 10.6$
heavily asymmetric	$\alpha_3 = 2$, $\alpha_4 = 8.6$	$\alpha_3 = 2, \alpha_4 = 9.8$	$\alpha_3 = 2$, $\alpha_4 = 15.8$

The three levels of skewness together with the three levels of kurtosis for each yielded nine different distribution types. Crossing the distribution types of the two samples yielded 9x9 = 81 different distribution combinations, all of them included in the simulation study. The asymmetric distributions listed in Table 1 are all positively skewed ($\alpha_3 > 0$). In order to investigate the appropriateness of the *t* test also for oppositely skewed distribution pairs, the six moderately and heavily asymmetric distributions appearing in Table 1 were crossed with six distributions of comparable skewness and kurtosis levels but with negative skewness. This yielded 6x6 = 36 more distribution pairs for the two samples.

The lambda distributions were generated in standardized form ($\mu = 0$, $\sigma = 1$) as is described in Ramberg et al. (1979). Therefore, the negatively-skewed distributions were generated in the same way as the positively-skewed distributions but with a simple multiplication by -1. In the presentation below of results involving oppositely skewed distribution pairs, the first distribution is always negatively skewed.

Thus, in the simulation the total number of different distribution pairs was 81 + 36 = 117.

2.2 Sample sizes

With respect to the sample sizes we varied the average sample size and the proportion of the two samples sizes. Average sample size was studied at four levels: 9, 12, 15, and 18. The ratio of the two sample sizes was either 1:1 or 1:2. Crossing these two factors yielded the following 8 sample size combinations:

(a) equal samples: (m = n = 9), (m = n = 12), (m = n = 15), (m = n = 18);

(b) unequal samples: (m = 6, n = 12), (m = 8, n = 16), (m = 10, n = 20), (m = 12, n = 24).

2.3 Technical details of the simulation

For each choice of the 117x8 = 936 simulation arrangements, 100,000 simulation iterations were used for assessing Type I error rates. At each iteration, $N = m + n$ random variates of the desired type were generated. The generalized lambda random variates were generated using the method described in Ramberg et al. (1979). In this generation process, Turbo Pascal's Random function was used to obtain pseudo-random uniform deviates. This is a linear congruential random-number generator that has turned out to be one of the most preferable in a recent study (Onghena,

1993), passing successfully ten criterion tests of randomness. Both the *t* and the Welch tests were then performed on the current set of *N* variates and evaluated in one- and two-tailed forms with significance levels .1, .05, and .01. Finally, the proportion of rejections was determined. This was an estimate of the Type I error rate. With the number of replications we were using (100,000), the standard deviation of an empirical Type I error rate was $[\alpha(1-\alpha)/100000]^{1/2}$, which yielded .00095, .00069, and .00031 if the true level were .1, .05, and .01 respectively.

3. RESULTS

3.1 The effect of the difference of the two skewness levels

First we analyzed the effect of the difference of the two skewness levels on the Type I error rates of the *t* test. Because in the simulation design we applied three levels of skewness (0, 1, 2), and the asymmetric distributions were combined to yield pairs with skewness in the same or opposite directions, this resulted in seven possible values of the difference between the two skewness levels: -4, -3, -2, -1, 0, 1, 2 (see section 2.1). For example one could get a difference of -4 for Skew1 - Skew2 if and only if the α_3 skewness values of populations 1 and 2 are -2 and 2 respectively. For the other Skew1 - Skew2 differences, the corresponding (Skew1, Skew2) pairs that were possible were: for -3, (-1, 2) or (-2, 1); for -2, (-1, 1) or (0,2); for -1, (0, 1) and (1, 2); for 0, (0, 0), (1,1) or (2,2); for 1, (1, 0), (2, 1); for 2, (2, 0).

Grouping the 117 distribution pairs according to this factor, the minimum, maximum, and mean values of the obtained Type I error estimates were calculated for each of the above seven groups. The resulting statistics at $\alpha = .05$ and an average sample size of 9 are presented in Figures 1 to 3 for lower one-tailed, upper one-tailed, and two-tailed tests, respectively.

(Insert Figures 1 to 3 about here)

These figures present convincing evidence that, at an average sample size of 9, a large difference in skewness causes a substantial loss of validity of the *t* test. Because a difference in skewness of the variables whose means are being compared causes the observed distribution of the *t* statistic also to be skewed, the bias in Type I errors is especially high if one uses a one-tailed form of the *t* test. In this case the Type I error rate can be decreased or increased by 40% relative

to the nominal level even if the difference of the two skewness levels is as low as 1 (see in Figure 1 the minimum value and in Figure 2 the maximum value corresponding to Skew1 - Skew2 = -1). If the difference of the skewness levels is larger, the bias can be far greater (see Figures 1 and 2).

The bias is milder in the two-tailed form of the *t* test (see Figure 3). It is worth noting that in this case severe bias occurs only in the positive direction (causing an inflation of the *t* test) and that only when the Skew1-Skew2 difference is -2 or more extreme. In the most extreme case, where Skew1 - Skew2 = -4, the estimated Type I error rate varies between .063 and .082, with an average of .075. In this case the range of the bias is 26-64%, but it can reach 54% if Skew1 - Skew2 = -3 (see Figure 3).

(Insert Figures 4 to 6 about here)

If the average sample size is doubled (from 9 to 18), one gets a similar but less pronounced bias (see Figures 4 to 6). Nevertheless the inflation in the Type I error rate can exceed the nominal level by more than 80% even in this case if one uses a one-tailed form of the *t* test (see Figure 5, results corresponding to Skew1 - Skew2 = -4). However, in the two-tailed case the bias of the Type I error rate does not exceed 40%, and it always remains in the ±20% region provided that the Skew1 - Skew2 difference is not less than -2 (see Figure 6).

(Insert Figure 7 about here)

The same mild degree of bias occurs also with an average sample size of 9 if one uses $\alpha = .1$ instead of $\alpha = .05$ (see Figure 7). By contrast, at the .01 alpha level the extent of bias of the Type I error rate exceeds every acceptable limit if the Skew1 - Skew2 difference is too negative (see Figure 8). Doubling again the average sample size (to $m + n = 36$) we can see that in this case the validity of the *t* test is already acceptable even at .01 alpha level, provided that the Skew1 - Skew2 difference does not exceed -1 in the negative direction (see Figure 9).

(Insert Figures 8 and 9 about here)

3.2 The effect of kurtosis

The effect of the difference of the skewness levels on the Type I error rate has been shown to be rather strong (see section 3.1). However, it is also striking that the difference between the minimum and maximum Type I error rates was rather large under certain conditions (e.g., with an average sample size of 9, and $\alpha = .05$). For example in Figure 3 when Skew1 - Skew2 is -2, the minimum of the Type I error estimates is only .050 (right at the nominal level), whereas the maximum is .068, which exceeds the nominal level by 36%. Similarly, at Skew1 - Skew2 = -3 the Type I error estimates fall into the range (.054 -.077), which corresponds in terms of percentage bias to: (8% - 54%).

These findings indicate that in addition to skewness other factors, such as kurtosis, may also play a nonnegligible role in determining the actual Type I error rate of the *t* test. To check this possibility all distribution pairs were divided into three groups. A pair was put into the group called "Normal-tailed pairs" if neither of the two distributions had a high kurtosis level in terms of the classifications in Table 1. By contrast the group called the "Long-tailed pairs" consisted of those pairs where both distributions had a high kurtosis level. The remaining distribution pairs formed the group of "Mixed-tailed pairs". The combined effect of this grouping factor and the Skew1 - Skew2 difference with an average sample size of 9 and $\alpha = .05$ is illustrated in Figure 10.

(Insert Figure 10 about here)

Figure 10 confirms that kurtosis also exerts a nonnegligible influence on the Type I error rate of the *t* test. It is regrettable, however, that this effect is to inflate the error rate in the group of "Normal-tailed pairs", which may occur most often in practice. Similar to what has been observed with the effect of kurtosis on the Type I error rate of the one-sample-t test (Vargha, 1996; Vargha & Delaney, in press-a), a high kurtosis level of both distributions may compensate for the strong inflating effect of skewness in the case of the two-sample *t* test as well. The effect of kurtosis is similar but milder in the case of larger samples (for average sample size = 15 see Figure 11).

(Insert Figure 11 about here)

In the case of the Welch test the combined effect of skewness difference and kurtosis level is presented in Figure 12. The picture here is much the same as in the case of the *t* test. For this reason the Welch test does not seem to be a reasonable alternative when the normality assumption of the *t* test is severely violated. Figure 13 shows that the combined distorting effect of the above two factors may be considerable even with larger samples (average sample size = 18).

(Insert Figures 12 and 13 about here)

3.3 Effects relating to sample size

Results that have been discussed in sections 3.1 and 3.2 indicate that the increase in the average sample size may substantially reduce the distortion caused by asymmetry (see Figures 1 vs. 4, 2 vs 5, 3 vs. 6, 8 vs. 9, 10 vs. 11, and 12 vs. 13). However, in the foregoing we have pooled the data from equal and unequal sample sizes at each level of average sample size. In the current section we focus on the effects of the factor of inequality of sample sizes.

From Figures 3, 6, and 7 to 11 one can discern that the validity of the *t* test begins to become unacceptably low only for those pairs of distributions where Skew1 is less than Skew2 by at least 2 units (i.e., when Skew1 - Skew2 $\leq$ -2). Since the bias of the Type I error rate is always less for Skew1 - Skew2 = 2 than for Skew1 - Skew2 = -2, we may presume that the inequality of the the signs of the two skewness parameters is itself a distorting agent of the validity of the *t* test (in our design Skew1 - Skew2 = 2 occurs if and only if Skew1 = 1 and Skew2 = 0, but Skew1 - Skew2 = -2 occurs if either Skew1 = 0 and Skew2 = 2 or Skew1 = -1 and Skew2 = 1). Based on this idea we determined the maximum Type I error rates of the *t* test at each level of average sample size for $m = n$ and $m \neq n$ separately, segregating also the oppositely skewed distribution pairs (i.e., where the sign of skewness is negative for one distribution and positive for the other distribution) and identically skewed distribution pairs (i.e., where the signs of skewness do not differ) (see Figure 14).

(Insert Figure 14 about here)

Figure 14 allows to draw the following conclusions.

1. The maximum possible Type I error rate is consistently higher for oppositely skewed distributions than for identically skewed ones.

2. The inequality of sample sizes is a second factor that tends to increase the Type I error rate.

3. If the average sample size increases, the maximal possible bias decreases. However, the Type I error rate at $\alpha = .05$ level can exceed .07 (i.e., 40% of the nominal level) even in the case of average sample size = 18, provided that the sample sizes are different and the distributions are oppositely skewed.

4. In that special case when the sample sizes are equal and the distributions are not oppositely skewed the Type I error rate of the *t* test at $\alpha = .05$ never exceeds .056 even if the average sample size is as low as 9. This confirms the robustness of the *t* test in the equal-*n* situation.

From Figure 15 one can draw the same conclusions with respect to the Welch test as well.

(Insert Figure 15 about here)

A new and relevant question arises now. Will this nice robustness of Student's *t* test and Welch-test under equal sample sizes and not oppositely skewed distributions remain if the theoretical variances are allowed to differ to a slight extent? O'Gorman (1995) carried out a similar simulation study with Student's *t* test, Welch test and some nonparametric tests for assessing and comparing the validity and power of these two-sample procedures. For the simulation O'Gorman applied the same generalized lambda-family of distributions as in the current study, but the two distributions to be compared always had identical skewness and kurtosis levels. The variances were allowed to differ only slightly: $\sigma_{max}/\sigma_{min} \leq 1.3$. Though O'Gorman's analysis concentrated on the power of the *t* test and its alternatives, the estimated Type I error rates were also reported. O'Gorman found that at $\alpha = .05$ the estimated Type I error rates of the *t* test always fell into the region (.035 - .065) (O'Gorman, 1995, p. 858). It must also be noted here that the average sample size levels used in that study were 12, 50, 200, and 800, and the sample size ratios used were 1:4, 1:1, and 4:1. O'Gorman also reported that the validity of the Welch test was similar to that of the *t* test under $m, n \geq 50$, but for $m = 12$, $n = 50$ and Skew1 = Skew2 = 3 the Type I error rate of the Welch test increased markedly.

The results of O'Gorman's Monte-Carlo study can briefly be summarized as follows. If the two distributions to be compared have identical skewness and kurtosis levels and the larger population variance is not greater than the smaller one by more than 30%, then the *t* test maintains its validity even with considerably different sample sizes.

However, the results of the present study revealed that the *t* and Welch tests can maintain their validity even if the shapes of the two distributions are different but not oppositely skewed, provided that the average sample size is not less than 15 (see Figures 14 and 15) and the population variances are equal. This latter requirement of variance homogeneity may probably be weakened under $m = n$, since according to the results of our study the equality of sample sizes substantially increases the robustness of both the *t* test and the Welch test even at an average sample size of 9 (see Figures 14 and 15).

In order to check this hypothesis we carried out a new simulation analysis, using only those 81 distribution pairs where the two distributions are not oppositely skewed (see section 2.1). In these arrangements the SD of the second population was always twice as large as that of the first distribution: $\sigma_2 = 2\sigma_1$. The applied sample sizes were m = n = 9 and m = n = 18. The results concerning the maximal obtained Type I error rates of the two tests grouped according to the maximal skewness level are presented in Figure 16. The maximal skewness level of a distribution pair is simply the larger of the two skewness levels. It is 0 if both distributions are symmetric (Skew1 = Skew2 = 0); 1 if Skew1 = 0 and Skew2 =1, or Skew1 = 1 and Skew2 =0, etc.

(Insert Figure 16 about here)

Based on Figure 16 one can draw the following conclusions.

1. The possible maximal Type I error rate of both tests is positively related to the maximal skewness level.

2. With larger samples ($m = n = 18$), the maximal Type I error rate of the *t* test is lower than with smaller samples ($m = n = 9$). However, if the maximal skewness level is as large as 2, the Type I error rate may still be as large as 140% of the nominal level. On the other hand, if neither distribution has a skewness level greater than 1 the bias of the *t* test never exceeds the acceptable ±20%.

3. The maximal possible Type I error rates of the Welch test are somewhat smaller than those of the t test. As a result, if $m = n = 18$ the Welch test seems to perform quite well: even if the maximal skewness level reaches the value of 2, the maximal Type I error rate does not exceed .068.

Summarizing: if the two sample sizes are equal and not low ($m, n \geq 18$), the skewness levels are not extreme (max(Skew1, Skew2) ≤ 1) and not opposite, and the populations' SD's do not differ to a large extent ($\sigma_{max}/\sigma_{min} \leq 2$), then the Type I error rates of the t test and the Welch test are close to the nominal level in that the bias does not exceed ±20%.

4. DISCUSSION

The effect of violation of the variance homogeneity assumption of Student's two-sample t test has been studied often and thoroughly by many authors. However, we cannot say the same with respect to its normality assumption though recent studies indicate that nonnormal distributions occur very frequently in practice (for examples from educational and behavioral investigations, see Micceri, 1989). Therefore in the present study the validity of the t test was assessed under several different conditions where the populations SD's were constrained to be equal ($\sigma_1 = \sigma_2$). We systematically varied the skewness and kurtosis levels, the average sample size (9, 12, 15, 18), the ratio of the sample sizes (1:1, 1:2), the type of alternative hypothesis (lower one-sided, upper one-sided, two-sided), and the α-level (.1, .05, .01). Along with the t test we analyzed the validity of its best known robust alternative, the Welch test.

In general, our simulation results show the practical consequences of the mathematical difficulties caused by differences in skew. Wilcox (1990) has reported that variation in skewness results in a lack of independence of the numerator and denominator of the t. It turns out that the correlation between the mean difference in the numerator and the square of the denominator of the test is a function of the skewness of each distribution. The correlation is larger when the two groups have distributions that are skewed in opposite directions, but the correlation decreases as sample sizes increase. The current simulation thus documents the extent to which the Type I error rate is influenced by this lack of independence of the numerator and denominator of the test statistic.

Based on our simulation results we can draw the following general conclusions.

Different distribution shapes may substantially lower the validity of the *t* test in the range of average sample sizes we investigated (9 - 18) which occurs frequently in practice. The most crucial invalidating factor observed was the difference between the two skewness levels. At its worst this factor may virtually invalidate the *t* test in spite of $\sigma_1 = \sigma_2$.

The one-tailed forms of the *t* test are much more sensitive to violation of the normality assumption than the two-tailed version. In the one-tailed case the extent of bias can be 100% greater than the nominal level (see Figure 2), and or as much as 60% less (see Figure 1).

In the case of the two-tailed test the bias is somewhat smaller, because the one-sided effects are opposite and therefore they partly neutralize each other. It is important to note that a substantial bias in this case occurs only in the upper direction (causing an inflation of the *t* test). If the two distributions are oppositely skewed, the Type I error rate can exceed the nominal level by more than 60% (see Figure 3).

The effect of nonnormality on the Type I error rate is milder at the .1 alpha-level, and is most pronounced at the .01 level (see Figures 8 and 9).

Regarding the effects of other factors, as with heterogeneity of variance, the smallness and inequality of sample sizes can also exacerbate the effects of skewness on the Type I error rate. The effects of non-normal kurtosis were not as pronounced as the effects of differential skewness, but were nonetheless present. Interestingly, high kurtosis distributions with their long tails tended to suppress the tendency of differences in skewness to inflate the Type I error rate. Differences in skewness had most pronounced effects when the kurtosis level of the distributions being compared was equal to that of the normal distribution.

An important result of the current study is that the above mentioned factors influence the validity of the Welch test the same way as that of the *t* test. Therefore the Welch test cannot be claimed to be a generally appropriate robust alternative to the *t* test. Nevertheless some usable guidelines can be formulated based on Figures 14 to 16. If the sample sizes are nearly equal and not very low (say m, $n \geq 15$), and the difference of the populations SD's is not striking (say $\sigma_{max}/\sigma_{min} \leq 2$), then the validity of the *t* test and Welch test will still be acceptable, provided that the two distributions ar not heavily and oppositely skewed.

In conclusion, the implication of this study is that differential skewness across groups can strongly influence the Type I error rate of the two-group *t* test. This effect can be exacerbated by

unequal sample sizes across the two groups. One rule of thumb that could be suggested is that if the frequency distributions in the two groups have opposite skewness the *t* test may not be valid, particularly with unequal *n*. Stated numerically, one should pursue an adjusted test if the product of the ratio of the sample sizes and the absolute value of the difference in skewness levels is greater than 4. Adjustments in the critical value of the test statistic that incorporate information about sample size, skewness and kurtosis are presented in Cressie and Whitford (1986). However, the adjustments are very complicated numerically and have been reported by Wilcox (1990) to sometimes "make matters worse" (Wilcox, 1996, p. 136).

Alternatively, in the case of heavily skewed distributions one might question the appropriateness of the mean as a location parameter, which should represent the bulk of the distribution. Some authors (e.g., Wilcox, 1996, 1998) recommend in this case modern robust alternative procedures based on other location parameters (such as medians or trimmed means), but the adequacy of these methods has not been empirically demonstrated for a wide range of nonnormal distributions yet. An alternative solution may also be to test the stochastic equality of the two populations rather than the equality of the two population means (or medians or trimmed means). Two populations are said to be stochastically equal with respect to a variable that is at least ordinally scaled if

$$P(x > y) = P(x < y),$$

where x and y are values taken at random and independently from populations 1 and 2 respectively (see Vargha & Delaney, 1998, 2000, in press-b). These authors report a recently developed and empirically validated robust test of stochastic equality denoted as the FPW test, which is a modification of a nonparametric two-group test due to Fligner and Policello (1981). The FPW test is available in the most recent version of the MiniStat program package (Vargha, 1999).

A limitation of the current study is that we have focused exclusively on Type I error rates, and thus have only examined the effect of nonnormality in situations where the null hypothesis of equality of means is true. We do not have information on the prevalence of variations in skewness across groups in multi-group studies. (Although Micceri (1989) indicated skewed distributions

were relatively common in education and psychology, he reported only data for the total sample rather than variation in skewness values across subsamples.) Such changes in skewness without changes in means or variances are likely rare, though perhaps conceivable in cases where the range of values on the dependent variable is strictly constrained, e.g. in the case of a Likert scale. But in reality differences in skewness seem more likely to arise in conjunction with mean differences (and perhaps heterogeneity of variance) across groups. For example, in an educational study, a floor effect (and the associated positive skewness) in a control group on an achievement measure may be removed in a treatment condition that raises average performance. Similarly, in an alcohol treatment study, the extreme positive skewness in drinking levels in an untreated sample may be greatly reduced in an effective treatment condition that lowers the mean level of alcohol consumption in part by greatly reducing the drinking levels of the heaviest drinkers. Thus, it will be important to supplement the current study of Type I error rates with studies of the effects of differences in skewness on power, and with investigation of alternative procedures that might not be as susceptible to inflation of Type I error rates as the two-sample *t* test.

REFERENCES

Algina, J., Oshima, T. C., & Lin, W. Y. (1994). Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics, 19*, 275-291.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis.* Mahwah, NJ: Lawrence Erlbaum.

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

Fligner, M. A., & Policello II, G. E. (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association, 76,* 323-327.

Maxwell, S. E. & Delaney, H. D. (1990). *Designing experiments and analyzing data. A model comparison perspective.* Belmont, California: Wadsworth Publishing Company.

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

Miller, R. G., Jr. (1986). *Beyond ANOVA: Basics of applied statistics*. New York: John Wiley.

O'Gorman, T. W. (1995). The effect of unequal variances on the power of several two-sample tests. *Communications in Statistics — Simulations, 24*, 853-867.

Onghena, P. (1993). A theoretical and empirical comparison of mainframe, microcomputer, and pocket calculator pseudorandom number generators. *Behavior Research Methods, Instruments, & Computers, 25,* 384-395.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62*, 223-241.

Posten, H. O. (1978). The robustness of the two sample *t*-test over the Pearson system. *Journal of Statistical Computation and Simulation, 6*, 295-311.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J., & Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics, 21*, 201-209.

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111*, 352-360.

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

Vargha A. (1996). Az egymintás t-próba érvényessége és javíthatósága. [The one-sample t test: validity and improvements] *Magyar Pszichológiai Szemle, LII (36), 4-6*, 317-345.

Vargha A. (1999). *MiniStat 3.1 verzió. Felhasználói kézikönyv.*[*MiniStat* 3.1 version. Manual] Budapest: Pólya Kiadó.

Vargha, A., & Delaney, H.D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics, 23*, 170-192.

Vargha, A., & Delaney, H. D. (in press-a). The effect of kurtosis on Student's one-sample t test. In Bruno D. Zumbo (Ed.), *Social Indicators and Quality of Life Research Methods: Methodological Devleopments and Issues, Yearbook 2000.* Boston: Kluwer Academic Publishers.

Vargha, A., & Delaney, H. D. (in press-b). A critique and improvement of the CL common language effect size statistic of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, in press.

Vargha, A., & Delaney, H. D. (2000). Comparing several tests of stochastic equality. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April, 2000.

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

Wilcox, R. R. (1996). *Statistics for the social sciences.* San Diego, New York: Academic Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods. *American Psychologist, 53*, 300-314.

Figure 1
Empirical Type I error rate of the two-sample t test
as a function of the difference of the two skewness levels
(lower one-tailed test, $\alpha = 5\%$, $m + n = 18$)

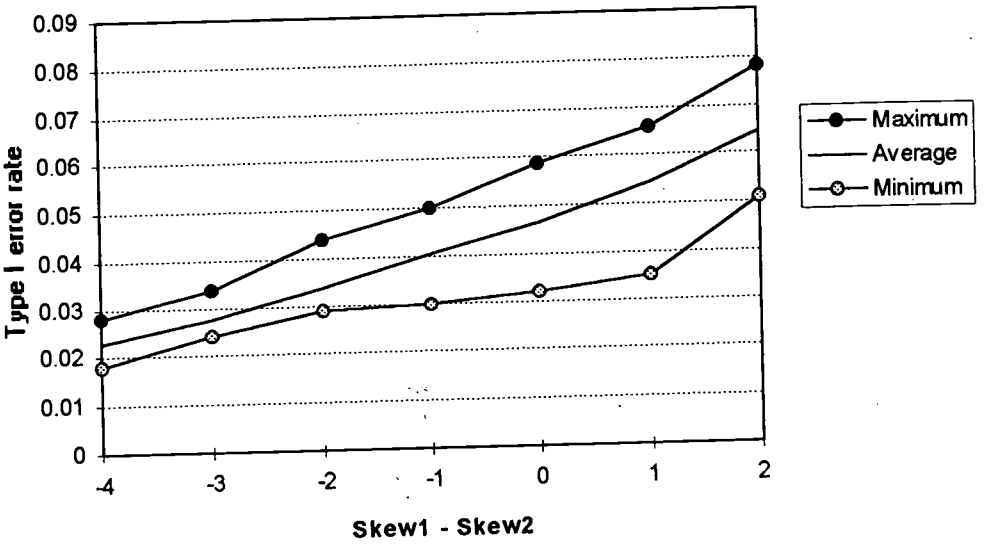


Figure 2
Empirical Type I error rate of the two-sample t test
as a function of the difference of the two skewness levels
(upper one-tailed test, $\alpha = 5\%$, $m + n = 18$)

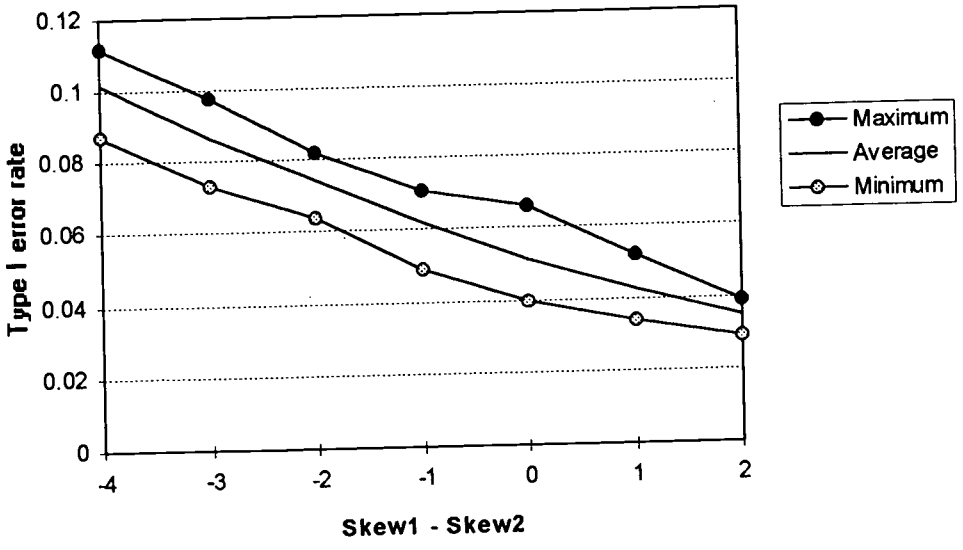


Figure 3
Empirical Type I error rate of the two-sample t test
as a function of the difference of the two skewness levels
(two-tailed test, $\alpha = 5\%$, $m + n = 18$)

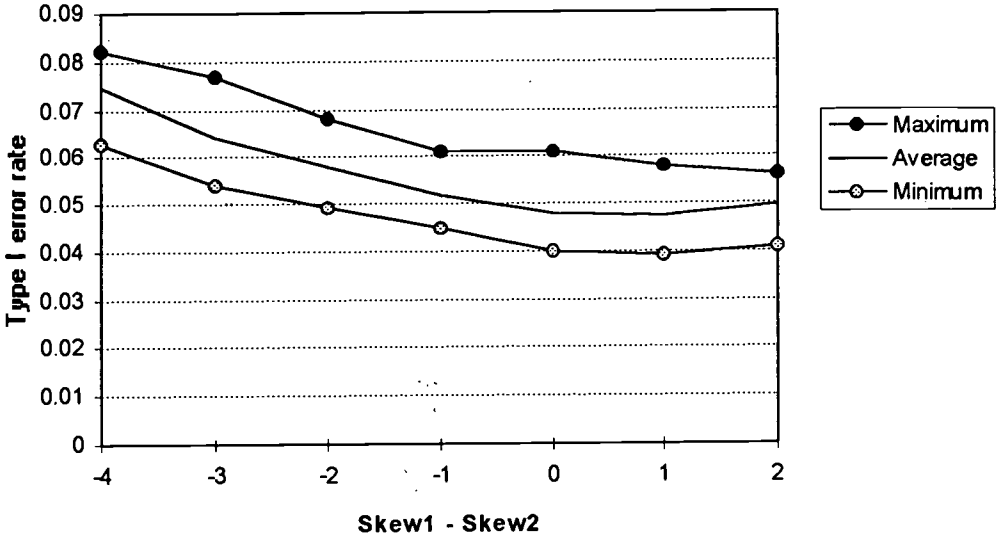


Figure 4
Empirical Type I error rate of the two-sample t test
as a function of the difference of the two skewness levels
(lower one-tailed test, $\alpha = 5\%$, $m + n = 36$)

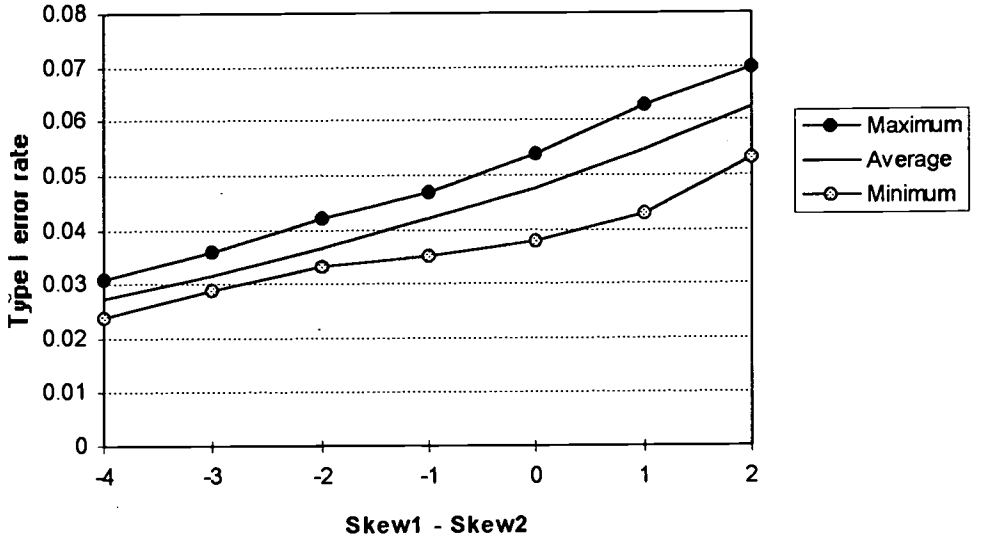


Figure 5
Empirical Type I error rate of the two-sample t test
as a function of the difference of the two skewness levels
(upper one-tailed test, $\alpha = 5\%$, $m + n = 36$)

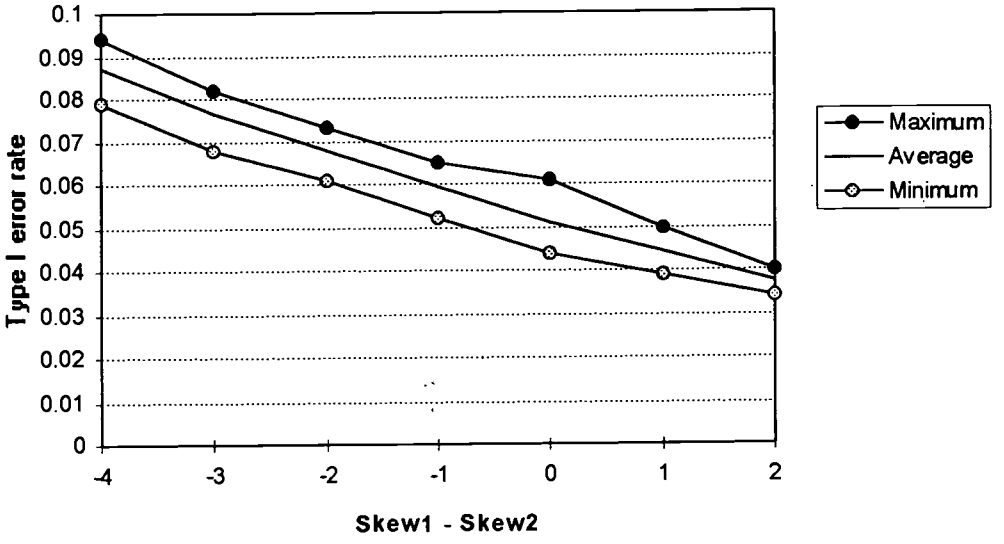


Figure 6
Empirical Type I error rate of the two-sample t test
as a function of the difference of the two skewness levels
(two-tailed test, $\alpha = 5\%$, $m + n = 36$)

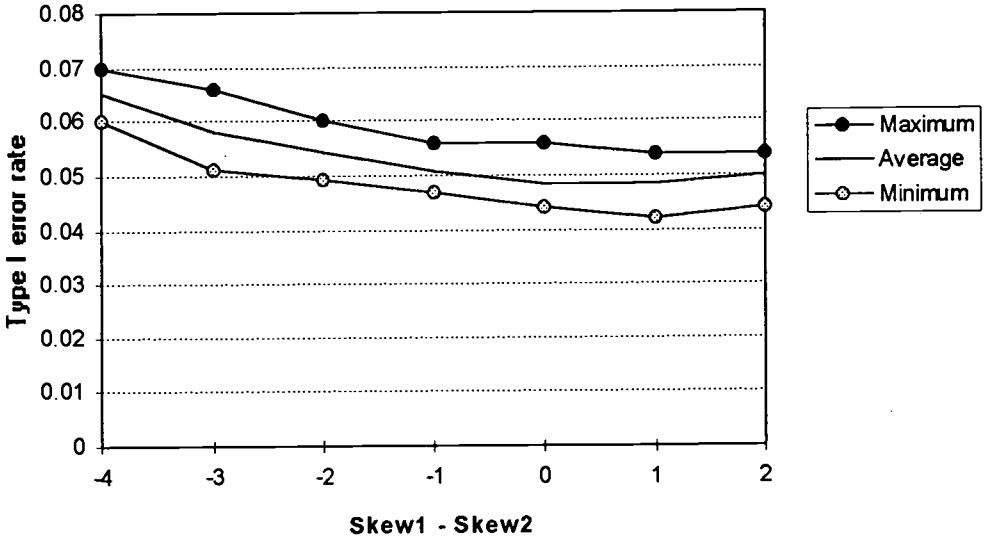


Figure 7
Empirical Type I error rate of the two-sample t test
as a function of the difference of the two skewness levels
(two-tailed test, $\alpha = 10\%$, $m + n = 18$)

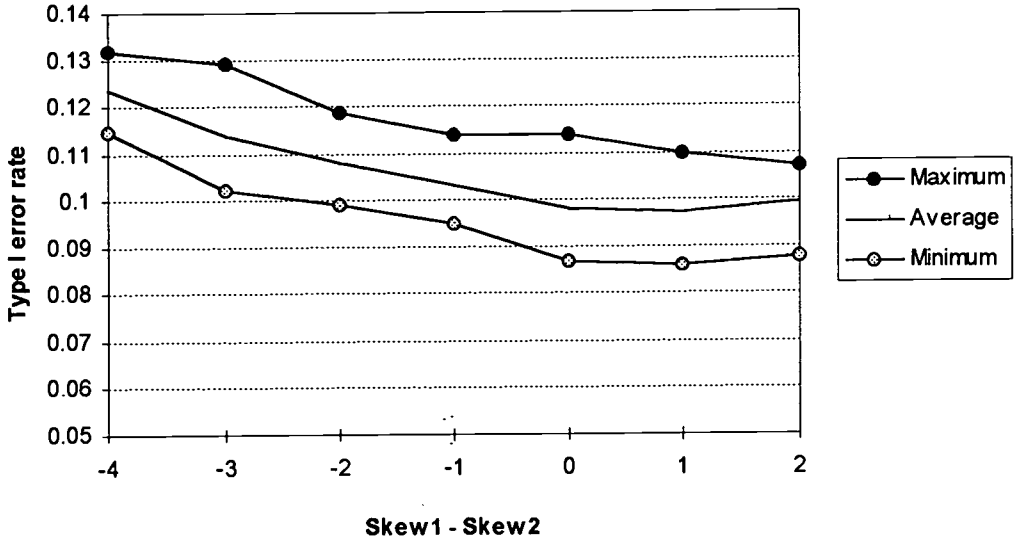


Figure 8
Empirical Type I error rate of the two-sample t test
as a function of the difference of the two skewness levels
(two-tailed test, $\alpha = 1\%$, $m + n = 18$)

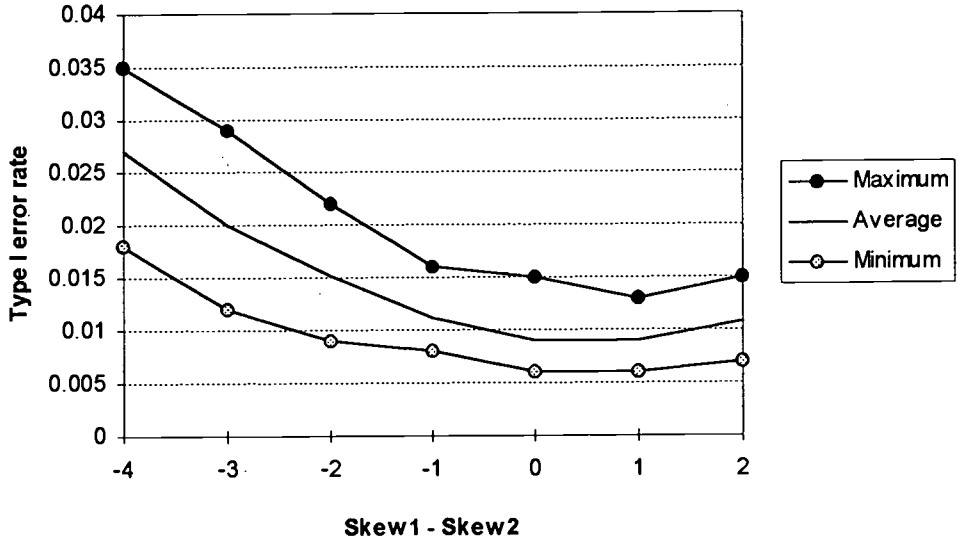


Figure 9
Empirical Type I error rate of the two-sample t test
as a function of the difference of the two skewness levels
(two-tailed test, $\alpha = 1\%$, $m + n = 36$)

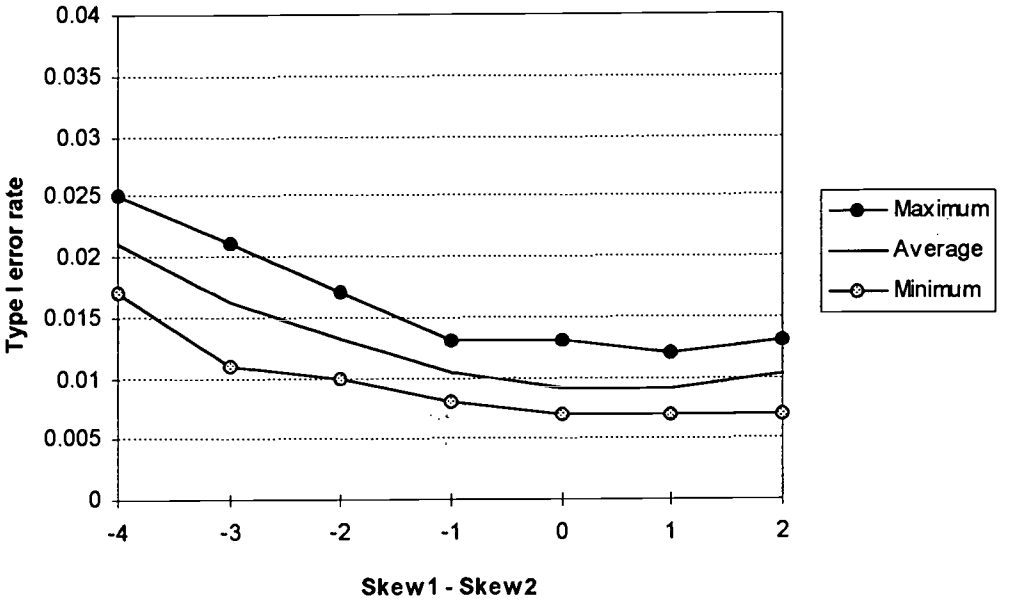


Figure 10
Empirical Type I error rate of the two-sample t test as a function of the difference
of the two skewness levels for different combinations of kurtosis levels
(two-tailed test, $\alpha = 5\%$, $m + n = 18$)

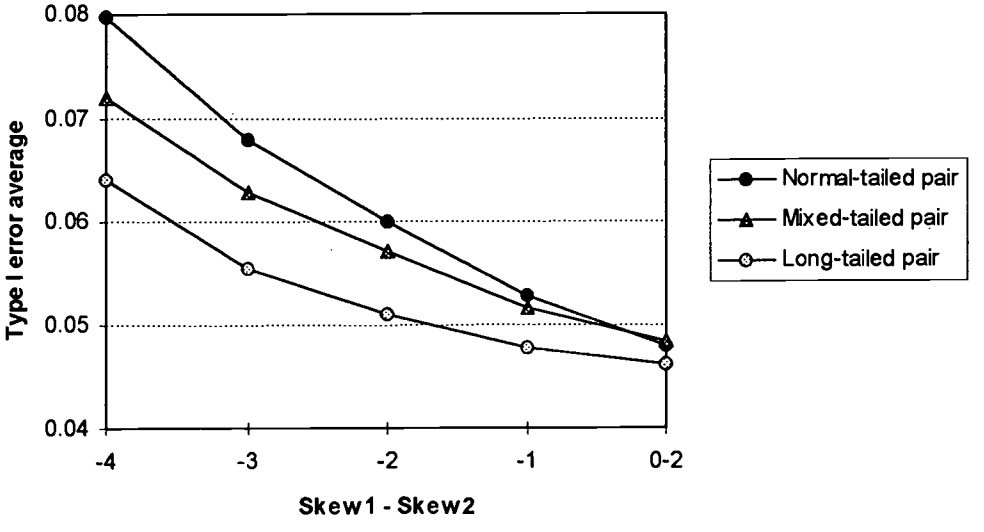


Figure 11
Empirical Type I error rate of the two-sample t test as a function of the difference of the two skewness levels for different combinations of kurtosis levels (two-tailed test, $\alpha = 5\%$, $m + n = 30$)

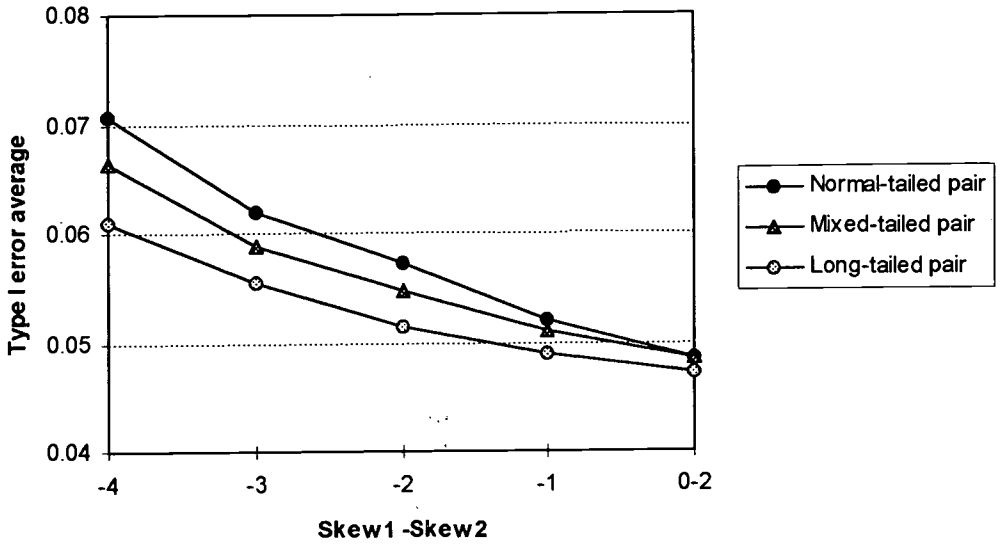


Figure 12
Empirical Type I error rate of the Welch-test as a function of the difference of the two skewness levels for different combinations of kurtosis levels (two-tailed test, $\alpha = 5\%$, $m + n = 18$)

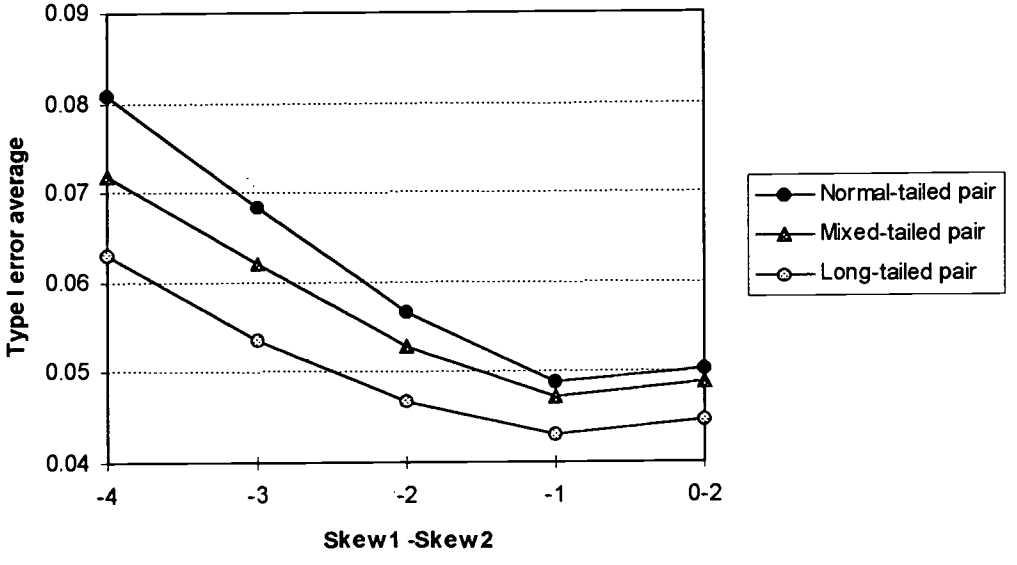


Figure 13
Empirical Type I error rate of the Welch-test as a function of the difference of the two skewness levels for different combinations of kurtosis levels (two-tailed test, $\alpha = 5\%$, $m + n = 36$)

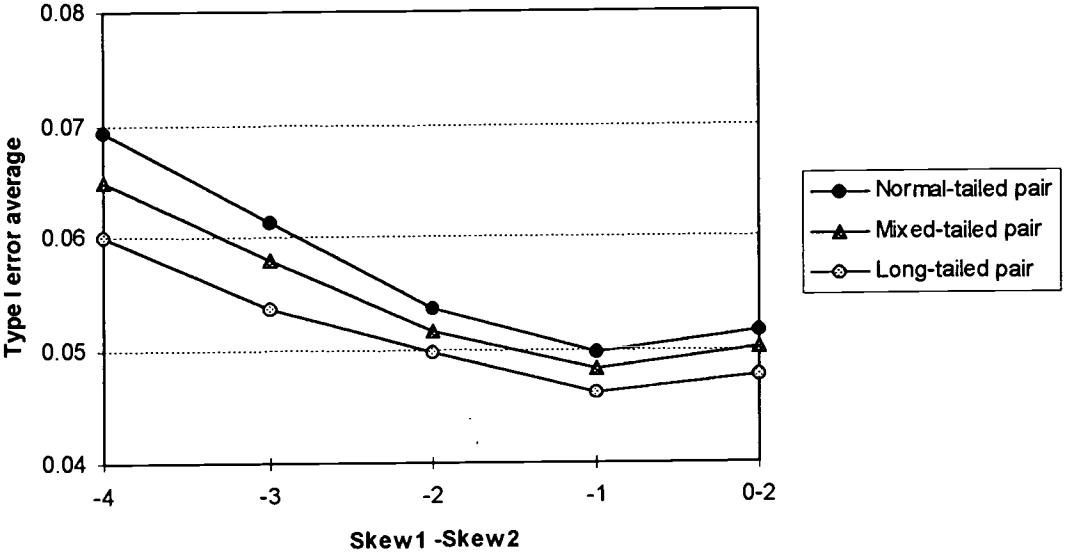


Figure 14
Maximal Type I error rate of the two-sample t test as a function of the average sample size for identically and oppositely skewed distribution pairs with identical and different sample sizes (two-tailed test, $\alpha = 5\%$)

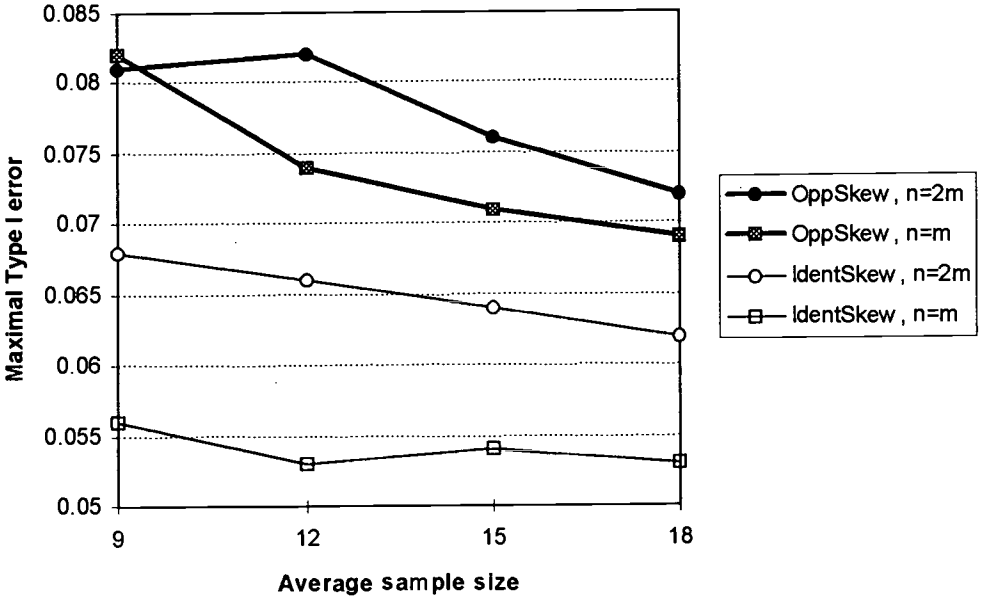


Figure 15
Maximal Type I error rate of the Welch-test
as a function of the average sample size for identically and oppositely skewed distribution pairs with identical and different sample sizes (two-tailed test, $\alpha = 5\%$)

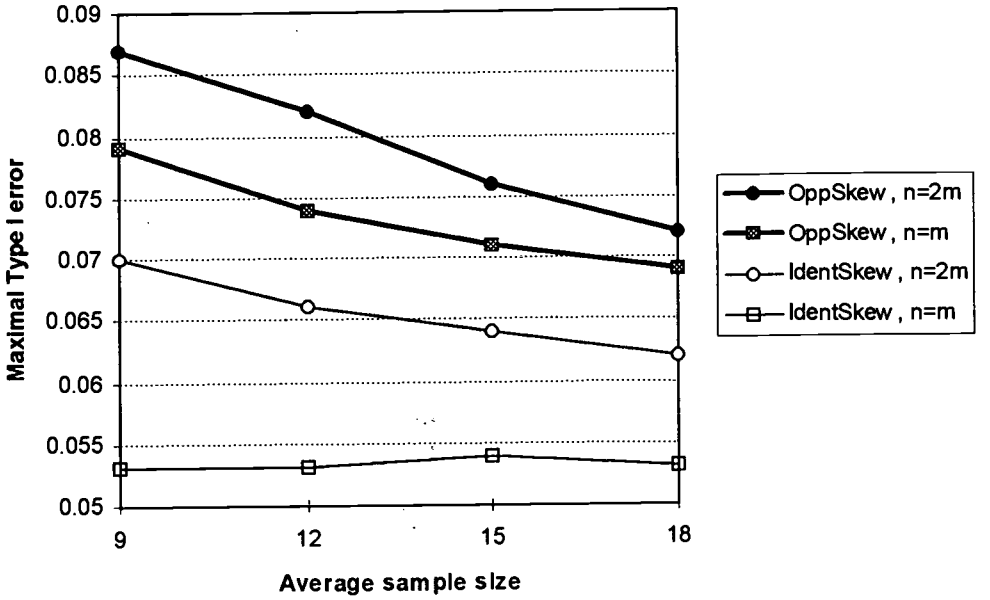
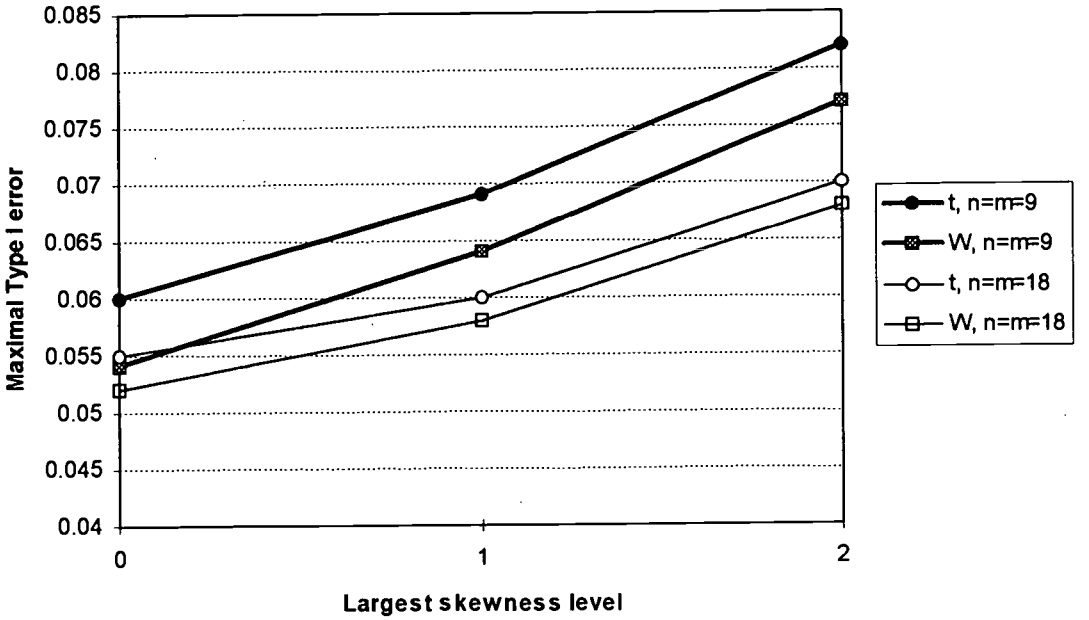


Figure 16
Maximal Type I error rate of the two-sample t and Welch tests
as a function of the largest skewness level for not oppositely skewed distribution pairs with identical sample sizes and $\sigma_2 = 2\sigma_1$ (two-tailed test, $\alpha = 5\%$)



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031495

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Effect of Nonnormality on Student's Two-Sample t Test

Author(s): Harold D. Delaney and András Vargha

Corporate Source: University of New Mexico Department of Psychology & ELTE Institute of Psychology (Hungary)

Publication Date: AERA Annual Meeting New Orleans, April, 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Harold Delaney	Printed Name/Position/Title: Harold Delaney, Professor
Organization/Address: Dept. of Psychology, University of New Mexico, Albuquerque, NM 87131	Telephone: 505-277-5224 FAX: 505-277-1394
	E-Mail Address: hdelaney@unm.edu Date: 6/6/2000

(over)

ERIC®

Clearinghouse on Assessment and Evaluation

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

May 8, 2000

Dear AERA Presenter,

Hopefully, the convention was a productive and rewarding event. As stated in the AERA program, presenters have a responsibility to make their papers readily available. If you haven't done so already, please submit copies of your papers for consideration for inclusion in the ERIC database. We are interested in papers from this year's AERA conference and last year's conference. If you have submitted your paper, you can track its progress at http://ericae.net.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the **2000 and 1999 AERA Conference**. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form enclosed with this letter and send **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can mail your paper to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 2000/ERIC Acquisitions
University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE