National Library of Canada

Acquisitions and Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada

Direction des acquisitions et des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Canada

An Investigation of the Robustness of the Type I Error Rate of the t, M-W-W, Welch and Welch on Ranks Tests Applied to Reaction Time Populations with Unequal Variances

by

Luba Mycio-Mommers

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfilment of the requirements for the
Masters of Arts (Measurement & Evaluation) Degree
Faculty of Education
University of Ottawa

ISBN 0-612-11584-4

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ABSTRACT

This thesis investigates the robustness of the Type I error rate of the t test, the t test on ranks (Mann-Whitney-Wilcoxon), the Welch test and the Welch test on ranks applied to reaction time populations with unequal variances. Reaction time is often encountered as a dependent variable in educational and behavioural research. Reaction time data are typically skewed and are commonly modelled on a family of distributions known as the ex-Gaussian. A Monte Carlo study compared the robustness of Type I error rates of the four tests under study under 36 conditions wherein four factors were observed: total sample size (N = 24 and 72), ratio of sample sizes (n_1:n_2 = 1:1, 1:2, and 1:3); ratio of population variances (var1:var2 = 1:1, 1:2, 1:4, and 1:9), and negative and positive conditions. In each condition, 5,000 scores were generated from Miller's (1988) most skewed distribution that represented a boundary condition of reaction time data. The results indicated that the t test was the preferred option under all simulated conditions, except the negative condition. Furthermore, under the negative condition, all four tests produced liberal Type I errors.

ACKNOWLEDGEMENTS

I wish to thank warmly my supervisor, Dr. Bruno Zumbo, for his assistance throughout the study. His guidance, support, observations, and recommendations were invaluable.

I am grateful for the the support and recommendations of Dr. Marc Gessaroli. The recommendations of Dr. Daniel Coulombe were also truly appreciated.

I would also like to thank Ms. Madeleine Lalonde, Academic Assistant at the Faculty of Education at the University of Ottawa, for her assistance.

The support of my husband, Dr. Alexander Mommers, and my son, Stephan, was greatly appreciated.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

APPENDIXES

CHAPTER 1

Introduction to the Problem

Let us consider the following scenario in experimental research:

> A researcher is interested in reaction times (recorded in milliseconds) as a dependent variable in an experiment involving letter recognition. The experimental design consists of two independent groups of unequal sample size, wherein group A is a control group (n = 30) and group B (n = 10) is an experimental group. As expected with reaction time data, the distribution of scores in each group is skewed (wherein the mean minus the median equals 92). Furthermore, the researcher finds himself in the situation wherein not only is the data skewed but the variance of the experimental group is four times larger than that of the control group.

Given that the researcher is interested in testing the equality of the two means, what are his options in this scenario? His first option may be to use the independent samples t test, however, it may be inappropriate. The reason why the t test might not be appropriate is because it requires that the scores arise from populations that are normally distributed and have equal variances (commonly referred to as homogeneity of variance). In my scenario these conditions are not met.

However, there is general agreement that the t

test can be used under moderate violations of homogeneity of variance provided that sample sizes are equal. The literature, however, does not provide any guidance as to what is meant precisely by the expression "moderate violations". It has been demonstrated, however, that Type I error rates are influenced by unequal variance in the following ways when sample sizes are unequal: (a) the Type I error rates are depressed when the larger variance is associated with the larger sample size (positive condition); and (b) the Type I error rates are elevated when the larger variance is associated with the smaller sample size (negative condition) (see, for example, Ramsey, 1980; Scheffé, 1959). In my scenario, the researcher is faced with the negative condition and thus, the t test may be inappropriate again because the Type I error rate may be inflated above the nominal value which is usually .05.

As a second option, many introductory statistical textbooks call for the use of nonparametric methods, such as the Mann-Whitney-Wilcoxon (M-W-W) test, or its k-group equivalent, the Kruskal-Wallis (KW) test, when there are violations of normality and/or homogeneity of variance. Recent studies have shown, however, that nonparametric tests are sensitive to unequal variances

in combination with unequal sample sizes and perform much like their parametric counterparts (Zimmerman, 1987; Zimmerman & Zumbo, 1993a, 1993b). Therefore, the M-W-W test may not be appropriate in my scenario above.

An alternative approach to the problem of unequal variances in the case of nonparametric tests has been proposed in the literature by Fligner and Policello (1981). These researchers devised what they refer to as a robust ranks test (Siegel & Castellan, 1988).

In a recent paper, Zumbo and Coulombe (1994) examined the performance of the robust ranks test (also known as the Fligner-Policello test) as a solution to the problem of unequal variance combined with nonnormality by using a skewed family of distributions, the ex-Gaussian. The ex-Gaussian distribution is commonly used as a model for reaction time or response time data. Zumbo and Coulombe concluded that for an ex-Gaussian distribution the robust ranks test performs inconsistently, that is, sometimes it performed quite liberally and other times quite conservatively. These results suggest that their test assumes symmetry of the population of data and does not work for skewed distributions. Interestingly, Siegel and Castellan did not discuss this assumption of symmetry but Fligner and Policello (1981) do mention it in their paper.

Clearly, then, the Fligner-Policello test is not a viable alternative in the research scenario described above.

As a third option, the researcher could use the Welch test. This test does not rely on variance equality. The Welch test was arrived at as a result of statisticians studying the sampling distribution of the *t* statistic when the assumption of equal variances was violated. The problem of heterogeneity of variance has become known as the Behrens-Fisher problem in honour of the two statisticians who were most influential in its investigation. In place of the usual *t* statistic, researchers have examined a modified statistic (*t'*),

$$t' = \frac{\overline{X_1} - \overline{X_2}}{(S_1^2/n_1 + S_2^2/n_2)^{1/2}},$$

which is calculated from variances that are not pooled and which is based on the sampling distribution of *t'* instead of *t*.

Simulation studies have shown that the Welch test is a good solution to the problem of unequal variances when the scores arise from symmetric distributions but is not effective when the problem of unequal variances is combined with skewed data. Therefore, the Welch test may not be appropriate in my scenario above. It

should be noted that the Behrens-Fisher problem was originally formulated only for the case of normal distributions (Scheffé, 1970) and is referred to as the generalized Behrens-Fisher problem in the context of nonnormal distributions.

As a fourth option, Zimmerman and Zumbo (1993a, 1993b) have recently developed and tested a technique that appears to address the generalized Behrens-Fisher problem. This test was motivated from Zimmerman and Zumbo's observation that: (a) a rank transformation of the scores before applying the t test (resulting in a test equivalent to the M-W-W) counteracts the problem of nonnormality for the two-sample case but does not counteract the problem of unequal variances; and (b) the Welch test counteracts the problem of unequal variances but does not counteract the problem of nonnormality. Therefore, if one first applies the rank transformation and then uses the Welch test, both nonnormality and unequal variances may be resolved. Preliminary evidence indicates that this procedure may work for skewed distributions in the positive condition. However, the procedure has not been investigated in the negative condition. Again, the Zimmerman and Zumbo technique, which will be referred to as the Welch test on ranks, may or may not be of use

in the above scenario involving reaction time data.

This thesis will further extend the work of Zimmerman and Zumbo (1993a, 1993b) and investigate the performance of their technique in scenarios like the one posed at the beginning of this section. More specifically, it will address the question: which of the four options--the t test, the Welch test, the t test on ranks (M-W-W), or the Welch test on ranks--could be used when one has unequal variances and reaction time data? To this end, literature on the generalized Behrens-Fisher problem will be reviewed with emphasis on those studies dealing with skewed distributions. Then, a simulation study will be proposed.

CHAPTER 2

Literature Review

The Behrens-Fisher problem is presented in this section followed by a review of empirical literature on the generalized Behrens-Fisher problem with emphasis on those studies that deal with skewed distributions.

The Behrens-Fisher Problem

The Behrens-Fisher problem can be illustrated by comparison with other alternatives. Table 1 lists six possible scenarios. The first column describes the approach to hypothesis testing. The second and third columns describe the null and alternative hypotheses, respectively. Column four describes the assumption regarding the shape of the population distribution of the scores while column five lists the appropriate test statistic(s). One can see that the classical parametric and nonparametric approaches (sections a and b) assume equality of variances for both the null and alternative hypotheses. Furthermore, the Behrens-Fisher problem (in its original or generalized forms, sections c and d) allows for unequal variances in both the null and alternative hypotheses. In contrast to the Behrens-Fisher problem, the Sawilowsky and Blair (1992) approach (sections e and f) assumes equal variances under the null hypothesis but unequal

Table 1

A Comparison of the Generalized Behrens-Fisher Problem with Alternative Approaches to Hypothesis Testing for the 2-group Case.

Approach	Null Hypothesis	Alternative Hypothesis	Distribution	Test
a. Classical Parametric	equal means; equal variance	unequal means; equal variance	Normal	t test
b. Classical Non-parametric	equal means; equal variance	unequal means; equal variance	Nonnormal	Mann-Whitney-Wilcoxon
c. Behrens-Fisher Problem	equal means; unequal variance	unequal means; unequal variance	Normal	Welch test
d. Generalized Behrens-Fisher Problem	equal means; unequal variance	unequal means; unequal variance	Nonnormal	Welch test on ranks or Fligner-Policello test
e. Sawilowsky & Blair Alternative Hypothesis	equal means; equal variance	unequal means; unequal variance	Normal	Modified t test
f. Generalized Sawilowsky & Blair Alternative Hypothesis	equal means; equal variance	unequal means; unequal variance	Nonnormal	Non-parametric modified t test

variances for the alternative. The focus of this thesis is the generalized Behrens-Fisher problem (section d).

Because the effects of unequal variances are central to an understanding of the Behrens-Fisher problem, it is useful to discuss the idea of pooling variances with respect to the t. The idea of pooling sample variances is to get a better estimate of the population variance than would be possible from either one of the samples used in an experiment. This idea requires that sample variances from the two independent groups both estimate the same quantity.

In the classical parametric approach, the idea of pooling the variances of two independent samples makes sense. Under this approach, the use of the t requires that samples come from populations with equal variances, regardless of the truth or falsity of the null hypothesis or whether two sample sizes are equal or unequal. This assumption is often reasonable in an experimental situation because a treatment effect can be regarded as an additive constant. When an experimental treatment is applied, scores are raised or lowered by an amount equal to the effect of the treatment (effect size). Adding or subtracting a constant to or from a set of scores has no effect on

variance and hence, one may assume that the variances would remain unaffected (that is, be assumed to be equal). One can say, then, that when the null hypothesis is true and when the variance is equal, then, in pooling the variances of the sample groups, the *t* ratio is distributed as *t* with $N_1 + N_2 - 2$ degrees of freedom (*df*).

When sample variances are not equal to begin with, it makes no sense to pool them because they are not estimating the same quantity. In this case, the researcher would be required to hypothesize that the variances are unequal under the null hypothesis. Under this condition, the *t* ratio would no longer be valid and thus the use of the t test would be inappropriate for significance testing.

Early in this century, statisticians examined the sampling distribution of the *t* statistic when variances are unequal. In place of the usual *t* statistic, they examined a modified statistic (*t'*) calculated from unequal sample variances that were not pooled. It is not known exactly what the sampling distribution of *t'* looks like; however, researchers have an idea of the approximate distribution of *t'*.

One of the first attempts to find the sampling distribution of *t'* was begun by Behrens and extended by

Fisher (Behrens, 1929; Fisher, 1935). Based on this work, the Behrens-Fisher distribution of t' was derived and was presented in a table in Fisher and Yates (1953). This table covered only a few degrees of freedom and thus was not regarded as especially useful.

A number of researchers, including Cochran and Cox (1957), Satterthwaite (1946), Smith (1936), and Welch (1938, 1947), have proposed different modifications of the t test using approximations. These are known as the Welch-Aspin, the Welch-Satterthwaite or the Smith-Satterthwaite approximations.

The most commonly used solution was developed independently by Welch (1938) and Satterthwaite (1946). It was based on a different approach to the problem and is referred to as the Welch solution or the approximate degrees of freedom test (Algina, Oshima, & Lin, 1994; Howell, 1992; Zimmerman & Zumbo, 1993a, 1993b). This method is included in SAS, BMDP, SPSS, SYSTAT, and MINITAB.

Under the Welch solution, t' is viewed as a legitimate member of the t distribution, but for an unknown number of degrees of freedom. The problem, then, becomes one of solving for the appropriate df, that is, df':

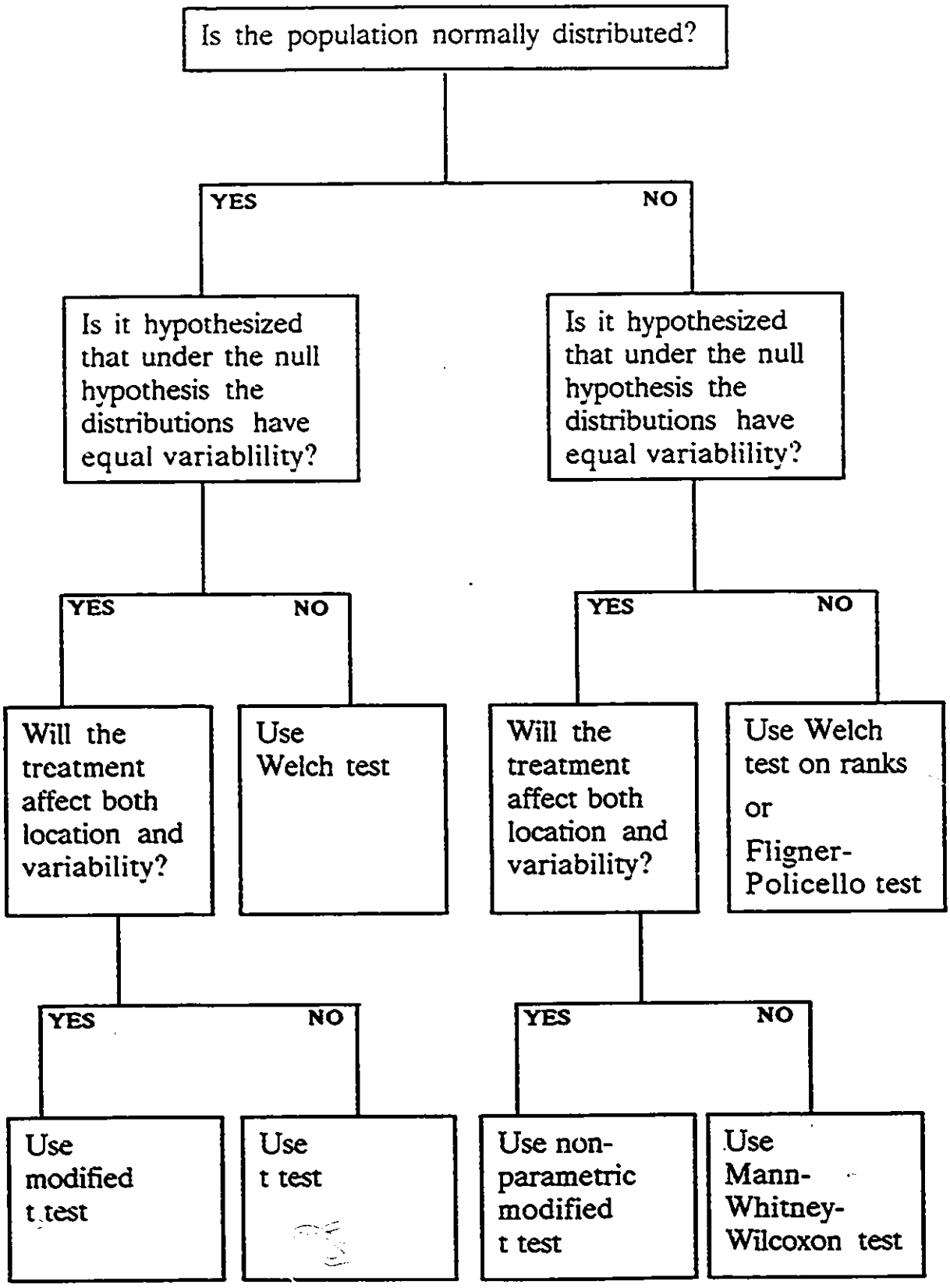
$$df' = \frac{(S_1^2/n_1 + S_2^2/n_2)}{S_1^4/n_1^2(n_1-1) + S_2^4/n_2^2(n_2-1)}.$$

The df' are taken to the nearest integer.

In summary, then, under the null hypothesis, the Behrens-Fisher problem assumes equal unknown means and unequal unknown variances. Under the alternative hypothesis, it assumes differences among means but the assumptions of normality and unequal variances remain as defined under the null hypothesis.

Figure 1 extends the information from Table 1 to provide a framework in a decision tree format. The decision tree illustrates particular combinations of conditions that a researcher may encounter while in the process of seeking an appropriate test that would account for assumptions about the population distribution and the equality of the population variance under a null hypothesis. Although the decision tree is preliminary, it puts into perspective the importance of the availability of a procedure to address the generalized Behrens-Fisher problem. For example, when a researcher encounters conditions such as the one in my scenario described earlier, the availability of a procedure such as the Welch test on

Figure 1. A decision tree for testing hypotheses of location shift in 2-group scenarios.



ranks would be important because it could protect his Type I error rate better than other available tests.

In summary then, the Behrens-Fisher problem is an important statistical issue in the literature (Algina et al., 1994; Mielke & Berry, 1994; Toothaker & Newman, 1994; Zimmerman & Zumbo, 1993a, 1993b; Zumbo & Coulombe, 1994). However, some authors have suggested that while the impacts of treatments may be seen in measures of scale, they always impact the mean as well (Sawilowsky & Blair, 1992). As Algina, Oshima, and Lin (1994) and Zumbo and Coulombe (1994) have argued, this premise is speculative. While there has been no systematic investigation providing evidence for the prevalence of the Behrens-Fisher problem such as that undertaken in Micceri's (1989) study regarding normality, the problem does occur in the real world of researchers (Mielke & Berry, 1994; Zumbo & Coulombe, 1994). In this respect, Toothaker and Newman (1994) have suggested that more research is needed for procedures under the conditions where the equal variance assumption has been violated, especially in combination with nonnormal distributions. At a more general level, Zumbo and Coulombe (1994) have indicated that the Behrens-Fisher problem is important in application because one needs to protect oneself from a

too liberal probability of Type I error or the possible deleterious effect of being too conservative. It should be noted that being conservative is only a problem if statistical power is also reduced.

Monte Carlo Studies

The validity of inferences from statistical tests performed on data that may violate underlying assumptions of those tests is an ongoing concern of quantitative methodologists. The literature associated with this concern is extensive. This thesis will limit its review to empirical literature on the generalized Behrens-Fisher problem with particular attention to skewed distributions.

Thirty-three studies were found which were relevant to this thesis. Of these, twenty-eight formed the population for a recent meta-analytic study conducted by Harwell, Rubinstein, Hayes and Olds (1992) who investigated the robustness of statistical tests for one-factor fixed effects designs and reported Type I error rates of these tests under conditions of nonnormality and unequal variances. Four recent studies were also found (Algina et al., 1994; Zimmerman & Zumbo, 1993a, 1993b; Zumbo & Coulombe, 1994) which address the generalized Behrens-Fisher problem and present empirical results on Type I error rates for

nonnormal distributions, including some skewed distributions. Algina et al. focused specifically on the effects of three tests on Type I error rates for three different skewed distributions whereas Zimmerman and Zumbo (1993a, 1993b) investigated the effects of four tests on seven nonnormal distributions, of which three were skewed. Finally, Zumbo and Coulombe examined the performance of the Fligner-Policello test as a solution to the generalized Behrens-Fisher problem by using ex-Gaussian distributions.

Given the comprehensiveness of the Harwell et al. (1992) study, it will be discussed in detail. It will also be used as a starting point toward constructing a framework within which one may compare the effects on Type I error rates of various tests under the condition of nonnormality and unequal variances. Harwell's et al. study included a literature review that will be summarized in this section.

Harwell et al. used meta-analytic methods to summarize and integrate the findings of a sample of Monte Carlo (MC) studies of the robustness of the F test in one- and two-factor fixed effects ANOVA models. Their study included results for the Kruskal-Wallis (Kruskal & Wallis, 1952) test (the nonparametric test counterpart of F) and the Welch (1947) test. Although

Harwell et al. used a form of the Welch test developed in 1947 by Welch, there is general agreement that it is essentially equivalent to the earlier version (1938) of the test (Algina et al., 1994; Howell, 1992). Only the results of the one-way ANOVA model will be considered in this review.

It should be noted that the use of the F test depends on the assumption of independent and normally distributed scores that share a common variance. The Welch test requires independent and normally distributed scores but does not require equal population variance (Welch, 1938). The Kruskal-Wallis test requires independent scores sharing a common variance but does not require normality (Pratt, 1964). The F and the Kruskal-Wallis tests are relevant for consideration because the two group problem is a special case of these tests (i.e., the t test and M-W-W test, respectively).

Harwell's et al. primary research question was as follows: What data conditions are associated with deviations from nominal Type I error rate and power levels for the F, Welch, and Kruskal-Wallis tests in the single-factor ANOVA model? Only results pertaining to the effects on Type I error rates will be discussed. The population of 28 empirical studies of MC literature

was screened for methodological faults and was derived from a search on the Educational Resources Information Center data base, Dissertation Abstracts International, the Current Index to Statistics, and periodicals such as the Journal of Educational Statistics and Communications in Statistics--Simulation and Computation. Each study was examined for inconsistent or unusual procedures and results by using the following criteria: (a) data generation method (for example, random number generator used); (b) evidence of the success of the data generation (for example, skewness and kurtosis statistics computed for the simulated data); and (c) pattern of Type I error rate and power values when underlying assumptions of the test were satisfied. A list of the studies is provided in Appendix 1.

The results of the study as they related to the effects of the F, Kruskal-Wallis, and Welch tests on the Type I error rates reported by Harwell et al. are listed in Table 2. The first column lists three types of assumption violations: nonnormality (section a); unequal variances (section b); and nonnormality in combination with unequal variances (section c). The second column summarizes findings for each violation under equal sample sizes while the third summarizes

Table 2

Consequences of Violations of Assumptions for Significance Testing

	Equal samples	Unequal samples
Type of violation	Effect on *a*	Effect on *a*
a. Nonnormality	.Negligible effect for F. .Negligible for KW. .Moderate inflation for W (especially skewness).	.Slight inflation for F (skewness more than kurtosis). .Negligible effect for KW. .More substantial inflation for W (skewness more than kurtosis).
b. Unequal variances	.Modest inflation for F that increases with increasing variance ratios. .KW test more erratic. .Modest effect for W up to ratios of 8:1.	.F seriously affected. .KW test somewhat less seriously affected; positive pairings produce conservative *a*'s; negative pairings produces inflated *a*'s. .Slight inflation for W.
c. Nonnormality and unequal variances	.Modest inflation for F. .KW more erratic. .Moderate inflating for W that depends on distribution and variance ratio.	.Negligible effect for F. .Negligible effect for KW. .Moderate inflation for W that depends on distribution and variance ratio.

Note. F = F test; KW = Kruskal-Wallis test; W = Welch test; *a* = Type I error rate.

findings for each violation under unequal sample sizes. Table 2 was adapted from Harwell's et al. (1992) Table 7 that, in turn, was drawn from Table 16 in Glass, Peckham, and Sanders (1972). In their table, Glass et al. reported Type I error effects of the single-factor F test as a function of normality and equal variance assumption violations for equal and unequal sizes. Harwell et al. updated Glass' et al. results and incorporated these with exact statistical theory results (Box, 1954; Gayen, 1949, 1950; Gronow, 1951; Horsnell, 1953; Hsu, cited in Scheffé, 1959; Ramsey, 1980; Scheffé, 1959; Tiku, 1964).

The primary focus of the discussion of Table 2 will be on the effects of the various tests considered by Harwell et al. on Type I error rates under the condition of nonnormality and unequal variances for unequal sample sizes (section c). The effects of these tests under other conditions will be mentioned, however, if they have a bearing on the primary area of interest. Furthermore, relevant results from studies conducted by Algina et al. and Zimmerman and Zumbo (1993a, 1993b) will be also included in the discussion. It should be noted that Algina et al. and Zimmerman and Zumbo focused on two-group tests and therefore they did not examine the F or Kruskal-Wallis tests.

Four important observations emerge from Table 2 regarding the performance of (a) the Welch test, (b) the F test, (c) the Kruskal-Wallis test, and (d) the effects of skewness on Type I error rates. With regard to the Welch test, Harwell reported that this test tended to have inflationary effects under nonnormality (section a, equal and unequal sample sizes). Similar results were found by Algina et al. and Zimmerman and Zumbo (1993a, 1993b) for unequal sample sizes but both found depressed rates when sample sizes were equal.

Harwell et al. reported that the Welch test appeared to counteract error rates when variances were unequal and when distributions were normal (section b, equal and unequal sample sizes). Zimmerman and Zumbo (1993a) found similar results for unequal sample sizes. In contrast, Harwell et al. found that the Welch test tended to have inflationary effects when nonnormality was combined with unequal variances, (section c, equal and unequal samples sizes). Algina's et al. results were consistent with those reported by Harwell et al. under the negative condition when sample sizes were unequal. Under the positive condition, however, both Algina et al. and Zimmerman and Zumbo (1993a; 1993b) reported generally depressed Type I error rates although Algina et al. found slightly inflated effects

when variance ratios increased to 9:1 for smaller sample size ratios. In summary, these results suggest that the Welch test does not protect the Type I error rates under the condition of nonnormality and unequal variances.

As summarized in Table 2 for the F test, Harwell et al. reported slight inflationary effects on the Type I error rates of the F test under nonnormality (section a) when sample sizes were unequal. More serious effects were found when variances were unequal and population distributions were normal (section b). Similar results were found by Zimmerman and Zumbo (1993a, 1993b) for the t test. However, substantial differences were found in Type I error rates when nonnormality was combined with unequal variances for unequal sample sizes. In this case, Harwell et al. reported negligible effects for the F test. However, Algina et al. found substantial inflationary effects for the t test (negative condition) and Zimmerman and Zumbo (1993a, 1993b) found substantial depressive effects (positive condition). This inconsistency brings Harwell's et al. results for the F test into question and requires further investigation.

As was the case with the F test, Harwell et al. also reported negligible effects of the Kruskal-Wallis

test on Type I error rates under nonnormality combined with unequal variance when sample sizes were unequal (section c). Zimmerman and Zumbo (1993a), however, found substantial depressive effects for the Kruskal-Wallis' 2-group counterpart, the t test on ranks (M-W-W), in the positive condition. This inconsistency suggests that Harwell's et al. results require further examination.

The fourth noteworthy observation from Table 2 is the effect of skewness on Type I error rates. As described in Table 2, skewness was reported as a factor contributing to, for example, inflationary effects on the Type I error rate for the F and Welch tests under the condition of nonnormality when sample sizes were unequal (section a) and for the Welch test when sample sizes were equal (section a).

Algina et al. examined three skewed distributions of varying degrees of skewness in which patterns of effects of skewness on Type I error rates could be observed in more detail for the Welch test. As Algina et al. demonstrated, when nonnormality was combined with equal sample sizes in the positive and negative conditions, skewness tended to amplify depressive effects regardless of total sample sizes. However, when sample sizes were unequal, skewness amplified

inflationary or depressive effects. Inflationary or depressive effects depended on sample size ratios and total sample sizes.

When nonnormality was combined with unequal variances, skewness tended to amplify inflationary effects across all distributions, regardless of total sample size or sample size equality in the negative condition. The same pattern was observed in the positive condition for equal sample sizes. However, when sample sizes were unequal in the positive condition, skewness tended to amplify depressive or inflationary effects, depending on variance ratios and sample size ratios.

The pattern of amplified effects across distributions was not observed when (a) large total sample sizes ($N > 80 <= 100$) were combined with the large sample size ratio ($n_2/n_1 = 3$) under nonnormality for unequal sample sizes (negative condition), and (b) when the larger sample size ratio ($n_2/n_1 = 3$) was combined with higher variance ratios (9:1) under nonnormality combined with unequal variances and unequal sample sizes (positive condition). However, in both cases, the most skewed distribution was more inflated than the least skewed distribution.

These observations suggest that (a) skewness may

amplify inflationary or depressive effects, and (b) these inflationary or depressive effects may be affected by variance and sample size ratios, and to a lesser extent, total sample sizes. Hence, skewness should be examined as a factor that may influence the effects of Type I error rates.

In summary, current evidence in the literature suggests the following: (a) that the Welch test is ineffective in controlling the Type I error rate under the condition of nonnormality combined with unequal variances with unequal sample sizes; (b) that the F and Kruskal Wallis tests might be effective in controlling Type I error rates under the same conditions but conflicting results with the t test and the t test on ranks (M-W-W) suggest that these tests need to be further investigated; and (c) that skewness appears to amplify Type I error rates and should be examined further.

These results must be viewed with some caution with respect to skewed distributions. First, there is the problem of the number of, and types of, skewed distributions observed in the studies and the wide range of total sample sizes upon which these observations were based. Altogether, the four studies dealt with only four skewed distributions (the

exponential, the lognormal, beta, and the half normal) and examined effects of various tests on total sample sizes which ranged from very small (N = 20) to very large (N = 700).

In Harwell's et al. study, only 6.8% of total distributions observed were skewed and all were exponential. Another 16.7% were identified as nonnormal distributions but their skewness was not identified. It is not known, therefore, if conclusions about effects of skewness on Type I error rates were attributed solely to exponential distributions or included others. Furthermore, observations of the effects on the Welch test were drawn from comparing results of three distributions--the normal, t (nonnormal but not skewed), and the exponential (skewed)--and were based on 215, 15, and 15 cases, respectively. Thus, it appears that conclusions about the effects of the Welch test on Type I error rates were based on a small number of cases of only one skewed distribution.

In contrast, Algina et al. focused their study on three skewed distributions that were selected to capture various degrees of skewness: the Beta distribution (least skewed), the exponential (moderately skewed), and the lognormal (most skewed).

However, their total sample sizes varied dramatically in the positive (for example, N = 20 and 40 for Beta; N = 40, 100, 160 for lognormal) and negative conditions (N = 20, 40, 60, 80 for Beta; N = 100, 200, through to 700 for lognormal). Clearly, their results would apply more appropriately for large sample studies. On the other hand, Zimmerman and Zumbo (1993a, 1993b) examined a total sample size of 24 and observed seven nonnormal distributions, of which only three were skewed: the exponential, the lognormal and the half normal. One could only suggest that, given this diversity of sample size on so few skewed distributions, any conclusions regarding the effects of skewness on Type I error rates would be preliminary.

Second, as Harwell et al. stated, there is the problem of the lack of a systematic examination of the Welch test across a range of nonnormal distributions, including skewed distributions. A thorough investigation of the Welch test under conditions of nonnormality and unequal variance is of particular importance because is does not rely on the assumption of equal variance. Although all studies appeared to support the premise that the Welch test is ineffective in controlling the Type I error rate, the situation remains problematic for skewed distributions. For

example, Harwell et al. examined the performance of the Welch test in a few cases while Algina et al. focused on effects on only three examples of skewed distributions. Zimmerman and Zumbo (1993a, 1993b) investigated the Welch test's effects across seven nonnormal distributions but their observations were limited to one total sample size of 24 examined under the positive condition only.

Similarly, there is a lack of a systematic investigation of the effects on Type I error rates on skewed distributions of other tests that were examined by these authors. Harwell et al. examined the F and the Kruskal-Wallis tests under the negative and positive conditions. However, as mentioned earlier, their results are in conflict with those found by Algina et al. and Zimmerman and Zumbo for the t test and the t test on ranks (M-W-W). This conflict, however, is based on limited evidence: Algina's et al. observations for the t test were limited to only one distribution--the lognormal--and only in the negative condition; Zimmerman and Zumbo's (1993a, 1993b) observations were based on the performance of the t test and the t test on ranks (M-W-W) across seven distributions but were limited, as noted earlier, to one sample size in the positive condition only.

Algina et al. also examined the James second-order test (a generalized series solution of Welch's 1947 series test) under positive and negative conditions. They found that the performance of the James second-order test was similar to the Welch test and recommended that other tests be explored. Zimmerman and Zumbo (1993a, 1993b) also investigated one additional test, namely the Welch test on ranks. Their preliminary examination of the Welch test on ranks suggested that it might correct for Type I error rates when nonnormality is combined with unequal variances and unequal sample sizes. Their finding, however, has yet to be examined systematically across skewed distributions.

In summary, empirical results concur that the Welch test and the F and t tests, as well as their nonparametric counterparts, the Kruskal-Wallis and the t test on ranks (M-W-W), may be ineffective in controlling Type I error rates in the context of skewed distributions. Evidence suggests that the Welch test on ranks may be effective in correcting error rates. However, these conclusions are, at best, preliminary as they are based on studies in which varieties of tests were used under varieties of distributions, sample sizes and conditions, all of which cannot be

systematically compared with one another.

Finally, there is also the problem of justification of a choice of distributions studied for effects on Type I error rates. With the exception of a study conducted by Zumbo and Coulombe (1994) who examined the ex-Gaussian distribution, none of the authors justified their choices of distributions as those that would typically occur in a research context. Indeed, their choices seemed without guidance and more related to convention in the academic field and convenience rather than to the study of distributions encountered in educational research. That is, there was no evidence of attempts to conduct a systematic study with nonnormal distributions, especially those that are skewed, that is representative of ones that are encountered in the socio-behavioural sciences.

To this end, this study will do a systematic investigation of the ex-Gaussian distribution. The ex-Gaussian is often used as a model for reaction time data. Reaction time is a commonly used dependent variable in the socio-behavioural sciences. This thesis, then, will address the following research question: For a skewed reaction time distribution (ex-Gaussian), what is the effect of unequal variances and unequal sample sizes (both positive and negative

conditions) on the robustness[1] of Type I error rate of the following tests:

a) t test;

b) t test on ranks (equivalent to the Mann-Whitney-Wilcoxon test);

c) Welch test; and

d) Welch test on ranks?

---

[1]When discussing Type I error rates, a significance test is considered robust if the probability of Type I error is close to the nominal value (e.g., .05) even though the assumptions of the procedure are violated. This will be operationalized in the Methodology section.

CHAPTER 3

Methodology

The methodology for the study conducted in this thesis is presented in this section. The discussion describes the factors considered in the design of the study, outlines the computer simulation methodology, and identifies the performance index that will be used in this thesis.

The Monte Carlo procedure used in this study is designed to empirically estimate Type I error rates by applying the t test, the t test on ranks, the Welch test, and the Welch test on ranks to an ex-Gaussianly distributed parent population. The ex-Gaussian family of distributions has played an important role in reaction time modelling (Luce, 1986) and was considered in detail by Hohle (1965), who reported good fits of the ex-Gaussian to empirical reaction time distributions of individual shapes. The ex-Gaussian has often been used to describe reaction time distributions (see for example, Heathcote, Popiel, & Mewhort, 1991; Hockley, 1984, Miller, 1988; Ratcliff, 1978; Ratcliff & Murdock, 1976). All the tests will be of non-directional hypotheses at an alpha level of .05.

Factors in the Study

Four factors will be considered in the simulation

as follows: (a) total sample size; (b) ratio of sample sizes; (c) ratio of variances; and (d) negative and positive conditions. The total sample sizes will consist of 24 and 72. The smaller total sample size was chosen so that our results would be comparable to Zimmerman and Zumbo's studies (1993a, 1993b) while the larger total sample size was chosen to test whether the results are generalizable to larger sample sizes. The ratios of sample sizes will be 1:1, 1:2, and 1:3. These ratios are the same as those found in Algina, Oshima, and Lin (1994) except that they used 1:1.857 where we will be using 1:2. The ratio of variances will follow Algina et al. and will be 1:1, 1:4, and 1:9. We will also include a variance ratio of 1:2 which Algina et al. did not use. The rationale to include this variance ratio is that there is a substantial leap from a variance ratio of 1:1 and 1:4 and one should investigate the variance ratio between these two values. Furthermore, two conditions will also be simulated as follows: one where the larger variance will be associated with the larger sample size (positive condition), and the other where the larger variance will be associated with the smaller sample size (negative condition).

Computer Simulation Methodology

The methodology which will be used in this study will be similar to that used by Zimmerman and Zumbo (1993a, 1993b) and Zumbo and Coulombe (1994). The ex-Gaussian distribution is defined as the sum of two stochastically independent random variables: (a) a normally distributed (that is, Gaussian) component with mean, MN, and standard deviation, SN; and (b) an exponentially distributed component with mean, ME. The normal deviates will be generated using the method of Box and Muller (1958). The exponential will be generated by $X = -\log_e(X_i) - 1$, where X_i is a uniform pseudorandom number on the interval [0,1]. Where necessary, the mean and variances of these distributions will be altered to the specified values by simple arithmetic transformations. The number of replications will be 5000.

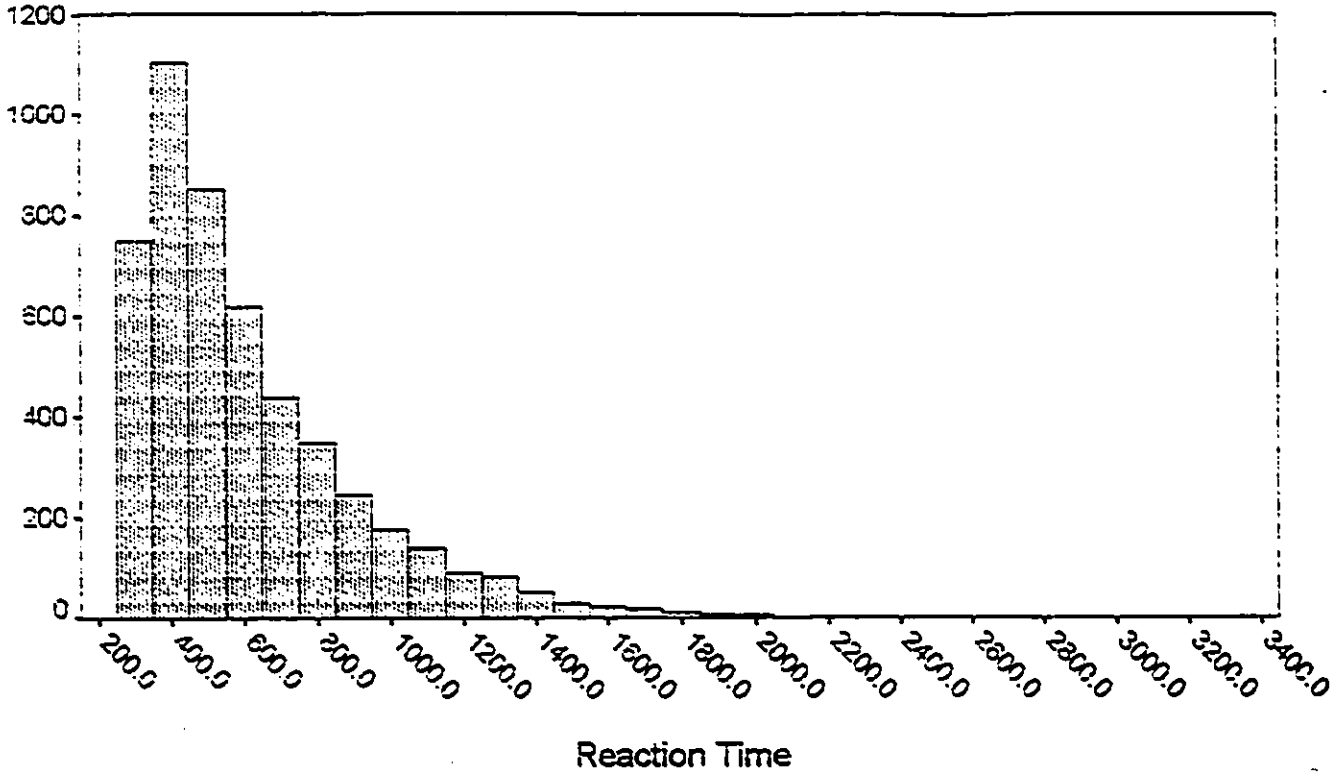
The three parameters MN, SN, and ME specify the shape of the ex-Gaussian allowing for a great deal of flexibility in shape. Miller (1988) listed twelve combinations of parameter values representative of those found in experiments using reaction time data as a dependent variable. As Miller states, the twelve reaction time distributions were defined by selecting the mean and standard deviation of the normal

distribution and mean of the exponential distribution from values reported in studies by Hockley (1984) and Ratcliff and Murdock (1976). Miller's distributions ranged from fairly symmetric to quite skewed. As in Zumbo and Coulombe (1994), Miller's most skewed distribution will be selected to represent a boundary condition or extreme shape of reaction time data. That is, this distribution is an extreme but, according to Miller (1988), a realistic distribution. By using the expression "boundary condition" I mean to highlight that my choice of distribution is not only extreme but also is a starting point for investigating the four statistical tests. By examining this extreme amount of skewness, it is implied that any of the four tests that perform adequately in this condition will perform at least as well in less skewed conditions.

The adequacy of the random number generator which will be used in the simulations has been evaluated using tests recommended by Lehman (1977) and Morgan (1984). The generator met accepted requirements of rectangularity, sequential independence, and lack of patterns in a sequence of numbers.

As a check for the accuracy of the proposed simulation methodology in this study, Figure 2 contains a histogram of 5,000 scores generated from the reaction

Figure 2. Histrogram of the ex-Gaussian.



time (in milliseconds) distribution which will be used in this study. As is illustrated in Figure 2, it is clear that the reaction time distribution is quite skewed. The mean of the distribution was 613.674 while the median was 521.729. For this particular ex-Gaussian distribution, Miller noted a mean-median difference of 92; the present data indicates a difference of 91.945. Furthermore, as expected, the scores observed in Figure 2 ranged from 249 to 3366, the variance was 93716.957 (standard deviation of 306.132), the skewness index reported by SPSS was 1.810 (standard error of skewness of .035), and a kurtosis of 4.976 (standard error of kurtosis of .069).

As a further check for the accuracy of the proposed simulation methodology, I also ran the four tests which will be used in this study and found that they were computed correctly. The use of my program was also investigated with the normal distribution. I found that the Type I error rates of the t test were as expected. That is, the Type I error rate was .05 when sample sizes and variances were equal. Under the positive condition, the Type I error rate was depressed and, under the negative condition, it was inflated.

Performance Index

Bradley (1978) discussed what it meant for a test

to be robust and concluded that a quantitative definition of robustness with regard to the Type I error rate could be achieved by stating, for a given level, the range of empirical levels for which the test would be considered robust. The present study will use Bradley's moderate criterion for robustness which is defined as the situation wherein the absolute value of the empirical Type I error rate minus .05 is less than or equal to .05 divided by 5 (i.e., the moderate criterion would require the empirical Type I error rate to lie between .04 to .06).

This performance index will be used to identify conditions where the Type I error rate is liberal (that is, greater than .06) and conditions where the Type I error rate is conservative (that is, less than .04).

CHAPTER 4

Results

The results of the computer simulations are presented in this chapter. The Type I error rates of each test--that is, the t test, the Welch test, the t test on ranks, and the Welch test on ranks--are reported in separate tables. Each table, then, will be treated as a graphical display (Moore & McCabe, 1993) of the relationship between the four factors in the study and the robustness, as defined in the Methodology Section by Bradley's criterion. As suggested by Moore and McCabe, the analysis of graphical displays should always follow the following steps: (a) look first for an overall pattern; and (b) then look for meaningful deviations from that pattern, such as particular cells in the design that do not adhere to the pattern of robustness.

By the nature of the problem of unequal variances, the study is an incomplete factorial design. That is, there are not negative or positive conditions for the cases of equal variances or equal sample sizes. Therefore, to facilitate presentation of the results, the results will be presented in three stages for each test. First, the equal sample sizes condition will be considered across the four levels of the equal and

unequal variances. Second, the equal variances condition will be considered across the three levels of equal and unequal sample sizes. Finally, elements of the table that represent the complete five-way factorial design, henceforth referred to as the complete factorial portion, will be considered. The complete factorial portion of the table would result in a 3 (variance ratio) x 2 (sample size ratio) x 2 (total sample size) x 2 (negative and positive pairings) design. Therefore, the analyses of the simulation results will proceed as follows: (1) for stages one and two, the analyses will rely solely on Bradley's criterion because the design is relatively simple (i.e., there are only two factors); (2) for the complete factorial portion, Bradley's criterion will be used to identify patterns (i.e., main effects) and then loglinear modelling will be used to ascertain whether interactions (i.e., deviations from the pattern) are meaningful. Loglinear modelling will be used in the complete factorial portion because of the complexity of the design.

The t Test

Table 3 presents the results for the t test. The columns describe the sample size ratio factor (n_1:n_2) wherein the first, second, and third columns represent

Table 3

Type I Error Rate of the t Test

$n_1:n_2$

var1:var2	1:1		1:2		1:3	
	N = 24	N = 72	N = 24	N = 72	N = 24	N = 72
1:1	.044	.048	.045	.054	.046	.048
1:2	.053	.047	neg. .075+ pos. .031−	neg. .079+ pos. .026−	neg. .088+ pos. .028−	neg. .094+ pos. .022−
1:4	.072+	.055	neg. .127+ pos. .033−	neg. .122+ pos. .023−	neg. .163+ pos. .024−	neg. .160+ pos. .018−
1:9	.081+	.060	neg. .178+ pos. .037−	neg. .152+ pos. .021−	neg. .244+ pos. .021−	neg. .211+ pos. .011−

Note. $n_1:n_2$ = sample size ratios; var1:var2 = population variance ratios. Type I error rate = .05.

the sample size ratios of 1:1, 1:2, and 1:3, respectively. The four rows along the far left-hand side list the variance ratio factor (var1:var2) as follows: 1:1, 1:2, 1:4, and 1:9 from top to bottom. The total sample size factor is represented within each cell such that one may compare Type I error rates for N = 24 and N = 72 for a given cell. Finally, the positive and negative pairings factor is displayed within the columns that represent unequal sample sizes and the rows that represent unequal variances. The negative conditions are represented by dividing the cell in two by a dashed line. Tables 4, 5, and 6 have the same layout as Table 3.

For the cases of equal sample sizes (the far left row), there is a total sample size by variance ratio interaction wherein for the variance ratios of 1:4 and 1:9 the smaller sample size resulted in an inflated Type I error rate and the large sample size did not. For the case of equal variances, the t test is robust for all of the conditions in our study. Finally, focussing on the complete factorial portion of the table, one can see that the negative pairing condition resulted in an inflation of the Type I error rate while the positive pairing condition resulted in a deflation of the Type I error rate. The pattern involving the

positive and negative pairing holds for the various levels of the variance ratio, sample size ratio, and total sample size factors in the complete factorial portion of the table.

The Welch Test

Table 4 presents the results of the Welch test. Focussing on the equal sample size condition, like the t test there is a sample size by variance ratio interaction wherein for the smaller sample sizes and larger variance ratios (1:4 and 1:9) the Type I error rate is inflated. Next, for the equal variance case, there is a sample size ratio by total sample size interaction. That is, the Welch test is robust for all cases considered except for the smaller sample size and the largest sample size ratio (1:3) wherein it is inflated.

Finally, in the complete factorial portion of the table, the Welch test is robust for positive pairing but liberal for the negative pairing. This pattern holds for all except two of the 24 Type I error rates reported in that portion of the table. The two exceptions are as follows: (1) for the case of a variance ratio and sample size ratio of 1:2, in the negative pairing condition for a total sample size of 72, wherein the Type I error rate is robust instead of

Table 4

Type I Error Rate of the Welch Test

var1:var2	$n_1:n_2$ = 1:1		$n_1:n_2$ = 1:2		$n_1:n_2$ = 1:3	
	N = 24	N = 72	N = 24	N = 72	N = 24	N = 72
1:1	.040	.048	.051	.060	.071+	.058
1:2	.049	.046	neg. .068+ pos. .041	neg. .057 pos. .043	neg. .093+ pos. .046	neg. .074+ pos. .052
1:4	.069+	.054	neg. .097+ pos. .042	neg. .071+ pos. .053	neg. .108+ pos. .043	neg. .079+ pos. .049
1:9	.077+	.059	neg. .101+ pos. .062+	neg. .073+ pos. .053	neg. .112+ pos. .049	neg. .082+ pos. .052

Note. $n_1:n_2$ = sample size ratios; var1:var2 = population variance ratios. Type I error rate = .05.

the expected inflated finding; and (2) for the case of variance ratio of 1:9 and a sample size ratio of 1:2, in the positive pairing condition for a total sample size of 24, wherein the Type I error rate is inflated instead of the expected robust finding.

The deviant results could be explained by the presence of a two-way interaction of positive or negative pairing by total sample size that would be conditional upon the level of variance ratio and upon the level of sample size ratio. In essence, the deviations suggest a four-way interaction. A loglinear model was fit to the complete factorial portion in Table 4 to investigate the possibility of a four-way interaction. A hierarchical loglinear model was fit with the program BMDP 4F and the resulting test of the hypothesis that the four-way interaction is zero resulted in a statistically nonsignificant result, likelihood ratio chi-square with two degrees of freedom equals 1.77, $p = 0.412$. Therefore, the predicted interaction was not found to hold in the analysis of the data.

The statistically nonsignificant interaction is interpreted to suggest that the deviant cells are not deviant enough statistically to nullify the general pattern that negative pairing resulting in an inflation

in Type I error rate while the Welch test is robust in the case of positive pairing.

The t Test on Ranks (Mann-Whitney-Wilcoxon)

Table 5 presents the results of the t test on ranks. Focussing on the equal variance case, the t test on ranks is robust in all cases. However, the equal sample sizes case shows that the t test on ranks is decidedly liberal for variance ratios of 1:2, 1:4, or 1:9.

Finally, in the complete factorial portion of the table, the Type I error rate of the t test on ranks is inflated for all cases except the following three positive pairings and total sample size of 24: (1) a variance ratio and sample size ratio of 1:2 where the test is robust; (2) a variance ratio of 1:2, sample size ratio of 1:3 wherein the Type I error rat[illegible] deflated; and (3) a variance ratio of 1:4, sample size ratio of 1:3 wherein the test is robust. As in the Welch test, these deviations would imply that the pairing by total sample interaction would be conditional upon the variance ratio and the sample size ratio; hence, a four-way interaction. A hierarchical loglinear model was fit to the complete factorial portion in Table 5 and the resulting test of the hypothesis that the four-way interaction is zero

Table 5

Type I Error Rate of the t Test on Ranks (Mann-Whitney-Wilcoxon)

var1:var2	$n_1:n_2$ = 1:1, N = 24	$n_1:n_2$ = 1:1, N = 72	$n_1:n_2$ = 1:2, N = 24	$n_1:n_2$ = 1:2, N = 72	$n_1:n_2$ = 1:3, N = 24	$n_1:n_2$ = 1:3, N = 72
1:1	.056	.047	.055	.055	.050	.049
1:2	.090+	.150+	neg. .108+ / pos. .055	neg. .170+ / pos. .114+	neg. .107+ / pos. .038−	neg. .174+ / pos. .090+
1:4	.151+	.297+	neg. .180+ / pos. .090+	neg. .293+ / pos. .244+	neg. .177+ / pos. .052	neg. .270+ / pos. .192+
1:9	.186+	.393+	neg. .221+ / pos. .134+	neg. .379+ / pos. .347+	neg. .213+ / pos. .070+	neg. .358+ / pos. .258+

Note. $n_1:n_2$ = sample size ratios; var1:var2 = population variance ratios. Type I error rate = .05.

resulted in a statistically nonsignificant result, likelihood ratio chi-square with two degrees of freedom equals 3.11, $p = 0.212$. And, again, the predicted interaction was not found to hold in the analysis of the data.

The statistically nonsignificant interaction is interpreted to suggest that the deviant cells are not deviant enough statistically to nullify the general pattern that the Type I error rate of the t test on ranks was inflated for all cases of unequal variances.

The Welch Test on Ranks

Table 6 presents the results of the Welch test on ranks. Of the four tests, these are the most straightforward results. In the case of equal variances, there is an interaction of total sample size by sample size ratio wherein the test is robust in all cases except for the smaller total sample size for a sample size ratio of 1:3 wherein the test is liberal. Furthermore, the test is decidedly liberal for all cases of unequal variances under investigation.

Table 6

Type I Error Rate of the Welch Test on Ranks

var1:var2	$n_1:n_2$ 1:1		1:2		1:3	
	N = 24	N = 72	N = 24	N = 72	N = 24	N = 72
1:1	.056	.047	.054	.057	.064+	.052
1:2	.090+	.150+	neg. .078+ pos. .080+	neg. .130+ pos. .153+	neg. .077+ pos. .081+	neg. .109+ pos. .151+
1:4	.147+	.297+	neg. .119+ pos. .145+	neg. .218+ pos. .342+	neg. .100+ pos. .142+	neg. .167+ pos. .355+
1:9	.180+	.391+	neg. .132+ pos. .215+	neg. .275+ pos. .488+	neg. .106+ pos. .218+	neg. .211+ pos. .509+

Note. $n_1:n_2$ = Sample size ratios; var1:var2 = population variance ratios. Type I error rate = .05.

CHAPTER 5

Discussion

The utility of the t test, the Welch test, the t test on ranks (Mann-Whitney-Wilcoxon) and the Welch test on ranks for significance testing purposes as applied to reaction time data (ex-Gaussian distribution) is discussed in this section. In addition, the results of the Monte Carlo simulations are related to earlier findings regarding the robustness of these tests. Finally, possible future applications of the tests are examined.

The focus of this thesis was to extend the work of Zimmerman and Zumbo (1993a, 1993b) and conduct a systematic investigation on reaction time data of the effect of unequal variances and unequal sample sizes on the robustness of the Type I error rate of the four tests under study. To this end, the present study compared the robustness of the four tests under the conditions of unequal sample sizes and unequal variances in the positive and the negative conditions. Two additional conditions were also compared to facilitate a more complete comparison into the performance of these tests, that is, the equal variance condition across equal and unequal sample sizes and the

unequal variance conditions across equal sample sizes. Although several conclusions can be drawn from the comparison conducted in this thesis, it should be remembered that the results are restricted to the specific simulations that were performed. No attempt was made, therefore, to generalize beyond these specific conditions.

It should further be noted that the results were based in part on nonsignificant findings with respect to the loglinear analyses. Statistical power, therefore, will need to be further addressed. Given that 5000 replications were conducted per cell, however, there should be enough power to detect a four-way interaction if it were present in the data.

It should be noted that the ex-Gaussian distribution was selected for the Monte Carlo simulation because it is encountered in educational and behavioural research settings. Reaction time is commonly used as a dependent variable in studies in social and behavioural sciences (see for example, Heathcote, Popiel, & Mewhort, 1991; Hockley, 1984) wherein the ex-Gaussian distribution is often used as a model for reaction time data (Luce, 1986) and has been studied in detail (Hohle, 1965). Hence, unlike the skewed distributions typically studied because they are

convenient to generate (Algina et al, 1994; Zimmerman & Zumbo, 1993a, 1993b), the ex-Gaussian distribution arises from real data. In this thesis, Miller's (1988) most skewed distribution was selected for the Monte Carlo simulation because it represents a boundary condition of reaction time data. This distribution was among the 12 combinations of parameter values listed by Miller (1988) which were representative of those found in empirical studies using reaction time as a dependent variable (for example, see Hockley, 1984; Radcliff & Murdock, 1976) and, thus, it does occur in research settings. As a boundary condition, this distribution provides a framework wherein effects on Type I error rates of the four tests under study could be observed under conditions of extreme skewness.

The criteria for the utility of a test will be guided by the following considerations: (a) that a liberal test is not usable (that is, it exceeds Bradley's (1978) criteria of a Type I error rate of .06); (b) that a conservative test is usable (that is, the Type I error rate is below .04 and power is not reduced); and (c) that a robust test is usable (that is, the Type I error rate falls within .04 and .06). It should be noted that this thesis did not examine power. Thus, further studies of power will be required

to validate claims of a test's usefulness for significance testing purposes.

Given that Bradley's criteria was used in this thesis to judge robustness of Type I error rates, Algina's et al. (1994) and Zimmerman and Zumbo's (1993a) data were re-examined wherein Bradley's criteria for robustness was applied to the Type I error rates for the six skewed distributions that they studied. This procedure was included to (a) establish common ground upon which one might determine expected general trends of Type I error rates arising from previous literature for various forms of skewed distributions, and (b) examine if there are any similarities between these general patterns of Type I error rates with reaction time data. The procedure is not meant to facilitate a direct comparison between skewed distributions and reaction time data. Given that substantial disparities between the skewness of distributions, sample sizes, sample size ratios, and variance ratios exist in the data reported in previous literature, a true comparative analysis was not possible. Data from previous literature is presented in Appendix 2.

In a research context, a researcher may typically consider the use of the t test first. For this reason,

this discussion section will initially focus on the performance of the t test, then it will compare the remaining three tests individually to the t test in order to draw conclusions about the merits of each test.

The t Test

The use of the t test requires that scores arise from populations that are normally distributed and have equal variances. In this study, however, the t test was subjected to scores arising from extremely skewed distributions and to unequal variances ranging from moderate (var1:var2 = 1:2) to extreme (var1:var2 = 1:9). In conditions where sample sizes were unequal and variances were unequal, the results indicated a pattern of inflated Type I error rates across all levels of variance inequality in the negative condition and depressed Type I error rates in the positive condition. This result did not support Harwell's et al. (1992) expected findings of negligible effects on Type I error rates (for the F test) under these conditions for nonnormal distributions. However, it does support Algina's et al. (1994) findings of substantial inflation of Type I error rates for the t test in the negative condition and Zimmerman and Zumbo's (1993a) reported results of depressed Type I

error rates in the positive condition for skewed distributions. This finding suggests that the use of the t test with reaction time data would be inappropriate in the negative condition but could be used in the positive condition, even under extreme variance inequality. Further study, however, will be required to investigate if power is also reduced under these conditions.

The results also indicated that for extreme reaction time data, the t test was robust and could be used even under extreme variance inequalities (up to var1:var2 = 1:9) for larger equal sample sizes, and under more moderate violations (var1:var2 = 1:2) for smaller equal sample sizes. This pattern did not support Algina's et al. finding for a different class of a skewed distribution, the lognormal, wherein inflated Type I error rates were observed for larger variance ratios (var1:var2 = 1:4 and 1:9) for larger equal sample sizes (N = 100). For the unequal sample sizes case where variance was equal across sample size ratios, the results also indicated that the t test for reaction time data remained robust and could be used even under large group size disparity (up to n_1:n_2 = 1:3). A similar trend of robustness was observed in Algina's et al. study of the lognormal distribution.

These findings suggest that the t test may remain effective for reaction time data even under extreme violations of normality and variance equality assumptions provided that sample sizes are equal. Furthermore, the t test could be used under extreme violations of normality when the disparity in sample sizes between groups is large.

These results have two important implications in a research context: (1) the general agreement in the literature that the t test can be used under "moderate violations" of the assumption of homogeneity of variance when sample sizes are equal may be too conservative; and (2) given that researchers may unwittingly tend to use the t test when assumptions of normality and/or variance equality are violated, the use of the t test might be appropriate under these conditions for reaction time data when sample sizes are equal and large. However, further study will be required to determine the levels of variance inequality and sample size ratios at which the t test remains within robust levels.

The t Test and the Welch Test

The patterns of the performance of the Welch test was similar to that of the t test. The Welch test does not rely on variance equality but, like the t test, it

does require that scores arise from symmetric distributions. For reaction time distributions, the results showed that the Welch test did not improve the Type I error rates of the t test when small, equal sample sizes were combined with increasing variance ratios. Furthermore, under equal variance conditions, however, the Welch test did not protect the Type I error rate as well as the t test did for large group size disparity when sample sizes were small. In this case, the t test would be the preferred option. In contrast to the t test, however, the Welch test corrected Type I error rates to robust levels in the positive condition but was not effective in reducing inflated results in the negative condition. The expected findings from Harwell's et al. (1992) study of moderate inflation that depended on distribution and variance ratio were not supported in the negative condition for this test. The overall results suggest that the Welch test may be a better option in the positive condition for reaction time distributions, but its performance across variance inequalities did not improve upon the t test nor did it match the t test's performance with respect to large group size disparity when sample sizes were small. In this case, the t test may be the preferred option provided that power levels

in the positive condition are not reduced.

The general patterns of Type I error rates of the Welch test with reaction time data appeared similar to those of the three skewed distributions that were examined by Algina et al. (1994) in their comparison of the Welch test under three levels of skewness--the Beta distribution (least skewed), the exponential (moderately skewed), and the lognormal (most skewed). For example, when equal sample sizes were combined with unequal variance ratios in Algina's et al. study, the Welch test was increasingly ineffective in protecting Type I error rates for smaller sample sizes as variance ratios and degrees of skewness increased. Furthermore, when unequal sample sizes were combined with equal variance in their study, the Welch test became less effective in controlling inflated Type I error rates when disparity in sample sizes between groups was large, especially for smaller sample sizes, as levels of skewness increased. Finally, when unequal sample sizes and unequal variances were combined, the Welch test did not correct for inflated Type I error rates in the negative condition. In the positive condition, the Welch test was robust for both the Beta and exponential distributions, and generally robust for the lognormal up to variance ratios of 1:4. These findings suggest

that further study of the Welch test with skewed distributions, including reaction time distributions, might not find any notable overall improvements over the use of the t test.

The t Test on Ranks (M-W-W) and the t Test

The results also indicated that applying the t test on ranks (M-W-W), that is, transforming scores into ranks and then applying the t test, did not appear to influence the robustness of the t test when variances were equal. In this case, the use of the t test or the t test on ranks may be appropriate but further study will be needed to determine if power of the t test is improved by a rank transformation under this condition.

Under conditions of variance inequality across equal and unequal sample sizes, the use of the t test on ranks was not a viable option for reaction time distributions. Transforming scores into ranks inflated Type I error rates even under "moderate violations" of the homogeneity of variance assumption across all equal samples, regardless of their size. That is, the t test became more sensitive to variance inequality when scores were changed into ranks.

Similarly, the rank transformation changed the conservative error rates associated with the t test in

the positive condition into a liberal test. Thus, under conditions where the homogeneity of variance assumption was violated, the t test would remain the preferred option. This finding suggests that the conventional view in the literature that calls for the use of nonparametric tests when there are violations of assumptions may not apply with reaction time data. It is notable that this finding was not supported for other forms of skewed distributions studied by Zimmerman and Zumbo (1993a) in the positive condition. They found that the rank transformation of the t test for the exponential and the half normal distributions did not affect conservative Type I error rates in this condition. Hence, the t test on ranks might be a viable option if power levels are not reduced for these forms of skewed distributions.

The Welch Test on Ranks and the t Test

Like the t test on ranks, the results indicated that the Welch test on ranks is not a viable option under the condition of variance inequality across all equal and unequal sample sizes. Like the t test on ranks, transforming scores into ranks inflated Type I error rates under even "moderate violations" of variance inequality for all equal sample sizes, regardless of their size. This finding suggests that

the rank transformation, as opposed to the Welch correction, may be the underlying factor in sensitizing the Welch test to variance inequality.

Similarly, the rank transformation changed robust Type I error rates associated with the Welch test into a liberal test in the positive condition. This finding further suggests that, like the t test on ranks, the rank transformation may be an underlying factor in sensitizing the Welch test to variance inequality in the positive condition.

In conditions where the variance was equal across sample sizes, however, the rank transformation did not change the performance of the Welch test when large group disparity was associated with small sample sizes. This finding suggests that the Welch correction may not be affected by a rank transformation in this condition.

This thesis investigated the effect of unequal variances and unequal sample sizes on the robustness of the Type I error rate of the t test, the Welch test, the t test on ranks (M-W-W) and the Welch test on ranks on reaction time data. The findings suggest that the t test would be the preferred overall option for significance testing purposes for reaction time data. Ironically, this finding is counter to trends in the literature which suggest the use of nonparametric

methods for significance testing purposes for nonnormal (and skewed) distributions (Harwell et al., 1992; Zimmerman and Zumbo, 1993a). However, in this study, it was shown that, in fact, a parametric test--in this case, the t test--may be a more suitable option for significance testing purposes for an extreme case of a skewed distribution (ex-Gaussian). Given that the t test is akin to the standard for significance purpose testing in educational research and is often considered as a first option, this finding has attractive implications for applied research. Researchers who are working with reaction time data may be able to use the t test even if their data is subject to the kind of extreme violations to normality and variance inequality which was evident in the scenario presented at the beginning of this thesis.

The suitability of the use of the t test, however, would only apply in the positive condition. It was clear that all the tests considered in this thesis, including the t test, were not viable options for significance testing purposes in the negative condition for reaction time data. In all cases, all tests demonstrated an inflated Type I error rate. Thus, the problem posed in our research scenario--which emphasized the negative condition situation--has yet to

be resolved. Hence, further study will be necessary to find an appropriate method to deal with reaction time data in the negative condition.

REFERENCES

Algina, J., Oshima, T. C., & Lin, W. (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, 274-291.

Behrens, W. U. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. Landwirtschaftliches Jahrbuch, 68, 807-837.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems I: Effects of inequality of variance in the one-way classification. Annals of Mathematical Statistics, 25, 290-302.

Box, G. E. P., & Muller, M. (1958). A note on the generation of random normal deviates. Annals of Mathematical Statistics, 29, 610-611.

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

Cochran, W. G., & Cox, G. M. (1957). Experimental designs (2nd ed.). New York: Wiley.

Fisher, R. A. (1935). The design of experiments. Edinburgh: Oliver & Boyd.

Fisher, R. A., & Yates, F. (1953). Statistical tables for biological, agricultural, and medical research (4th ed.). Edinburgh: Oliver & Boyd.

Fligner, M. A., & Policello, G. E. (1981). Robust rank procedures for the Behrens-Fisher problem. Journal of the American Statistical Association, 76, 162-168.

Gayen, A. K. (1949). The distribution of 'Student' t in the random samples of any size drawn from 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.

Glass, G. V., Peckham, P. D., & 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.

Gronow, D. G. C. (1951). Test for the significance of the difference between means in two normal populations having unequal variances. Biometrika, 38, 252-256).

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. Journal of Educational Statistics, 17, 315-339.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response time distributions: An example using the Stroop task. Psychological Bulletin, 109, 340-347.

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 598-615.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. Journal of Experimental Psychology, 69, 382-386.

Horsnell, G. (1953). The effect of unequal group variances on the F-test for the homogeneity of group means. Biometrika, 40, 128-136.

Howell, D. C. (1992). Statistical methods for psychology (3rd ed.). Boston: PWS-Kent.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47, 583-621.

Lehman, R. S. (1977). Computer simulation and modelling: An introduction. Hillsdale: Lawrence Erlbaum Associates.

Luce, R. D. (1986). Response times: Their roles in inferring elementary mental organization. Oxford: Oxford University Press.

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

Mielke, P. W., & Berry, K. J. (1994). Permutation tests for common locations among samples with unequal variances. Journal of Educational and Behavioral Statistics, 19, 217-236.

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

Moore, D. S., & McCabe, G.P. (1993). Introduction to the practice of statistics. New York: W.H. Freeman.

Morgan, B. J. T. (1984). Elements of simulation. London: Chapman & Hall.

Pratt, J. W. (1964). Robustness of some procedures for the two-sample location problem. Journal of the American Statistical Association, 59 665-680.

Ramsey, P. H. (1980). Exact Type I error rates for robustness of student's t test with unequal variances. Journal of Educational Statistics, 5, 337-349.

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85, 59-108.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. Psychological Review, 83, 190-214.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin, 2, 110-114.

Sawilosky, 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.

Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. Journal of the American Statistical Association, 65, 1501-1508.

Siegel, S., & Castellan Jr., N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd. ed.). New York: McGraw-Hill.

Smith, H. F. (1936). The problem of comparing the results of two experiments with unequal errors. Journal of Scientific and Industrial Research, 9, 211-212.

Tiku, M. L. (1964). Approximating the general non-normal variance-ratio sampling distributions. Biometrika, 51, 83-95.

Toothaker, L. E., & Newman, D. (1994). Nonparametric competitors to the two-way ANOVA. Journal of Educational and Behavioral Statistics, 19, 237-273.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika, 34, 29-35.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. Biometrika, 34, 29-35.

Zimmerman, D. W. (1987). Comparative power of the Student *t* test and Mann-Whitney U test for unequal sample sizes and variances. Journal of Experimental Education, 55, 171-174.

Zimmerman, D. W., & Zumbo, B. D. (1993a). Rank transformations and the power of the Student t and Welch t' test for non-normal populations with unequal variances. Canadian Journal of Experimental Psychology, 47, 523-539.

Zimmerman, D. W., & Zumbo, B. D. (1993b). The relative power of parametric and nonparametric statistical methods. In G. Keren, & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological issues (pp. 481-517).

Hillsdale: Lawrence Erlbaum Associates.

Zumbo, B. D., & Coulombe, D. (1994). Investigation of a statistical method for the generalized Behrens-Fisher problem. Manuscript under review.

Appendix 1

The following list identifies the single-factor case studies which comprised the population of 28 Monte Carlo studies used in Harwell's et al. (1992) study:

Bishop, T. A. (1976). Heteroscedastic ANOVA, MANOVA, and multiple-comparisons. Dissertation Abstracts International, 37, 3822B.

Blair, R. C., Higgins, J. J., & Smitley, W. D. S. (1980). On the relative power of the U and t tests. British Journal of Mathematical and Statistical Psychology, 33, 114-120.

Boehnke, K. (1984). F- and H-test assumptions revisited. Educational and Psychological Measurement, 44, 609-617.

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

Budescu, D. V., & Applebaum, M. I. (1981). Variance stabilizing transformations and the power of the F test. Journal of Educational Statistics, 6, 53-74.

Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. Journal of Educational Statistics, 1, 207-214.

Dijkstra, J. B., & Werter, P. S. P. J. (1981). Testing the equality of several means when the population variances are unequal. Communications in Statistics--Simulation and Computation, 10, 557-569.

Donaldson, T. S. (1968). Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. Journal of the American Statistical Association, 63, 660-676.

Feir-Walsh, B. J., & Toothaker, L. E. (1974). An

empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. Educational and Psychological Measurement, 34, 789-799.

Games, P. A., & Lucas, P. A. (1966). The analysis of variance of independent groups on non-normal and normally transformed data. Educational and Psychological Measurement, 26, 311-327.

Havlicek, L. L., & Peterson, N. L. (1974). Robustness of the t test: A guide for researchers on effect of violations of assumptions. Psychological Reports, 34, 1095-1114.

Kanji, G. K. (1976). The study of robustness of power in the analysis of variance. International Journal of Mathematical Education in Science Technology, 7, 401-407.

Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. Journal of Experimental Education, 43, 61-69.

Levine, D. W., & Dunlap, W. P. (1982). Power of the F test with skewed data: Should one transform or not? Psychological Bulletin, 92, 272-280.

Lin, L. I., & Sanford, R. L. (1983). The robustness of the likelihood ratio test, the nonparametric rank sum test and F-ratio tests when the populations are from the negative binomial family. Communications in Statistics--Simulation and Computation, 12, 523-539.

McSweeney, M., & Penfield, D. (1969). The normal scores test for the c-sample problem. British Journal of Mathematical and Statistical Psychology, 22, 177-192.

Nath, R., & Duran, B. S. (1981). The rank transform in the two-sample location problem. Communications in Statistics--Simulation and Computation, 10, 383-394.

Neave, H. R., & Granger, W. J. (1968). A Monte Carlo study comparing various two-sample tests for differences in means. Technometrics, 10, 509-532.

Norton, D. W. (1952). An empirical investigation of the effects of nonnormality and heterogeneity upon the F-test of analysis of variance. Unpublished doctoral dissertation. University of Iowa, Iowa City.

Olejnik, S. (1984). Conditional ANOVA for mean differences when population variances are unknown. Journal of Experimental Education, 53, 141-148.

Penfield, D. A., & Koffler, S. L. (1985, March). A power study of selected nonparametric K-sample tests. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Randolph, E., Robey, R., & Barcikowski, R. (1990, April). Type I error of the ANOVA revisited using power analysis criteria. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Rasmussen, J. L. (1985). An evaluation of parametric and non-parametric tests on modified and non-modified data. British Journal of Mathematical and Statistical Psychology, 39, 213-220.

Rasmussen, J. L. (1985). The power of student's t and Wilcoxon statistics. Evaluation Review, 9, 505-510.

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA F-test robust of variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. American Educational Research Journal, 14, 493-498.

Tomarkin, A., & Serlin, R. C. (1986). A comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychological Bulletin, 99, 90-99.

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F* statistics. Communications in Statistics--Simulation and Computation, 15, 933-943.

Zimmerman, D. W. (1987). Comparative power of student t

test and Mann-Whitney U test for unequal sample sizes and variances. Journal of Experimental Education, 55, 171-174.

Appendix 2

This Appendix provides tables of data from Algina et al. (1994) and Zimmerman and Zumbo (1993a) studies.

Table 7

Type I Error Rate of the t Test for the Lognormal Distribution (Algina et. al., 1994)

$n_1:n_2$

var1:var2	1:1	1:~2	1:3
1:1	N=100 .042	N=100 .043	N=100 .045
1:2		neg. pos.	neg. pos.
1:4	.065+	neg. .116+ pos.	neg. .157+ pos.
1:9	.082+	neg. .164+ pos.	neg. .232+ pos.

Note. $n_1:n_2$ = sample size ratios; var1:var2 = population variance ratios. Type I error rate = .05.

Table 8

Type I Error Rate of the Welch Test for the Beta Distribution (Algina et al., 1994)

$n_1:n_2$

var1:var2	1:1			1:2			1:3		
	N=20	N=40	N=60	N=20	N=40	N=60	N=20	N=40	N=60
1:1	.045	.049	.049	.052	.053	.053	.067+	.057	.056
1:2				neg. pos.			neg. pos.		
1:4	.055	.053	.050	neg. .072+ pos. .051	.062+ .048	.059	neg. .087+ pos. .043	.066+ .048	.062+
1:9	.067+	.058	.059	neg. .079+ pos. .062+	.066+ .051	.060+	neg. .084+ pos. .050	.069+ .055	.065+

Note. $n_1:n_2$ = sample size ratios; var1:var2 = population variance ratios. Type I error rate = .05.

Table 9

Type I Error Rate of the Welch Test for the Exponential Distribution (Algina et al., 1994)

var1:var2	$n_1:n_2$ = 1:1			1:2			1:3		
	N=20	N=40	N=60	N=20	N=40	N=60	N=20	N=40	N=60
1:1	.038−	.046	.048	.049	.053	.052	.069+	.062+	.062+
1:2				neg. pos.			neg. pos.		
1:4	.071+	.063+	.056	neg. .094+ pos. .046	.049	.072+	neg. .113+ pos. .035−	.045	.082+
1:9	.091+	.073+	.067+	neg. .099+ pos. .065+	.058	.080+	neg. .116+ pos. .047	.051	.082+

Note. $n_1:n_2$ = sample size ratios; var1:var2 = population variance ratios. Type I error rate = .05.

Table 10

Type I Error Rate of the Welch Test for the Lognormal Distribution (Algina et. al., 1994)

var1:var2	$n_1:n_2$ = 1:1, N=40	1:1, N=100	1:~2, N=40	1:~2, N=100	1:3, N=40	1:3, N=100
1:1	.038-	.041	.050	.051	.066+	.061+
1:2			neg. pos.		neg. pos.	
1:4	.077+	.066+	neg. pos. .041	.090+ .048	neg. pos. .036-	.103+ .046
1:9	.102+	.085+	neg. pos. .075+	.101+ .065+	neg. pos. .059	.117+ .050

Note. $n_1:n_2$ = sample size ratios; var1:var2 = population variance ratios. Type I error rate = .05.

Table 11

Type I Error Rates for the Exponential, Half-Normal, and Lognormal Distributions (Zimmerman & Zumbo, 1993a)

var1:var2	$n_1:n_2$ = 1:1				1:2	1:3
	N=24				N=24	N=24
	t	W	tr	tW		
1:1						
Exponential	.053	.050	.052	.052		
Half-normal	.047	.046	.044	.044		
Lognormal	.009-	.008-	.050	.050		
1:2					neg.	neg.
					pos.	pos.
1:4					neg.	neg.
					pos.	pos.
						t W tr tW
Exponential						.001- .021- .001- .007-
Half-normal						.010- .114+ .019- .062+
Lognormal						0 .067+ .163+ .158+
1:9					neg.	neg.
					pos.	pos.

Note. $n_1:n_2$ = sample size ratios; var1:var2 = population variance ratios; t = t test; W = Welch test; tr = t test on ranks (M-W-W); tW = Welch test on ranks. Type I error rate = .05.