SECCIÓN METODOLÓGICA

Psicológica (2002), 23, 415-450.

The Robustness of Validity and Efficiency of the Related Samples t-Test in the Presence of Outliers

Bruno D. Zumbo* and Martha J. Jennings**

*University of British Columbia **University of Ottawa

The performance of the related samples t-test (a one-sample t-test applied to the difference scores) given data which are essentially normal but contain outliers is largely unknown. In this Monte Carlo study the robustness of validity and efficiency for both the paired and one-sample t-tests are investigated. The Type I error rate and power of these tests given a normal underlying population are compared with the performance of these tests given a systematic range of outlier contamination in the underlying population. Sample sizes of 8, 16, 32, 64, and 128 are included in the design. Robustness of validity results are explored using regression models. Robustness of efficiency results are expressed using a proposed fairly stringent criterion for power. The results indicate that the t-test demonstrates fairly stringent robustness of validity for the range of symmetric contamination explored. When contamination is asymmetric the Type I error rate becomes inflated as the proportion of contamination increases. If robustness of validity is intact, power is not greatly affected when medium or large effect sizes are examined. This is not necessarily true for small effect sizes and the problems are further exacerbated when sample sizes are also small. Finally, a model with practical relevance for data analysts confronted with outlier contaminated data is developed using a novel index of contamination. This model is compared with a model using skewness and kurtosis values as disributional measures.

The objective of this Monte Carlo (MC) study is to provide a thorough examination of the effect of outlier contamination on the robustness of validity and efficiency of the related samples t-test. The related samples t-test is also referred to as the paired samples, repeated measures, or matched samples t-test. Although the paired-samples experimental design involving either

* Send correspondence to: Professor Bruno D. Zumbo, Dept. of ECPS, 2125 Main Mall, University of British Columbia, Vancouver, B.C. CANADA V6T 1Z4. E-mail: bruno.zumbo@ubc.ca This paper was initiated while the first author was Professor of Mathematics and Psychology at the University of Northern British Columbia. He is particularly grateful to his colleagues in Mathematics and Computing Science, and particularly Professor Lee Keener, who were both encouraging and supportive of this research program.

repeated measures or matching is not as frequently used in research as repeated measures designs involving more that two dependent groups (or repeated measures), studies periodically appear in which the paired samples t-test is the primary statistical test (for example, a 1993 study by Fallon et al. investigating a psychopharmacological intervention for in hypochondriacal concerns in the *Journal of Clinical Psychopharmacology*).

The related samples t-test is also used in the post-hoc analysis of more complex repeated measures analyses. A recent study by Zimmerman (1997) discusses the often overlooked advantages and disadvantages of paired-samples experimental designs. Furthermore, since the related samples t-test is a one-sample t-test applied to difference scores, our results also contribute to the understanding of the performance of the one-sample t-test.

Although a great deal of research has focussed on the psychometric properties of difference scores (e.g., Zimmerman, Williams, and Zumbo, 1993; Zumbo, 1999), little is known about the impact of outlier contamination on the robustness of the related samples t-test. In a repeated measures design an outlier can occur, for example, if one participant's baseline score changes (either gain or decline) markedly more than any other participant's change. An understanding of the impact of such outliers is necessary to ensure appropriate interpretation of the test results. The paired samples and one-sample t-tests are perfectly robust to violations of normality at infinitely large sample sizes (Bradley, 1980 a,b; Scheffé, 1959) but at some unknown sample size this robustness begins to break down if the underlying population distribution is not precisely normal. A greater understanding of this robustness is offered in this paper.

Three aspects of this study which make it particularly comprehensive should be highlighted. First, a systematic range of nonnormal population distributions is generated in this study using a contamination model. Previous researchers investigating the one-sample t-test have explored a limited range of nonnormal distributions, often using standard probability densities like the exponential or Cauchy. Second, both robustness of validity (Type I error rate) and robustness of efficiency (power) are discussed and a method of quantifying the degree of power robustness is suggested. Earlier studies have focused on only one type of robustness and no means of quantification was provided for those studies concerned with power. The third novel aspect of this paper is the approach adopted in the examination and expression of MC study results. MC studies in this area have principally relied upon tabulation and narrative description in the analysis of results. In addition to tabulation, the results of the robustness of validity portion of this study are examined using regression techniques inspired by Harwell (1992, 1993) and demonstrated in Zumbo and Harwell (1999). The use of regression, while not feasible for the power portion of this study, permits a thorough examination of the variables and their interactions in the robustness of validity portion of the study. These regression techniques are an integral part of response surface modeling (Khuri & Cornell, 1987) which has not been widely applied to MC study results (see Zumbo & Harwell, 1999).

The paper begins with a review of the literature pertaining to this problem. At the outset, the meaning of the term 'nonnormal' is clarified. Relevant concepts and terminology from the robustness literature are reviewed

and the concept of contamination models is outlined. Existing studies of the robustness of the paired and one-sample t-tests are discussed. While the previous robustness studies reviewed focus on the one-sample t-test, the results apply equally to the paired samples t-test. A detailed description of the methodology of the study is presented. The final sections of the paper present the results obtained in the study and discuss the implications of these results.

RELEVANT LITERATURE

Types of Nonnormality

The faith of the research community in the tendency of populations to be normally distributed has waxed and waned ever since the development of the normal distribution in the early 1800's. Mosteller and Tukey (1968) state "the history of statistics and data analysis is a messy mixture of healthy skepticism and naive optimism about the exact shapes of the distributions of observations" (pp. 86-87). It is beyond the scope of this paper to trace the historical development of this argument over the past two centuries but the interested reader is directed to Stigler (1973) for a fascinating description.

Both the Micceri (1989) and Stigler (1977) studies contribute to a body of evidence which suggests that nonnormality in one form or another is common in research data sets (Bradley, 1977; Mosteller & Tukey, 1968; Rosenthal, 1978, Wilcox, 1995a, 1995b). What is often unclear, is the manner in which the data analyst should proceed given these data sets. Are the sample distributions essentially normal with some aberrant points or do they reflect inherently nonnormal underlying population distributions? In this light, a clear difference between Stigler's data and Micceri's data can be seen. Stigler's data are essentially normal with heavy tails and some outliers. In contrast, Micceri's data descriptions suggest a more extreme nonnormality. When the term 'nonnormal' is applied to both types of distributions ambiguity results. For the purposes of this paper two types of nonnormality are identified. Truly nonnormal data sets include those described by Micceri. The expression 'normal with outliers' is used to describe data sets such as those encountered by Stigler.

Research concerned with the effect of violations of the assumption of normality on parametric tests can be divided into two groups based on these two types of nonnormality. Extensive research has been conducted to determine what happens when samples from truly nonnormal underlying distributions are unwittingly subjected to parametric tests (see Zimmerman & Zumbo, 1993, for a review). However, this research is not the main concern of this paper. The principle concern of this paper is with distributions that are essentially normal but contain some outliers. As stated earlier, in paired samples studies, outliers may arise due to aberrant amounts of change (either gain or decline). Outliers may also result from data entry errors or from atypical subject responses, which result from factors such as fatigue, motivation or failure to understand the task or test item. The tails of these distributions may be heavier than the normal distribution or if most of the errors are located on one side of the mean, the tails may be asymmetric.

Mosteller and Tukey (1968) identify this type of nonnormality as the most important because it is hard to detect, frequently ignored and yet drastically effect the sample mean and variance.

Whereas a number of studies have demonstrated that outliers are extremely common in research (Mosteller & Tukey, 1968; Rosenthal, 1978), relatively few studies of the paired samples or one-sample t-test which investigate robustness to this type of nonnormality were located. Accordingly, there is a pressing need for more systematic study of the impact of these distributions on the test performance.

Key Concepts in Robustness

The term 'robust', introduced by Box in 1953, refers to tests which are not greatly affected by violations of their assumptions (Boneau, 1960). At present, researchers typically refer to two types of robustness: robustness of validity and robustness of efficiency. Zimmerman and Zumbo (1993) and Sawilowsky and Blair (1992) among others have stressed the importance of considering both types of robustness in a comprehensive description of a statistical test. In addition, researchers have begun to label statistical procedures not simply as robust or nonrobust, but also to provide a measure of the degree of each type of robustness. These are the key concepts of robustness, which must be explored in order to thoroughly examine the behavior of the t-test in the presence of outliers.

Robustness of validity is said to occur if the accuracy of a statement made from a statistical procedure is not highly dependent on the assumptions of the underlying model being perfectly met (Wainer, 1982). Robustness of validity exists if the probability statements (expressed as the Type I error rate) made under violation of the assumption of normality are as accurate as those statements made when samples are drawn from the normal population. The robustness of validity issue is stated in terms of the alpha value as follows: Is the alpha value obtained from the t-test when the assumption of normality is violated the same as that obtained when this test is conducted from normal distributions?

A measure of the degree of robustness of validity must also be explored. Bradley (1978) proposes that a quantitative definition of robustness of validity can be achieved by stating, for a given alpha value, the range of simulation-based empirical Type I error rates for which the test would be considered robust. To exemplify this approach Bradley (1978,1980b) identifies three different levels of robustness which he terms fairly stringent, moderate and very liberal. The fairly stringent criterion is defined as the situation when the absolute value of p minus alpha is less than or equal to alpha divided by 10. Thus for an alpha level of .05, the fairly stringent criterion for robustness of validity would require obtained values of p to lie between .045 and .055. The moderate criterion is defined as the situation when the absolute value of p minus alpha is less than or equal to alpha divided by 5. Accordingly, for an alpha level of .05, the moderate criterion would require the obtained values of p to lie between .04 and .06. Bradley's very liberal criterion is defined as the situation when the absolute value of p minus alpha is less than or equal to alpha divided by 2. For an alpha level of .05 the very liberal criterion would require the obtained values of p to lie between .025 and .075. Bradley's

criteria for robustness of validity are applied to the Type I error rates in this study.

Robustness of efficiency refers to the power of a statistical procedure to find significant differences when the underlying assumptions are violated. Robustness of efficiency is said to exist when the power of a statistical procedure is the same under violation of the assumption of normality as it would be when normal distributions are used. To determine robustness of efficiency the power of the t-test obtained when the sample is drawn from a nonnormal population is compared with the power obtained when the sample is drawn from a normal population. The researcher begins by calculating the Type II error rate of the test, denoted as beta, and defined as the failure to reject the null hypothesis when it is false. Beta can be calculated once the sample size, alpha value and effect size (ES) are specified. Power is then calculated as 1-beta and the obtained value is the probability of rejecting a false null hypothesis.

The sample size and alpha value are easily specified in a Monte Carlo study. However, the choice of an appropriate ES is more problematic. In applied research studies substantive theory is used in determining the ES, in a MC study the choice of ES is more abstract. Cohen (1992) has defined small, medium and large ES indices for a number of different statistical tests. For the independent samples t-test the ES index is referred to as 'd' and is calculated by finding the difference between means and dividing by the within population standard deviation. The same process can be applied to the paired samples t-test. The resulting values are then classified as small (d=. 20), medium (d=. 50) and large (d=. 80) ESs. The medium ES for the t-test is equivalent to one half of a standard deviation and the small and large ESs are equal distances above and below the medium ES. In this study the power of the t-test for all three ESs is compared for samples from normal and nonnormal population distributions. At this point it is important to distinguish the measure of effect size (ES) from the non-centrality parameter used in directly evaluating the power via a non-central distribution function and a corresponding non-centrality parameter. For example, in classic examples of evaluating the power of a F-test via a non-central F distribution with a corresponding non-centrality parameter, one specifies the non-centraliy parameter and then computes one minus the cummulative probability of the test statistic's distribtion function, such as the F-test. In this manner, using the complement of the cummulative (non-central) distribution one can create a power chart for various values of the non-centrality parameter of the test statistic (e.g., the F-test). Unfortunately, this procedure is most ofen only useful under conditions where the test statistic's assumptions are true making it difficult to use in many situations where one is exploring the robustness of the test stastistic to assumptional violations, like non-normality. Therefore, it is importantly to note that evaluating the complement of a non-central distribution function is different than what is done in Monte Carlo simulation studies. In the typical Monte Carlo simulation we are mechanically duplicating the scientific use of hypothesis testing, while controlling the population values of dependent variable(s) and counting the number of false rejections over the replications. In this context we are not making use of a non-central distribution function to compute the statistical power directly. MC studies are

a more indirect mechanical method of computing the power against a non-zero effect size (i.e., a population mean, or in research settings with more than one group, the mean difference).

No standard method of quantifying robustness of efficiency was evident in the literature. For the purposes of this study a fairly stringent level of robustness of efficiency is suggested. This fairly stringent criterion is defined as a power difference of + or - 10% between the normal and contamination populations. This criterion is similar to Bradley's criterion for robustness of validity and is suggested as a useful starting point for quantifying robustness of efficiency.

Contamination Models

The data for this study were generated using contamination models, also known as mixed normal distributions or compound normal distributions. A contamination model is created by drawing the majority of data points from a parent distribution, denoted P_p and the remainder from a contamination distribution, denoted P_c. For example, P_p may be normal with a mean of 0 and a standard deviation of 1. P_c may have the same mean but a different standard deviation than P_p. In this case the contamination is symmetric. When P_c has a different mean than P_p, the contamination is asymmetric. Increasing the difference between the mean of P_p and P_c creates greater degrees of asymmetric contamination. In addition, greater degrees of outlier contamination can be created by increasing the proportion of sampling from P_c.

From this description of a contamination model, three parameters of contamination can be defined. The first parameter is the proportion of sampling from P_c. The second parameter is termed the mean shift and refers to the difference between the mean of P_p and the mean of P_c. The third parameter is the standard deviation of P_c. These three parameters are independent variables in this study and are selected to create a systematic range of symmetric and asymmetric contamination.

The notation used to describe contamination distributions is outlined by Mosteller and Tukey (1968). For example, if a parent distribution with a mean of 0 and a standard deviation of 5 is used for 90% of the sample values, it is denoted as N(0,5), p=.9. The contamination distribution with a mean of 1 and a standard deviation of 15 would be denoted N(1,15), 1-p=.1. Caution should be exercised when reading this notation in published studies since some researchers use the second value in the parentheses to refer to the variance in the contamination distribution rather than the standard deviation. This can create confusion when reviewing these studies.

The importance of contamination models as population models in a number of research settings is discussed by Blair and Higgins (1980). These authors point out that mathematical statisticians have suggested mixed normal distributions as a model for outliers as they may occur in various research domains. Thus, the use of a contamination model is consistent with the type of nonnormality being explored in this paper. In addition, the use of a contamination model provides a panoramic view of the performance of the t

test because a continuous range of nonnormal distributions can be generated by changing the parameters of the contamination distribution.

Evidence of Robustness for the t-Test

While a large number of studies have been published for the independent samples t-test, few studies of the paired samples or one-sample t-test were found. Those studies which could be located are divided into two groups in this literature review; those dealing with robustness to truly nonnormal distributions and those dealing with robustness to outliers. By far the majority of studies belong to the former group and these studies are reviewed because they provide some insight into the factors to consider in the design of a study of robustness to outliers.

In a simulation study, Chaffin and Rhiel (1993) investigated the effect of skewness and kurtosis on the Type I error rate of the one-sample t-test. No significant impact of kurtosis was found. However, with respect to skewness they showed that two-tailed tests are more appropriate than one-tailed tests given the levels of skewness investigated. For extreme skewness, two-tailed tests are only appropriate for large sample sizes and an alpha of .01. For moderate skewness, two-tailed tests have adequate robustness of validity at the .05 level even if the sample size is small. However, Lee and Gurland (1977) showed analytically that the Type I error rate of the one-sample t-test may differ greatly when sampling from distributions which have the same skewness and kurtosis. The authors used three different contaminated normal distributions all of which produce distributions with the same mean, variance, skewness, and kurtosis. The Type I error rates for these three distributions differed as a function of the fifth and sixth moments of the distributions. This paper demonstrates that skewness and kurtosis "provide only partial information about a distribution, but it is the whole structure of the nonnormal distributions which may effect the behavior of the t-test" (Lee and Gurland, 1977; p. 806). Investigations of skewness and kurtosis may provide a limited image of the robustness of the one-sample t-test. Other aspects of nonnormality must be considered and the Chaffin and Rhiel (1993) results should be interpreted with some caution.

The most extensive series of studies of robustness to truly nonnormal distributions is the work of Bradley (1977, 1978, 1980a, 1980b, 1980c). These studies used a distribution of response time data which the author described as L-shaped. Although he did not treat it as such, it should be noted that Bradley's (1977) L-shaped distribution can be conceptualized as a generalization of a contamination distribution which has several contaminating populations with varying proportions of contamination, location, and scale The performance of both the one-sample and the independent samples t-test was investigated using this data. Bradley compared the performance of the t-test when sampling from this L-shaped distribution with the performance when sampling from a bell-shaped (essentially normal) distribution. He identifies four factors as important in investigations of the one-sample t-test: the size of alpha, the location of the rejection region, sample size, and the shape of the population from which the sample was drawn (Bradley, 1978).

With reference to alpha values, Bradley (1978) demonstrated that the left tailed one-sample t-test did not meet the liberal criterion for robustness of

validity until N=256 at an alpha of .05 and did not meet this same liberal criterion at any N less than 1024 at alphas of .01 or .001. As the alpha value is decreased from .05 to .001 the robustness of the one-sample t-test diminishes. Similar results were obtained for the L-shaped distribution under various conditions in the Bradley 1980b and 1980c studies. These results prompted Bradley to conclude that an alpha value of .05 is the most robustness conducive alpha value. With reference to the location of the rejection region, Bradley investigated three situations in his studies: left-tailed, right-tailed and two-tailed rejection regions. He concludes that for the symmetric bell shaped distribution robustness is worse for two-tailed than for one-tailed t-tests. However, for the L-shaped distribution robustness for a two-tailed test is either superior to or intermediate between the robustness of right-tailed and left-tailed tests at the same alpha level. Bradley's results with reference to rejection region may be highly specific to the L-shaped distribution he explored but it is useful to note that when a distribution is markedly skewed the location of the rejection region may be an important factor in establishing the robustness.

Bradley's studies investigate sample sizes of 2, 4, 8, 32, 64, 128, 256, 512, and 1,024. In general he concludes that no N value below 512 ever brought the simulation-based empirical Type I error rate to within 10% of the alpha value for any combination of rejection regions and alpha values when sampling from the L-shaped population (Bradley 1980a). In addition, a sample size as great as 128 was required to bring the deviation of the simulation-based empirical Type I error rate from alpha to within 50% of alpha for the two-tailed test at an alpha of .05. He states "it clearly was not typical for the true probability of a Type I error to become statistically indistinguishable from alpha at small or moderate N values" (1980b, p.335). Furthermore, the obtained simulation-based empirical Type I error rates did not always deviate from the proposed alpha values in a conservative manner, as was observed by Boneau (1960) for the independent samples t-test when sampling from the exponential distribution. Rather, Bradley found that the simulation-based empirical Type I error rates were sometimes far greater than the alpha value and sometimes far less.

The fourth factor identified by Bradley as important in robustness studies of the t-test is the shape of the population from which the sample is drawn. The only nonnormal shape which he investigated is the L-shaped distribution. He concludes that the t-test is nonrobust under all circumstances when sampling from this distribution unless sample sizes are quite large. Bradley's results clearly indicate that very liberal definitions of robustness are obtained with this distribution only when sample sizes exceed 128 and are never achieved under some combinations of conditions with samples as large as 1024. There is some suggestion in his conclusions that these nonrobust results are the result of the highly skewed nature of the L-shaped distribution. In support of this contention, Sawilowsky and Blair (1992) demonstrated that while the independent samples t-test is reasonably robust for a number of nonnormal distributions, decidedly nonrobust results were obtained when distributions with extreme skew were used. This situation may also apply to the one-sample t test.

Two general conclusions about the robustness of the t-test are made by Bradley as a result of this series of studies. First, Bradley states that "robustness was strongly influenced by all of the factors investigated, and interactions among the influencing factors were often strong and complex" (1980b, p.333). Bradley also concludes that any statement concerning the robustness of a statistical test must be highly qualified and include the precise conditions under which the robust results were obtained. This seems like prudent advice in light of the varied results obtained under each of the conditions in Bradley's studies.

While Bradley's series of studies is arguably the most thorough exploration of the robustness of the t-test, there are two important limitations. First, the only nonnormal population he has considered is the L-shaped distribution. This distribution is clearly a truly nonnormal distribution and a researcher confronted with such a distribution would be compelled, in theory at least, to use nonparametric procedures. The second limitation of this study is that Bradley has examined only the Type I error rate. As stated earlier, a thorough examination of the robustness of parametric tests must include a discussion of the robustness of efficiency of the test through an examination of power.

The second group of studies, those exploring robustness of the t-test to the presence of outliers is of greater relevance to this paper. However, apart from the analytical work of Lee and Gurland (1977) which was discussed earlier, no systematic studies of this type of nonnormality were located for the paired or one-sample t-test. Indeed, the absence of studies is further testimony to the need for this study.

The Contamination Index

When using contamination models in a simulation study an awkward situation arises. The three parameters of contamination (mean shift, standard deviation of P_C, and proportion of contamination) are well suited to creating a systematic range of outlier contamination. However, these parameters are of no practical use for the data analyst confronted with a data set containing an unknown degree of contamination. That is, a data analyst has no way of knowing the values for any of these parameters for a given data set. Thus, the parameters of contamination are useful variables for the methodologist seeking a better understanding of the robustness of a test but they have no direct relevance for data analysts because these parameters cannot be estimated from sample data.

Researchers who are confronted with a data set containing an unknown degree of contamination often adopt an implied continuity principle when interpreting the results of a statistical analysis (Lind & Zumbo, 1993). That is, it is often assumed that data that deviate only slightly in form from the normal curve, will then only slightly distort the usual estimates of means, variances, correlations, and associated hypothesis tests. Likewise, with increasing departure from an underlying normal curve, the greater it is assumed, will be the inaccuracy of the computed statistics. Over the past several decades, research in statistics has demonstrated that such a continuity principle is invalid. The classical estimates of mean, variance, and correlation

have been shown to be highly sensitive to even small departures from an underlying normal curve. A single outlier, for example, can strongly bias these statistics and thereby provide misleading or invalid results (Huber, 1981; Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Lind & Zumbo, 1993; Zimmerman & Zumbo, 1993).

The poor performance of classical statistics in the presence of small departures from normality has led some statisticians (Hogg, 1977; Tukey, 1977) to warn that the routine use of classical statistics is unsafe. They recommend that classical estimates only be used in conjunction with alternative methods that are robust with respect to departures from normality. Although there is an increasing amount of statistical software that incorporates robust methods, these methods are seldom used in applied research. This is, in part, due to the lack of operational guidelines which inform the researcher as to their use.

As Lind and Zumbo (1993) state, a general procedure for using robust statistics in practice has been suggested by Hogg (1977) and Tukey (1977), henceforth referred to as the Hogg-Tukey procedure. This procedure involves three steps. The first step is to conduct the usual classical analysis of the data. The second step is to perform an analysis of the data using robust statistics. Finally, the results obtained from these two analyses are compared. If the results are similar, then Hogg and Tukey recommend reporting the results associated with the classical analysis in the usual manner. If the results obtained from classical and robust methods fail to agree, then the data should be re-examined for the presence of errors. If obvious errors are found, they can be corrected and the data re-analyzed. If obvious errors are not found then Hogg and Tukey recommend reporting the results from the anlaysis using robust statistics.

An example of the above procedure will help in motivating the discussion. Let us consider the problem of computing a confidence interval around a sample mean. A researcher confronted with fifty observations would compute the classical arithmetic mean, and due to reasons of optimality (Huber, 1981; Lind & Zumbo, 1993) an M-estimator such as the bi-weight (Beaton & Tukey, 1974). The arithmetic mean and the bi-weight are then compared. Due to sampling instability, the two estimators are not necessarily expected to be equal. However, there are presently no available guidelines as to how different the bi-weight and arithmetic mean should be before the robust estimator should be used. Unfortunately, the enormous contribution that the development of robust statistics can make to the scientific community is hampered by the lack of clear, practical, guidelines that can be easily employed by the scientist. Furthermore, one does not want to use the robust estimator exclusively because it is less optimal than the arithmetic mean when normality exists (Huber, 1981). A decision-making mechanism needs to be developed which can be easily computed from the data at hand and can be standardized for use with any sample (i.e. it should take into account the variability in the sample). The contamination index (CI) proposed in this paper may provide data analysts with an indicator of the extent of contamination in a data set.

The CI can be easily computed using the arithmetic mean and the bi-weight. Both the arithmetic mean and the bi-weight are available options on

many standard statistical packages such as SPSS or SAS. Standardization can be achieved by using a commonly recommended robust estimate of the standard deviation, referred to as the median absolute deviation (MAD) and calculated using

$$s_{med} = \frac{med\,|x_i - med_i\,x_i|}{0.6745}\,, \tag{1}$$

where: med_i denotes the median of a sample; x_i denotes a score in the sample; and 0.6745 is the constant value required to make the s_{med} unbiased (Huber, 1981). The MAD is the median value of the deviation of each score from the median of the sample and is easily calculated by applying a few simple mathematical procedures to the median which is already generated by standard statistical packages. Because the contamination index is standardized in this manner, a researcher can compare the index from one sample of data to another.

The proposed contamination index (CI) is calculated using

$$CI = \frac{|Mean_c - Mean_r|}{s_{med}}, \tag{2}$$

where: $Mean_c$ denotes the classical mean; $Mean_r$ denotes the robust mean; and s_{med} denotes the median absolute deviation. The numerator of this formula is the absolute value of the difference between the classical arithmetic mean and a robust estimate of the mean. Equation 2 represents a standardized deviation of the robust and classical estimators, and intuitively reflects the amount of contamination by outliers in a sample of data. Furthermore, CI is a data analytic measure (Mosteller & Tukey, 1977) and does not appear to have a workable mathematical distribution theory.

Therefore, CI provides the data analyst with a single number representing the degree of contamination present in a given data set. The larger the magnitude of the index of contamination the greater the degree of outlier contamination. Analysis of the results of this paper will show that this single number may be a more efficient method of characterizing the nonnormality present in a sample than the use of measures such as skewness and kurtosis.

RESEARCH QUESTIONS

Three research questions are examined in this paper using four dependent variables and four independent variables. The dependent variables are: Type I error rate (i.e. robustness of validity), and three levels of power (i.e. robustness of efficiency) corresponding to the small, medium, and large ESs discussed earlier. The first three independent variables are the parameters of the contamination model described earlier: the proportion of contamination, the mean shift and the standard deviation of the contamination distribution.

The values of these three variables are selected to create varying degrees of outlier contamination. There are three levels of each of these independent variables. The fourth independent variable is sample size. Sample sizes of 8, 16, 32, 64, and 128 are included in the design.

The first research question is: How is robustness of validity (Type I error) affected by variations in the parameters of the contamination distribution? Results for the first research question are expressed using Bradley's criterion for robustness of validity followed by fixed effects regression modeling. Harwell (1992) and Zumbo and Harwell (1999) suggest that the results of MC studies can be more readily understood by the use of regression techniques. They point out that numerous tables of values are difficult to synthesize in a meaningful manner. In addition, the use of narrative description can result in ambiguity and misinterpretations. Harwell states "the problem is one of correctly modeling variation in the empirical Type I error rates and power values as a function of study characteristics. Educational and psychological researchers would be well served by summaries of the effects of assumption violations for a test that would result from such modeling" (1992, p.300). Logistic regression has also been suggested for the analysis of MC study results. Harwell (1992) has shown that logistic regression techniques provide similar results to the fixed effects regression models. Fixed effects regression techniques are used in this study because they are more widely understood than logistic regression.

The second research question is: How is robustness of efficiency (power) affected by variations in the parameters of the contamination distribution? The fixed effects regression models used for the analysis of the robustness of validity are more problematic for the analysis of power. The reasons for this are outlined in the presentation of results. Accordingly, the second research question is examined using a fairly stringent criterion for robustness of efficiency.

The third research questions is: What distributional measures (e.g., skewness, kurtosis, contamination index) are useful for the data analyst confronted with potentially outlier contaminated data? Regression modeling techniques with skewness, kurtosis, and the contamination index (CI) as explanatory variables are presented in this part of the study.

METHODOLOGY

Selecting Parameter Values

At the outset of this study, the values for the parameters of the contamination distribution were selected to ensure that a range of outlier contamination in the sample would result. The selection process was guided by previous studies which used contamination models. The values selected by Rasmussen (1985) which included a mean shift of 33, a standard deviation of 10 for the P_C and a 5% proportion of sampling from P_C were considered too extreme to have practical application in research settings. In contrast, Mosteller and Tukey (1968) provide an example of a distribution which is sampled at 1% from a contamination distribution with a mean of 0 and a standard deviation of 3 (relative to the parent with a mean of 0 and standard

deviation of 1). This contamination model is used by the authors to determine the relative efficiency of some statistical procedures. Mosteller and Tukey describe this contamination as 'extreme' within the context of their example. Given this wide range of 'extreme' degrees of contamination, it was difficult to choose the values of the parameters for this study. Ultimately, the parameter values used by Blair and Higgins (1981) were selected and then slightly modified to create equal intervals between the levels of each parameter.

Three levels of each of the parameters of contamination were chosen. For proportion of sampling from the contamination distribution, values of 1% (.01), 8% (.08) and 15% (.15) were selected. For mean shift, values of 0, 1.5 and 3.0 were selected. The mean shift value of 0 indicates symmetric contamination while the other two mean shift values represent increasing degrees of asymmetric contamination. For standard deviation of the contamination distribution values of 0.5, 1.75 and 3.0 were chosen. The standard deviation of 0.5 for P_C is actually less than the standard deviation of 1.0 in the parent distribution. This means that the spread of the contamination distribution is actually less than the spread of the parent distribution. Very few outliers are likely to occur in this situation. The standard deviations of 1.75 and 3.0 for P_C are greater than the standard deviation in the parent distribution and have the effect of introducing increasing degrees of outlier contamination into the distribution. The values selected for parameters of the contamination model are shown in Table 1.

Table 1 establishes the basic design of this study. Each numbered box in the table represents a distinct population with a specific combination of parameters of contamination. A total of 27 different degrees of outlier contamination have been generated in this study. Each population can be described in terms of the parameters associated with that population. For example, population 2 is a distribution which has a proportion of sampling from the contamination distribution of .01 or 1%. The mean shift for this cell is zero so the contamination is symmetric. Finally, the standard deviation of the contamination distribution is 1.75 relative to the parent distribution with a standard deviation of 1. In contrast, population 27 is a distribution which has a proportion of sampling from the contamination distribution of .15 or 15%. The mean shift for this population is 3.0 so the contamination is asymmetric. The standard deviation of the contamination distribution is 3.0. For each of the 27 populations, five cells are included. to these contaminated populations, a normally distributed population is included in the design. Each cell represents one of the five sample sizes in the design. In addition.

The performance of each of the contaminated populations is compared to the performance of the normal population throughout the study. It is important to note that when considering the paired samples t-test herein, the simulation model is designed so that difference scores follow each of the investigated contaminated distributions.

In our simulation design, we are not making any statement concerning the marginal distributions of each set of scores, ie. the initial scores before the difference is computed.

Table 1. Parameter values and resulting population distributions in study.

			Standard Deviation of P_c		
Proportion	Mean Shift	n	0.5	1.75	3.0
.01	0	8 16 32 64 128	1	2	3
	1.5	8 16 32 64 128	4	5	6
	3.0	8 16 32 64 128	7	8	9
.08	0	8 16 32 64 128	10	11	12
	1.5	8 16 32 64 128	13	14	15
	3.0	8 16 32 64 128	16	17	18
.15	0	8 16 32 64 128	19	20	21
	1.5	8 16 32 64 128	22	23	24
	3.0	8 16 32 64 128	25	26	27

Generation of the Data

Pseudo random numbers were generated using a well-known and thoroughly tested prime-modulus multiplicative congruential generator described by Lewis and Orav (1989). The pseudo random numbers were transformed to a normal distribution using the Box-Muller method (1958). Evidence that the Box-Muller transformation is functioning as expected can be found in the values obtained for the normal distribution[1]. The mean, skewness, and kurtosis values for the normal distribution in theory should be close to zero. The values obtained for a sample size of 15,000 in the simulation were .0110, -.0253, and -.0119 respectively. The expected number of outliers for a sample size of 15,000 would be about 104. The number of outliers for the normal population in the simulation was 99. Each of the 27 contaminated populations was created by applying a transformation to the normal distribution. This transformation applies the mean shift and standard deviation of the specific contamination distribution to the normal distribution for the appropriate proportion of sampling from P_C (i.e. .01, .08, or .15). The accuracy of this method was tested by generating 15,000 cases for each of the contaminated distributions and for the normal distribution. The mean, skewness and kurtosis values were calculated for each of the populations from these 15,000 values. In addition, stem and leaf diagrams were plotted using SPSS. The hardware which was used for the simulation was unable to generate stem and leaf diagrams for samples greater than 15,000. Undoubtedly even greater accuracy would be demonstrated if larger sample sizes were used. This procedure generates a list of outliers for each stem and leaf diagram. Outliers or extreme values are identified, arbitrarily, in this program as beyond about 2.7 standard deviations from the mean.

Evidence that the proportion of contamination was increasing as expected in the study can be found by comparing the total number of outliers for populations 5, 14, and 23. Population 5 is contaminated at 1% and contains 146 outliers. Population 14 is contaminated at 8% and contains 282 outliers. Population 23 is contaminated at 15% and contains 443 outliers. These three populations have the same values for all parameters except proportion of contamination. Thus, the number of outliers increases as the proportion of contamination increases. It is important to recognize that the number of outliers contained in contaminated distributions provide only a crude indication of this type of nonnormality. The difficulty arises from the fact that outliers are identified by SPSS as data points beyond 2.7 standard deviations from the mean. However, the standard deviation is inflated when contaminated populations are being explored and the number of outliers is underestimated.

---

[1] As was pointed out by one of the reviewers Brately, Fox, and Schrage (1983, p. 210-211) suggest that care must be taken when using the multiplicative congruential pseudo-random number generator in conjunction with the Box-Muller method to generate normal random deviates. However, as also suggested by the reviewer, an exploration of pairs of normal deviates shows that the simulation method is valid for our use and that Brately et al.'s words of caution do not apply. The conclusion is comforting given that our simulation methodology is standard in the discipline.

Evidence that the mean shift parameter is functioning in the specified manner can be found by comparing the mean values for populations 20, 23 and 26. For population 20 the mean shift is 0 and the obtained mean is -.0053. For population 23 the mean shift is 1.5 and the obtained mean is .2183. For population 26 the mean shift is 3.0 and the obtained mean is .4638. Clearly, as the mean shift increases the value obtained for the mean also increases. Since the effect of increasing the mean shift is to create asymmetry, the number of outliers in each tail of the distribution is another useful method of assessing the effectiveness of the algorithm. For population 20 the mean shift is 0 indicating symmetric contamination. This population has 76 outliers in the left tail and 134 in the right tail. For population 23 the mean shift is 1.5 and 48 outliers are found in the left tail versus 395 in the right tail. For population 26 the mean shift is 3.0 and 12 outliers are found in the left tail versus 1017 in the right tail. Populations 20, 23, and 26 are identical for every parameter except the mean shift. Once again, it must be noted that the number of outliers contained in contaminated distributions is only a rough indication of this type of nonnormality. Despite this limitation, increasing degrees of asymmetric contamination are evident when comparing these three populations. These values indicate that the method of generating symmetric and asymmetric contamination is functioning as intended. Additional support for the methods used to generate the contamination models in this study is found in Tukey (1960) who demonstrated the analytical accuracy of these models. Having established that the method of data generation is sound the next step in the methodology is to determine the Type I error rates.

Individuals interested in the computer code used in the simulation should contact the first author.

Determining Type I Error Rates

The process of conducting the t-tests in this study is described for one cell in the expermental design (see Table 1). Consider population 1 with a sample size of eight. One sample of 8 values is drawn using the contamination model. A t-test is calculated to determine if the mean of this sample differs from the population mean which is set at 0. If the sample mean differs significantly, then a Type I error has occurred and the counter is incremented by one. The critical values for the two-tailed t-test were entered into the program for each of the five sample sizes at an alpha value of .05. A two-tailed test was chosen to allow the researcher to identify a significant result in either direction. Less information is available to the researcher with the use of a one-tailed hypothesis test. In addition, the use of a one-tailed test enhances the power of the test. While this may be desirable to a researcher looking for significance, it is not advantageous in a simulation study. The arguments in support of the use of two-tailed hypothesis testing are clearly outlined by Pillemer (1991).

This process is repeated 2000 times for sample size 8 from population 1. The total number of Type I errors on the counter after the 2000 replications have been completed is then divided by the number of replications to provide the probability of a Type I error. This number is recorded as one data point for that cell in the design. However, each cell in the design requires more than

one observation in order to estimate the parameters and test the fit of the regression model. Therefore, this process is repeated 15 times for each cell. Thus, the Type I error rate obtained for each cell in the design is based on 15 batches of 2000 replications of the t-test. The accuracy of the simulation program for Type I error was tested by examining the results for the normal population. The Type I error rate was close to the expected value of .05 in all cases. This indicates that the program functioned as intended.

Note that due to the type of outlier contaminatoin being modeled, for some of the cells in the experimental design the population distribution had to be centered to a mean of zero so that one could study the Type I error rates. Of course, this centering was also helpful in the statistical power portion of this project because we had control over the specified levels of effect size in studying the statisical power of the test. Therefore, all of the populations in Table 1 were centered to zero and then either the Type I error rate was studied or a specified effect size was modeled to study the statistical power.

Determining Power Values

Power values were calculated for three ESs: small (d=.20), medium (d=.50), and large (d=.80). ESs were introduced into the program by offsetting each sample value by the amount of the effect size. As in the Type I error program, a counter is created at the outset and set to zero. The t-test is then conducted. A significant result indicates that the difference has been detected and the counter is incremented by one. The power of the test to detect a given ES is determined by dividing the number on the counter by the number of replications in the design. As with the Type I error program, 15 batches of 2000 replications were conducted for each cell in the design.

The accuracy of the simulation algorithm for power was tested using the values obtained with the normal population. The expected power values for the one-sample or paired t-test were calculated using the method outlined in Cohen (1977, pp. 46-48). This method provides an expected power value for each sample size at each of the three ESs being investigated assuming that the underlying population is normally distributed. In all cases the power value obtained for the normal population in the simulation was within rounding error of the expected value calculated from Cohen. This indicates that the power portion of the simulation algorithm functioned as intended.

Obtaining Population Values for the Contamination Index, Skewness and Kurtosis

The previous sections of the methodology have described how the Type I error and power values were obtained in the study. It was also necessary to obtain a value for the contamination index (CI) for each of the 27 populations in the study. These population analog values were obtained using samples of 30,000 values for each of the 27 contaminated populations as well as the normal population. To calculate the population analog values the median absolute deviation was determined for each population in the design by applying Equation 1 to the median value computed from a sample of 30,000 values. The classical mean for each of the populations was obtained as well as the robust mean (biweight with a weighting constant set to 4.685) using SPSS.

Table 2. Values for the Contamination Index, Skewness and Kurtosis for Each Population Distribution.

Population	Contamination Index	Skewness	Kurtosis
1	0.000899333	.0035	-.0073
2	0.005081167	-.0297	.1810
3	0.002490768	-.0709	.9547
4	0.007505417	.0255	-.0920
5	0.012909570	.1232	.4473
6	0.099360828	1.2168	6.3921
7	0.033476657	.1940	.3957
8	0.028204539	.3923	1.5040
9	0.022417496	.7682	5.4638
10	0.006383281	-.0366	.2635
11	0.000678376	.0139	.4733
12	0.006388022	.1354	5.4500
13	0.001822973	-.0214	-.2901
14	0.071743140	.5299	1.4138
15	0.093086676	1.2596	6.8248
16	0.178228205	.5610	.3626
17	0.190677675	1.2486	3.2055
18	0.197227215	2.0290	8.7864
19	0.001563411	.0037	.3052
20	0.003823077	-.0030	.8766
21	0.002680132	.0041	4.8745
22	0.015724013	-.1027	-.3556
23	0.121087101	.7088	1.5650
24	0.168975390	1.2283	5.0908
25	0.217835562	.5146	-.2602
26	0.297612716	1.2162	2.1692
27	0.341866350	1.8664	5.6331
normal	0.002831484	-.0253	-.0119

These values were then entered into the formula for the CI (Equation 2). The resulting CI is a single number which indicates the degree of outlier contamination present in each of the populations. These values are provided as Table 2.

The accuracy of the method for calculating CI was assessed in three ways. First, the obtained population analog values should increase as the degree of outlier contamination in the population is increased. This was found to occur. Second, the CI was calculated for the normal distribution. The expected value of the index for the normal distribution should be very close to zero. This was found to occur. The skewness and kurtosis values obtained for samples of 15,000 values are also shown in Table 2. Comparison of these values with the CI also provides evidence that the program for calculating the CI values functioned as intended.

Table 3. Type I Error Rates for Each Population Distribution Under Study.

			Standard Deviation of P_c		
Proportion	Mean Shift	n	0.5	1.75	3.0
.01	0	8	.054633	.053733	**.056533***
		16	.051433	.051300	.048600
		32	.047733	.049567	.050467
		64	.050267	.049967	.050567
		128	.051267	.051900	.049500
	1.5	8	.054567	.052567	.052633
		16	.052900	.051333	.050700
		32	.047800	.051767	.047867
		64	.049167	.049033	.049533
		128	.051267	.051400	.047333
	3.0	8	.052833	.052967	.052900
		16	.050700	.052867	.047333
		32	.048600	.049767	.050400
		64	.047500	.051200	.049500
		128	.051133	.050933	.049133
.08	0	8	.054100	.053167	.049633
		16	.049400	.047967	.048467
		32	.049267	.048867	.047000
		64	.048933	.051200	.050167
		128	.050367	.050933	.051033
	1.5	8	**.056033***	.051400	.052967
		16	.051200	.051500	.052500
		32	.050000	.051700	.053733
		64	.049433	.049167	.053100
		128	.047733	.049100	.054933
	3.0	8	**.058667***	**.063233****	**.064233****
		16	**.056600***	**.063500****	**.068167****
		32	.051433	**.058400***	**.063800****
		64	.050467	**.056867***	**.058233***
		128	.050567	**.056600***	.054167
.15	0	8	.053967	.051467	**.044200***
		16	.051200	.052333	**.043700***
		32	.049800	.047867	.045933
		64	.049967	.050333	.051533
		128	.049833	.051467	.051800
	1.5	8	.054667	**.057300***	.054333
		16	.052533	**.055467***	**.058867***
		32	.050100	.053667	**.057267***
		64	.050667	.053933	**.055600***
		128	.052800	**.055100***	**.055067***
	3.0	8	**.070533****	**.077800*****	**.090467*****
		16	**.059433***	**.072800****	**.090067*****
		32	.052367	**.061400****	**.073967*****
		64	.052533	**.058300***	**.064967****
		128	.052367	.054867	**.055533***

Note: plain type-fairly stringent , * -moderate ,**-very liberal ,***- beyond very liberal

RESULTS AND CONCLUSIONS

Research Question One

To determine how robustness of validity is affected by variations in the parameters of the contamination distribution the Type I error rates (Table 3) are examined. Each tabled value is the mean of the 15 data points collected for that cell and is based on a total of 30,000 replications of the t-test (i.e. 15 batches of 2000 replications). These results are examined initially by applying Bradley's criterion and then using regression techniques. Bradley's fairly stringent criterion for
robustness of validity requires Type I error values to lie between .045 and .055. Values in this range are indicated in Table 3 in plain type. Values which satisfy the moderate criterion, between .04 and .06, but fail to satisfy the fairly stringent criterion are indicated in Table 3 using a single asterick (*). Values which satisfy the very liberal criterion, between .025 and .075, but fail to satisfy the moderate criterion are indicated in the table with an asterisk (**). Values which fail to meet even the very liberal criterion are indicated with a double asterisk (***).

The vast majority of Type I error values (75.5%) in Table 3 meet the fairly stringent criterion. An additional 14.8% of the values meet the moderate criterion while the very liberal criterion accounts for an another 6.67% of the Type I error values. A small proportion of the values (2.96%) fail to meet even the very liberal criterion. With the exception of two borderline values for samples of size 8, the robustness of validity of the t-test does not begin to deviate from the fairly strict criterion until population 17 in the design. This population has 8% asymmetric contamination, a mean difference of 3.0 and a standard deviation of 1.75. Symmetric contamination at 15% results in the fairly stringent criterion being met with the exception of sample sizes of 8 and 16 which meet the moderate criterion. The Type I error rate does not encounter serious inflation until the final two populations in the design. These populations have asymmetric contamination at a rate of 15%. The effect is reduced for sample sizes of 64 and 128.

These results indicate that the Type I error rate is quite stable for most of the degrees of contamination investigated in this study. Further, only asymmetric contamination creates a serious change in Type I error rate. When the Type I error rate is affected it tends to be inflated. This is not the 'conservative' effect reported in much of the literature for the independent samples t-test. The use of Bradley's criterion for robustness of validity is useful as a first step in the analysis of the results of this MC study. However, it is difficult to form any conclusions concerning the specific impact of each of the parameters of the contamination model and their interactions from Table 3. In order to fully answer this research question regression modeling techniques are used.

In the regression models the Type I error rate (TYPE I) is the outcome variable and the parameters of the contamination model along with sample size (N) are the predictor variables. These parameters are proportion of contamination (% CONTAM), mean shift (MEAN SHIFT) and standard

deviation of P_C (STD DEV P_C). The regression equations explored were of the form:

TYPE I = %CONTAM + MEAN SHIFT + STD DEV P_C + N.

Since the predictor variables are uncorrelated, an examination of the correlation matrix is all that is necessary to determine the direction and magnitude of the relationship between Type I error rate and each predictor variable (Budescu, 1993; Thomas, Hughes, & Zumbo, 1998). An examination of the correlation matrix allows for the ordering of the predictor variables in terms of their influence on Type I error rates.

The variable which correlates most highly with Type I error rate is MEAN SHIFT (.377). The positive correlation indicates that increases in MEAN SHIFT are associated with increases in Type I error rate. The second largest correlation is between % CONTAM and Type I error rate (.282). This correlation indicates that an increase in the % CONTAM is associated with an increase in Type I error rate. The third most important variable is sample size with a correlation of -.178. The negative correlation between sample size and Type I error rate indicates that as N increases the Type I error rate decreases. This is the expected direction of relationship between these two variables. The STD DEV P_C is the least important variable among the parameters of the contamination model (.139).

Table 4. B-Weights and Their Standard Errors for the Regression of the Parameters of Contamination Model on Type I Error.

Variable	b-weight	Standard Error
Constant	.0541	.00095
MEAN SHIFT	-.0018	.00046
STD DEV P_c	-.0011	.00045
% CONTAM	-.0241	.00918
N	-.00006	.00001
N*STD DEV P_c	.000025	.000006
N*% CONTAM	.00053	.00012
N*MEAN SHIFT	.000036	.000006
STD DEV P_c* % CONTAM	-.0081	.0043
STD DEV P_c*MEAN SHIFT	.00048	.00021
% CONTAM*MEAN SHIFT	.0292	.0042
% CONTAM*STD DEV P_c*MEAN SHIFT	.0202	.0018
N*STD DEV P_c *MEAN SHIFT	-.000016	.000002
N*% CONTAM * MEAN SHIFT	-.0005	.00004

In order to create a more parsimonious model, variables were selected according to two statistical criteria. First, those variables for which the b-weight was not statistically significant were removed from the model. Second, variables which had a significant b-weight but which accounted for less than 1% of the variance in Type I error rates were also removed. The complete

regression model including the three parameters of contamination and sample size accounts for 27.3% of the variance in Type I error rate. The addition of the six two-way interactions results in an increase of 20.8% in the variance explained. While the addition of the four three-way interactions results in an increase of 7.1% in the variance explained. Since the four-way interaction resulted in a small increase (1%) in the variance explained, the model including the three-way interaction terms is preferred. This model accounts for 55.2% of the variance in Type I error rates.

The three-way interactions were examined to determine if any were statistically non-significant. The t values which test the significance of each variable in the model showed that the interaction of N*% CONTAM*STD DEV P_C was not statistically significant and this interaction was removed from the model. The b-weights and their associated standard errors for this final model are shown in Table 4.

As an indicator of the appropriateness of the model the value of the constant is reasonably close to the expected value of .05. It should also be noted that the value of the b-weights is scale bound which means that the units for each variable must be considered when interpreting these values (Darlington, 1990).

B-weights must be interpreted carefully when interaction terms are included in a regression model. A two-way interaction means that the size of a conditional effect changes with another variable (Darlington, 1990). A three-way interaction means that the size of a two-way interaction changes with another variable. For this reason all of the two-way interactions (as well as the main effects) must be maintained when three-way interactions are included in the model. A three-way interaction can also be defined as the change in a two-way interaction associated with a 1-unit change in a third variable.

Each of the three-way interactions in the final model are discussed separately. The % CONTAM* STD DEV P_C* MEAN SHIFT interaction has a b-weight of .0202. This b-weight indicates the extent of the change in the two-way interaction of % CONTAM*STD DEV P_C associated with a 1-unit change in MEAN SHIFT. Specifically, a 1-unit increase in MEAN SHIFT results in an increase in the effect of % CONTAM * STD DEV P_C on the Type I error rate. More simply, the effect of the two-way interaction of % CONTAM * STD DEV P_C is less when the MEAN SHIFT is small than when the MEAN SHIFT is large. This interaction is best described by referring to Table 3 wherein it can be seen that when the MEAN SHIFT is zero increases in the STD DEV P_C do not have an effect on the Type I error rate even when the proportion of sampling from contamination increases. However, when the MEAN SHIFT is 3.0, increases in the STD DEV P_C do result in an increase in the Type I error rate for the 8% and 15% proportions of contamination.

A similar approach can be used to interpret the other two three-way interactions. The N*STD DEV P_C *MEAN SHIFT can be interpreted to mean that when the MEAN SHIFT is zero, increases in the STD DEV P_C do not have an effect on Type I error rates as the sample size decreases. When

the mean shift is 3.0, increases in the STD DEV P_C do result in an increase in the Type I error rate as the sample size decreases. The N*% CONTAM*MEAN SHIFT interaction can be interpreted to mean that when the mean shift is zero, increases in the % CONTAM do not have an effect on Type I error rates as the sample size decreases. When the mean shift is 3.0, increases in the % CONTAM result in an increase in the Type I error rate as the sample size decreases. For example, the Type I error values from Table 3 for population 27 are more inflated for the sample size of 8 (.0905) than for the sample size of 128 (.0555).

The residuals from these regression models were examined using scatterplots to determine if there was an obvious presence of asymmetry or outliers. The scatterplots did not reveal any trends. This finding suggests that the use of linear regression techniques is appropriate for this data set. However, since the Type I error rates have been recorded as proportions, a transformation of the data values might provide more meaningful results. One transformation which is appropriate for this situation is the logit transformation (Darlington, 1990). The utility of the logit transformation for this data set was examined by transforming all of the Type I error values and then running the same regression models discussed above. While the R^2 values tended to be slightly lower (approximately 1-2%) using the logit values, the results were very similar and there was no clear evidence that the logit transformation was advantageous. Given that the transformed values are more difficult to interpret, the results have been expressed using only the untransformed values.

Research Question Two

To determine how robustness of efficiency is affected by variations in the parameters of the contamination distribution the power values shown in Tables 5, 6, and 7 must be analyzed. Power values are only reported for cells in the design which satisfy the fairly stringent criterion for robustness of validity. When robustness of validity is not intact, the Type I error rate is not protected and the obtained values cannot be interpreted as true power values. A dash (-) is used to indicate these cells in the tables. The calculation of regression models for power is seriously hindered by these empty cells in the data set. Regression models using the parameters of contamination as predictor variables would be difficult to interpret and odd interactions might occur as a result of the pattern of empty cells. For these reasons, regression models were not computed for the power results.

The analysis of power results was undertaken using a method similar to Bradley's robustness of validity criterion. A fairly stringent criterion for robustness of efficiency can be devised by applying the same criterion as Bradley suggested for the Type I error rate. Therefore, power values which fall beyond + or - 10% of the power values actually obtained for the normal distribution are highlighted in Tables 5, 6, and 7. An upward pointing arrow symbol (⇑) indicates that the power value for the contaminated population exceeded the normal value by more than 10%. A downward pointing arrow (⇓) symbol indicates the power value for the contaminated population was below the normal value by more than 10%.

Table 5. Power Values for Each Population Distribution Under Study (Small Effect Size). ⇑ - power is above normal by > 10%, ⇓ - power is below normal by >10%.

			Standard Deviation of P_c		
Proportion	Mean Shift	n	0.5	1.75	3.0
.01	0	8	.085333	.087367	-
		16	.119400	.125233	.127900
		32	.198000	.209167	**.214333** ⇑
		64	.355067	.375933	.374767
		128	.622433	.652233	.648667
	1.5	8	.084967	.079500	.081200
		16	.118067	.113733	.117300
		32	.194933	.187267	.202000
		64	.348933	.332800	.365900
		128	.610800	.590400	.633167
	3.0	8	.079733	**.074933** ⇓	.079167
		16	.120500	**.105200** ⇓	.118700
		32	.198700	**.169333** ⇓	.200600
		64	.369033	**.310700** ⇓	.371033
		128	.641733	.560033	.654100
.08	0	8	.084667	.089933	.089767
		16	.111533	**.131733** ⇑	**.132200** ⇑
		32	.186067	**.217700** ⇑	**.213267** ⇑
		64	.329800	**.396733** ⇑	.372200
		128	.584400	.668433	.614967
	1.5	8	-	**.067733** ⇓	**.064667** ⇓
		16	.112367	**.105533** ⇓	**.094800** ⇓
		32	.186633	.187267	**.169667** ⇓
		64	.340833	.352600	.329533
		128	.594733	.637333	.605533
	3.0	8	-	-	-
		16	-	-	-
		32	.177833	-	-
		64	.355467	-	-
		128	.638867	-	.640433

Table 5. ***(continue)***

Proportion	Mean Shift	n	0.5	1.75	3.0
.15	0	8	.086700	.082500	-
		16	.121567	.120667	-
		32	.200667	.189567	.199133
		64	.351567	.344000	.335300
		128	.614700	.598000	.573133
	1.5	8	.085133	-	**.055300** ⇓
		16	.118033	-	**-**
		32	.194467	**.158667** ⇓	-
		64	.348900	**.304333** ⇓	-
		128	.605233	-	-
	3.0	8	-	**-**	-
		16	-	-	-
		32	**.164167** ⇓	**-**	**-**
		64	.326433	-	**-**
		128	.597233	.606600	-

⇑ - power is above normal by > 10% ⇓ - power is below normal by >10%

An examination of Tables 5, 6, and 7 reveals a number of interesting trends. For the small ES (Table 5) the power of contaminated populations is sometimes less than the power of the normal and sometimes greater than the power of the normal (shown in Table 8). Symmetric contamination results in a power advantage over the normal distribution. Asymmetric contamination results in a power loss relative to the normal distribution. For the medium ES (Table 6) only two cells are beyond the fairly stringent criterion for robustness of efficiency. This means that robustness of efficiency is greater for medium ESs than for small ESs. Both of the cells in the medium ES table which do not meet the fairly stringent criterion involve symmetric contamination and result in an increase in the power value relative to the normal distribution. In addition, both of these cells reflect a standard deviation of P_C of 3.0. The sample sizes for these cells are 8 and 16 respectively. For the large ES (Table 7) three cells lie beyond the fairly stringent criterion for robustness of efficiency. All of these cells are for sample sizes of 8 and have a standard deviation of P_C of 3.0.

Table 6. Power values for each population distribution under study (Medium Effect Size) ⇑ - power is above normal by > 10% ,⇓ - power is below normal by >10%.

			Standard Deviation of P_c		
Proportion	Mean Shift	n	0.5	1.75	3.0
.01	0	8	.249500	.252867	-
		16	.468900	.483567	.492800
		32	.784400	.797600	.796633
		64	.975933	.980000	.977533
		128	.999933	.999900	.999700
	1.5	8	.245667	.239400	.254767
		16	.465733	.458867	.486767
		32	.779633	.775500	.800700
		64	.974867	.975133	.979233
		128	.999833	.999800	.999933
	3.0	8	.240867	.236367	.255767
		16	.484200	.455967	.498100
		32	.792000	.778233	.815700
		64	.981267	.974533	.984267
		128	.999867	.999833	.999900
.08	0	8	.247100	.271033	**.299867** ⇑
		16	.462400	.501967	**.528567** ⇑
		32	.767533	.804100	.785033
		64	.973300	.981667	.967367
		128	.999867	1.00000	.999467
	1.5	8	-	.234433	.251500
		16	.456733	.473067	.508833
		32	.774133	.803967	.822600
		64	.975100	.982967	.985800
		128	.999967	.999933	.999867
	3.0	8	-	-	-
		16	-	-	-
		32	.815867	-	-
		64	.985867	-	-
		128	1.00000	-	1.00000
.15	0	8	.260867	.255667	-
		16	.474667	.475733	-
		32	.786067	.774233	.768133
		64	.976700	.972333	.961600
		128	.999900	.999800	.999633
	1.5	8	.238500	-	.251767
		16	.459200	-	-
		32	.776067	.798433	-
		64	.975167	.983600	-
		128	.999867	-	-
	3.0	8	-	-	-
		16	-	-	-
		32	.803167	-	-
		64	.986033	-	-
		128	.999967	1.00000	-

Table 7. Power Values for Each Population Distribution Under Study (Large Effect Size) ⇑ - power is above normal by > 10% ,⇓ - power is below normal by >10%

			Standard Deviation of P_C		
Proportion	Mean Shift	n	0.5	1.75	3.0
.01	0	8	.521733	.526967	-
		16	.849400	.860900	.863900
		32	.991967	.993367	.988300
		64	1.00000	.999967	.999967
		128	1.00000	1.00000	1.00000
	1.5	8	.515600	.512233	.533467
		16	.847067	.848833	.862633
		32	.991467	.991767	.992500
		64	.999900	1.00000	1.00000
		128	1.00000	1.00000	1.00000
	3.0	8	.517633	.518067	.549200
		16	.867533	.854467	.881967
		32	.993600	.993567	.996167
		64	1.00000	1.00000	1.00000
		128	1.00000	1.00000	1.00000
.08	0	8	.515700	.548900	**.605367** ⇑
		16	.846100	.864467	.852867
		32	.990700	.992267	.978067
		64	1.00000	1.00000	.999800
		128	1.00000	1.00000	1.0000
	1.5	8	-	.529067	**.588733** ⇑
		16	.848267	.874133	.895467
		32	.991600	.995500	.994400
		64	1.00000	1.00000	1.00000
		128	1.00000	1.00000	1.00000
	3.0	8	-	-	-
		16	-	-	-
		32	.997300	-	-
		64	1.00000	-	-
		128	1.00000	-	1.00000
.15	0	8	.531733	.535033	-
		16	.850533	.846800	-
		32	.991467	.990167	.978267
		64	1.00000	1.00000	.999767
		128	1.00000	1.00000	1.00000
	1.5	8	.504833	-	**.610100** ⇑
		16	.843500	-	-
		32	.992100	.996167	-
		64	1.00000	1.00000	-
		128	1.00000	-	-
	3.0	8	-	-	-
		16	-	-	-
		32	.998100	-	-
		64	1.00000	-	-
		128	1.00000	1.00000	-

Table 8 . Type I Error and Power Values for the Normal Distribution

N	Type I Error	Power - Small	Power - Medium	Power - Large
8	.054900	.083033	.246567	.519067
16	.051067	.118600	.474567	.853233
32	.050167	.193567	.780767	.992067
64	.049900	.351200	.975867	1.00000
128	.048700	.614467	.999867	1.00000

The results in Tables 4, 5, and 6 can be summarized with a few general statements. The robustness of efficiency of the t-test decreases as the effect size becomes smaller. At small ESs asymmetric contamination results in a power loss. Paradoxically, symmetric contamination results in a power advantage. For the medium and large ESs power differences are only noted for small samples sizes (i.e.. 8 and 16). For these sample sizes the power of the contaminated distributions exceeded the normal. Therefore, researchers should be aware that power differences for contaminated distributions will be most noticeable at small effect sizes or when small sample sizes are being used for medium and large effect sizes.

A second form of analysis was applied to the power data in order to better understand the results. The power values were converted into power curves using the program MacCurveFit (Raner, 1993). A second order polynomial curve was fit to the data values using the equation $Y= ax^2 + bx + c$, where Y denotes the power and x the effect size. A power curve for each cell was plotted on one graph together with the normal power curve for the same sample size, facilitating an immediate comparison of power under normal and contaminated populations.[2]

In only one cell, population 12 with a sample size of 8, was a power advantage greater than 10% noticed. Most of the remaining cells had differences of less than + or - 2%. The limitation inherent in the use of power curves to analyze robustness of efficiency lies in the fact that the importance of effect size as a variable cannot be determined because the ES is, in essence, integrated over or averaged out. The examination of power curves allows only a rough visual examination of the power differences at each ES. This is unsatisfactory given the importance of ES for the data analyst. For this reason the analysis of the power values through tabulation of the results and application of the fairly stringent criterion is preferable.

Research Question Three

The final research question concerns the utility of distributional measures (i.e. skewness, kurtosis, contamination index) for data analysts confronted with data containing outliers. The regression models reported earlier are of little pragmatic use for data analysts as they would have no means of determining the parameters of contamination in a data set. For this reason,

[2]Examples of these power curves can be obtained from the authors.

two additional regression models have been created for the data analyst. The first model uses skewness and kurtosis values as the predictor variables and the second model uses the CI. Given that skewness and kurtosis are readily available on statistics packages, these distributional measures provide one method for the data analyst to characterize the shape of the population distribution. Comparison of the first model with the results for the contamination index model is also a useful metric for this proposed index.

In the first model, both skewness and kurtosis are included. Horswell and Looney (1993) show that skewness and kurtosis coefficients when used jointly may provide a better method of assessing normality than skewness alone. The use of skewness tests alone is problematic and the authors show that these tests do not possess good specific diagnostic properties. They summarize research from a number of sources which demonstrate that some skewness coefficients have a high probablility of misdiagnosing non-skewed distributions as skewed. In addition to skewness and kurtosis, the first model includes sample size (N). Sample size was included in the model both because of the demonstrated correlation between N and Type I error rate (as shown in Table 4), and because it has conceptual importance to the data analyst.

The first model accounting for 35.93% of the variance in Type I error rates is expressed conceptually and with b-weights as follows. The standard error associated with each b-weight is shown in brackets below each variable.

TYPE I = CONSTANT + SKEW + KURT + N

TYPE I = .0526 + .0107*SKEW -.0011*KURT -.000035*N
(.00027) (.00036) (.000087) (.000003)

The two-way interactions for this model have also been examined. The omnibus model including skewness, kurtosis, sample size and the three two-way interactions which result from these variables accounts for 43.1% of the variance in Type I error rates. The interaction terms account for an increase of 7.2% in variance explained. The t-test of these interactions indicates that the SKEW*KURT interaction is not significant so this interaction was dropped from the model. The N*KURT interaction was shown to account for less than 1% of the variance and was also dropped from the model. Only the interaction of N*SKEW need be included in the model as it accounts for about 5% of the total variance. It should be noted that the three-way interaction of N*SKEW*KURT resulted in an R^2 change of less than 0.5% of the variance in Type I error; therefore the three-way interaction was not included in the model.

The final model is shown conceptually and with b-weights as follows. The standard error of the b-weights is shown in brackets below each variable.

TYPE I = CONSTANT + SKEW + KURT + N + N*SKEW

TYPE I = .0509 + .0142*SKEW - .0011*KURT - .0000004*N - .00007(N*SKEW)
(.0003) (.0004) (.00008) (.000004) (.000005)

Once again, the b-weights must be interpreted with care when interaction terms are included in the model. The b-weight for skewness (.0142) indicates the estimated conditional effect of skewness on Type I error rates when all other regressors are zero. The b-weight for the interaction of N*SKEW (-.00007) indicates that the conditional effect of skewness on Type I error rates changes with changing levels of sample size. Specifically, decreasing the sample size increases the effect of skewness on Type I error rates. This model, including the two-way interaction term accounts for approximately 41% of the variance.

The second model which was computed for the data analyst used the contamination index (CI) value as a predictor variable along with the sample size. This model is expressed as TYPE I = CONSTANT + N + CI. Since the values of CI and N are uncorrelated the magnitude of the b-weights can be used directly to indicate the variable ordering (Darlington, 1990). The model which results is expressed as

TYPE I = .0512 -.00003*N + .0530*CI
(.00025) (.000003) (.0015)

The R^2 value which results from this equation is .4042. Therefore, this third set of models using the contamination index accounts for about 40% of the variance in Type I error rate. It should also be noted that the value of the constant at .0512 is the value which would be expected for Type I error given a contamination index of 0. This is close to the nominal value of .05 given a normal distribution and provides further evidence that the simulation algorithm and population analog values for the CI are functioning as intended.

As with the skewness and kurtosis model, the interaction of the CI and N variables was examined. This two-way interaction results in an R^2 change of .0822 over the model with no interaction term. Since this value indicates that over 8% additional variance is accounted for by the interaction between sample size and CI, the interaction term should be included in any model used by data analysts. The use of the contamination index accounts for a greater amount of variance in Type I error rates (48%) than does the use of skewness and kurtosis model (41%).

The b-weights and associated standard errors for the CI model are shown as follows,

TYPE I = .0489 + .00001*N + .0813*CI - .00057(N*CI)
(.0003) (.000004) (.0021) (.00003)

These b-weights are interpreted in the same manner as for the previous model. For example the b-weight .0813 for CI indicates the estimated conditional effect of the contamination index on Type I error rate when all the

other regressors are zero. The interaction term N*CI has a b-weight of -.00057. This indicates that the conditional effect of CI on Type I error rate changes with changing levels of sample size. Specifically, the effect of CI on Type I error rate increases as the sample size decreases.

DISCUSSION

Three aspects of this study which contribute to its comprehensive nature were highlighted in the introduction. The results concerning the robustness of validity and efficiency of the t-test are discussed within the framework of these three aspects. First, the systematic range of 'essentially normal with outliers' populations which were generated using a contamination model was effective. These models provide an excellent method for investigating robustness in MC studies for a number of reasons. Previous researchers typically used methods of data generation which are more relevant to investigations of truly nonnormal underlying population distributions (i.e. standard probability densities such as the exponential or Cauchy). The use of a contamination model allows for a much more panoramic view of the factors which influence the robustness of the t-test to outlier contamination. For example, both symmetric and asymmetric contamination of varying degrees can be readily simulated. This is an important asset in light of the results observed for the paired samples t-test in this study. That is, both robustness of validity and robustness of efficiency functioned differently under conditions of symmetric versus asymmetric contamination. MC studies of robustness should strive to investigate both symmetric and asymmetric conditions and the use of a contamination model makes this feasible. In addition, the parameters of the contamination model can be useful in the identification of complex interactions among the factors which may influence robustness. These interactions can be identified using regression techniques This advantage is more thoroughly explored in the discussion of the regression techniques which follows. One final advantage to the use of contamination models should be mentioned; these models facilitate the replication of MC studies and provide a framework for future research. Specifically, the expansion of the parameters of the contamination model would permit a researcher to examine different degrees of outlier contamination.

It should also be noted that in future simulation studies that for the present study the results of the simulation for the t-test under conditions of true normality are used in order to validate to validate the simulation program and act as a baseline comparison. However, under adequate simulation design, these simulation results may also be used as a variance reduction technique as “control variates” in order to obtain even more accurate estimates of the Type I error rates and power.

The second aspect of this study which contributes to the comprehensive nature of the study is that both robustness of validity and efficiency were examined and a method of quantifying the degree of robustness of efficiency is suggested. Specifically, the results indicated that the paired samples or one-sample t-test satisfies a fairly stringent criterion for robustness of validity for

most of the degrees of contamination examined in this study. Robustness of validity is only seriously compromised when contamination is asymmetric and the proportion of contamination is 15%. The effect of contamination on robustness of validity in this study is to increase the Type I error rates. This finding is contrary to the conservative effect noted by Boneau (1960), Rasmussen (1985), and others, for the independent samples t -test. Bradley (1980a, 1980b, 1980c) found the Type I error rates were sometimes far greater and sometimes far less than the nominal rates for the t test when sampling from the L-shaped distribution. The results from the present study indicate that an inflation of Type I error occurs quite consistently when contamination is asymmetric and the proportion of contamination is 15%.

With reference to robustness of efficiency, the results of this study indicate that when robustness of validity is inflated, the power of the t-test cannot be determined. This observation highlights the need to consider both types of robustness within one study. If robustness of validity is intact then power values are maintained when medium and large ESs are examined. This means that given protected Type I error rates, the power values are also reasonably close to the expected normal values for medium and large ESs. For these ESs power differences are noted only for sample sizes of 8 and 16. However, when small ESs are being investigated, the power values are not as expected. Specifically, at small ESs, asymmetric contamination results in a power loss. Again, paradoxically, symmetric contamination results in a power advantage over the normal distribution for these small ESs. These effects are exacerbated when sample sizes are small. These results differ from the reduced power noted by Rasmussen (1985) and Zimmerman and Zumbo (1993) for the independent samples t-test. Both of these studies showed that power is improved when outliers are removed from the data set. Clearly, the inclusion of effect sizes in MC studies of robustness of efficiency is an important factor in developing a full understanding of these relationships.

Since no method of quantifying robustness of efficiency was evident in the literature, a criterion for robustness of efficiency similar to the Bradley criterion for robustness of validity is proposed in this paper. This criterion provided the most useful method of summarizing the power results in this study. The fairly stringent criterion for robustness of efficiency proposed in this paper requires the power difference between the contaminated population and the normal population to be within + or - 10% of the normal value. The application of this criterion to the tabled power values permitted a quick identification of contaminated populations which seriously affected the power of the t test.

The third important aspect of this paper is the examination of results through regression modeling. Regression modeling was particularly useful in examining the robustness of validity results in this study. The criterion for robustness of validity introduced by Bradley provides a useful starting point for examining the Type I error rates. However, this criterion cannot be used to investigate the complex interactions among the variables which influence Type I error rates. Regression models of the parameters of contamination indicate that mean shift and proportion of contamination account for the greatest portion of variance in Type I error rate. Sample size is negatively correlated with Type I error rate. As the sample size decreases the Type I

error rate increases. The real benefit of applying regression techniques to this study is that the models guided our interpretation of Type I error rates. The three-way interactions which became evident through modeling were not readily discernible from the tabled values and their narrative description.

Some limitations of the use of regression models for the analysis of Type I error results must be discussed. The use of R^2 values as a method of assessing these models is somewhat problematic for two reasons. First, the R^2 values have a slight positive bias (Darlington, 1990). The second difficulty with the use of these R^2 values is that there is little variance in the results of this study. This may be true in other MC studies. When very little variability exists in the results a large proportion of the variability may be due to sampling variability and not to any of the variables under investigation. When regression modeling is used for these results most of the variance is due to sampling or error variability and the R^2 value is attenuated. Therefore it is difficult to assess the appropriateness of these models. Finally, regression modeling is of no use in the analysis of robustness of efficiency because of the existence of a large number of empty cells in the design. These empty cells are the result of inflated Type I error rates. If regression modeling is attempted with these empty cells included in the design, the results are difficult to interpret. One avenue of future research is to find a means of using regression techniques in the analysis of robustness of efficiency.

The regression model which resulted from this process is informative for statisticians and methodologists seeking a better understanding of the performance of the paired samples and one-sample t- test. Unfortunately, this model is of little use to data analysts confronted with outlier contaminated data because the parameters of contamination cannot be determined. A set of models which would be of practical use for data analysts were created as part of the third research question. These models use a proposed 'contamination index' which can be readily computed by a data analyst using a standard statistics package. That is, the use of the contamination index (CI) enables a data analyst confronted with outlier contaminated data to quantify the degree of contamination present. In addition, the robustness of validity of the t-test can be modeled effectively using this index. The use of the CI together with sample size accounted for about 40% of the variance in Type I error rates. The addition of the CI*N interaction results in a total of about 48% of the variance being accounted for. The CI model has three advantages over the model using skewness and kurtosis. First, a greater proportion of variance in Type I error rate is accounted for using the CI. Second, since the CI and sample size are uncorrelated the model can be more easily interpreted than the skewness and kurtosis model. The third advantage is that the CI is conceptually linked to the presence of outliers. Skewness and kurtosis are more appropriate when true nonnormality is being considered. The advantages of the CI model lend support for the continued investigation of this proposed method for quantifying contamination.

For the researcher the results of the robustness of validity portion of this study indicate that a data set with a CI beyond about 0.20 will result in an unacceptable inflation in the Type I error rate. This effect is most serious

when small samples (i.e.. n < 16) are being used. This observation can be clearly demonstrated by inserting the values for the CI located in Table 2 into the regression model. For example, the CI value for population 1 is 0.0009. Given the equation TYPE I = .0489 + .00001*N + .0813*CI - .00057(N*CI) and a sample size of 8, the Type I error for this cell would be .0490. In comparison, for population 27 the CI is .3419 and the Type I error rate which would be associated with this degree of contamination for a sample size of 8, according to the model would be .0752. Interestingly, for this sample size and a CI value of 0.20 the resulting Type I error rate predicted by the model is .0714. The considerable inflation in Type I error rates for values beyond a CI of 0.20 is unacceptable because it can result in false claims of statistical significance. The values from the regression model using CI are intuitively correct. That is, population 1 is characterized by 1% symmetric contamination and results in little change in Type I error rate. By contrast, population 27 is characterized by 15% asymmetric contamination and results in a serious inflation of Type I error rate. This consistency in the results lends further support for the continued development of the CI as a useful measure of contamination. In this study population values of the CI were used in the regression modeling. Future research needs to investigate the use of sample CI values as a diagnostic for performance of the t-test.

In summary, this study provides further support for the apparent sensitivity of normal theory tests to the asymmetry of distributions. Harwell and Serlin (1989) state that there is a difficulty in checking the normality assumption of normal theory tests. The methodology used in this MC study provides one possible solution to this problem. The related samples t-test was chosen for this study because it is a logical starting point for a systematic empirical investigation of robustness. The issues of unequal sample sizes and variances are not relevant for this test and in this sense the related samples t-test is the simplest case. However, this methodology could be readily applied to other frequently used statistical tests to gain a better understanding of the impact of outliers on the performance of parametric procedures.

REFERENCES

Beaton, A. E., & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, *16*, 147-185.

Blair, R.C. & Higgins, J.J. (1980). The power of t and Wilcoxon statistics: A comparison. *Evaluation Review*, *4*, 645-656.

Blair, R.C. & Higgins, J.J. (1981). A note on the asymptotic relative efficiency of the Wilcoxon rank-sum test relative to the independent means t test under mixtures of two normal distributions. *British Journal of Mathematical and Statistical Psychology*, *34*, 124-128.

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

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. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician*, *31*, 147-150.

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

Bradley, J.V. (1980a). Nonrobustness in one-sample Z and t tests: A large-scale sampling study. *Bulletin of the Psychonomic Society*, *15*(1), 29-32.

Bradley, J.V. (1980b). Nonrobustness in Z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society*, *16*, 333-336.

Bradley, J.V. (1980c). Nonrobustness in classical tests on means and variances: A large-scale sampling study. *Bulletin of the Psychonomic Society*, *15*, 275-278.

Bratley, P., Fox, B. L., & Schrage, L.E. (1983). *A guide to simulation*. Springer Verlag.

Budescu, D.V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*, 542-551.

Chaffin, W. W., & Rhiel, G. S. (1993). The effect of skewness and kurtosis on the one-sample t-test and the impact of knowledge of the population standard deviation. *Journal of Statistical Computation and Simulation, 46*, 79-90.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Darlington, R.B. (1990). *Regression and linear models*. New York, NY: McGraw-Hill.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W.A. (1986). *Robust statistics: The approach based on influence functions.* New York: Wiley.

Harwell, M.R. (1992). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics*, 17, 297-313.

Harwell, M.R. (1993, July). *Analyzing and reporting the results of Monte Carlo Studies in item response theory*. Paper presented at the meeting of the European Psychometric Society, Barcelona, Spain.

Harwell, M.R. & Serlin, R.C. (1989). A nonparametric tests statistic for the general linear model. *Journal of Educational Statistics*, *14*, 351-371.

Hogg, R.V. (1977). An introduction to robust estimation. in R.L. Launer, & G.N. Wilkinson (Eds.), *Robustness in Statistics*. New York, NY: Academic Press.

Horswell, R.L. & Looney, S.W. (1993). Diagnostic limitations of skewness coefficients in assessing departures from univariate and multivariate normality. *Communications in Statistics: Simulation and Computation*, *22*, 437-459.

Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.

Khuri, A. I., & Cornell, J. A. (1987). *Response surfaces: Designs and analyses*. New York: Marcel Dekker, Inc..

Lee, A. F. S., & Gurland, J. (1977). One-sample t-test when sampling from a mixture of normal distributions. *Annals of Statistics*, *5*, 803-807.

Lewis, P. A. W., & Orav, E. J. (1989). *Simulation methodology for statisticians, operations analysts, and engineers, Vol. 1*. Pacific Grove, CA: Wadsworth.

Lind, J.C. & Zumbo, B.D. (1993). The continuity principle in psychological research: An introduction to robust statistics. *Canadian Psychology*, *34*, 407-412.

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

Mosteller, F. & Tukey, J.W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *The Handbook of Social Psychology: Vol. 2 Research methods* (pp. 80-203). Reading, MA: Addison-Wesley.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression.* Reading, Mass.: Addison Wesley Publishing Company.

Pillemer, D.B. (1991). One- versus two-tailed hypothesis tests in contemporary educational research. *Educational Researcher, 20*, 13-17.

Raner, K. (1993). *MacCurveFit* Version 1.0.3. Author.

Rasmussen, J.L. (1985). The power of Student's t and Wilcoxon W statistics: A comparison. *Evaluation Review, 9*, 505-510.

Rosenthal, R. (1978). How often are our numbers wrong? *American Psychologist, 33*, 1005-1008.

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, 3*, 352-360.

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

Stigler, S.M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885-1920. *Journal of the American Statistical Association, 68*, 872-879.

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

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, 45*, 253-275.

Tukey, J.W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, S.G. Ghwyne, W. Hoeffding, W.G. Madow, & H.B. Mann (Eds.), *Contributions to Probability and Statistics. Essays in Honour of Harold Hotelling* (pp. 448-485). Stanford: Stanford University Press.

Tukey, J.W. (1977). Robust techniques for the user. In R.L. Launer & G.N. Wilkinson (Eds.), *Robustness in Statistics*. New York, NY: Academic Press.

Wainer, H. (1982). Robust statistics: A survey and some prescriptions. In G. Keren (Ed.), *Statistical and Methodological Issues in Psychology and Social Sciences Research* (pp. 187-214). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Wilcox, R. R. (1995b). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology, 48*, 99-114.

Zimmerman, D. W. (1997). A note on the interpretation of the paired samples t-test. *Journal of Educational and Behavioral Statistics, 22*, 349-360.

Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement, 17*, 1-9.

Zimmerman, D.W., & Zumbo, B.D. (1993). Relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences. Volume 1: Methodological Issues* (pp. 481-517). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In Bruce Thompson (Ed.). *Advances in Social Science Methodology, Volume 5, (pp. 269-304)*. Greenwich, CT: JAI Press.

Zumbo, B. D., & Harwell, M. R. (1999). *The Methodology of Methodological Research: Analyzing the Results of Simulation Experiments* (Paper No. ESQBS-99-2). Prince George, B.C.: University of Northern British Columbia. Edgeworth Laboratory for Quantitative Behavioral Science.

(Manuscript received: 21/12/00; accepted: 27/4/01)