National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

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

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

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

AVIS

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

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

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

Canada

The Robustness of Validity and Efficiency of the One-sample t Test in the Presence of Normal Contamination

Martha Jennings B.A., B.Ed.
Faculty of Education

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfillment of the requirements for the
Master of Arts degree in Education

University of Ottawa

ISBN 0-612-00473-2

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Abstract

The performance of parametric tests given data which are essentially normal but contain outliers is largely unknown. In this Monte Carlo study the robustness of validity and efficiency for the one-sample location problem are investigated. The Type I error rate and power of the one-sample t test given a normal underlying population are compared with the performance of this test 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. The robustness of validity results are explored using three sets of regression models. The first set of models is constructed using the parameters of the contamination model and is intended to inform the social science methodologist. The second set of models is constructed using skewness and kurtosis values. A third set of models is developed using an index of contamination proposed by Zumbo (1993). This set of models has practical relevance to the data analyst confronted with outlier contaminated data. Robustness of efficiency results are expressed using both power curves and a proposed fairly stringent criterion for power. In general, the results indicate that the one-sample t test demonstrates fairly stringent robustness of validity for all the 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.

Acknowledgement

The author wishes to express her gratitude to Dr. Bruno D. Zumbo. His knowledge and enthusiasm throughout this endeavour were greatly appreciated.

Table of Contents

List of Figures

List of Tables

Introduction to the Problem

Parametric tests are some of the most frequently used statistical tools in educational research. The basic logic underlying these tests is that sample statistics can be used to estimate unknown population parameters and test hypotheses provided that certain assumptions are met. Two assumptions must be considered when using tests such as the z, t or F; the sample observations must be independently and identically distributed, and the underlying population distribution must be normal. Optimal performance of these tests requires strict adherence to these assumptions. The assumption of independence is often met by a randomization mechanism such as random assignment. Satisfying the assumption of normality is generally more problematic (Zumbo & Zimmerman, 1993). In most cases, the shape of the underlying distribution is not known with any degree of certainty. In addition, there is considerable evidence which suggests that nonnormality in one form or another is very common in educational research data sets (Bradley, 1977; Micceri, 1989; Mosteller & Tukey, 1968; Rosenthal, 1978). When the assumption of normality is violated the researcher cannot be certain that the results obtained from a parametric test are accurate.

A great deal of research has been conducted to determine the effect of violations of the assumption of normality on various parametric tests. This research can be divided into two distinct groups. The first group of studies is concerned with what happens when samples from truly nonnormal underlying distributions are unwittingly subjected to parametric tests. Clearly, if a researcher believes that the underlying population distribution is truly nonnormal then the assumption of normality is not tenable and nonparametric methods should be used. For this reason, while studies from this group are summarized in the literature review, they are not the main focus of this thesis.

The type of nonnormality considered herein is one of an essentially normal distribution which contains outliers. Far less research has been conducted on the impact of this type of violation than has been conducted with truly nonnormal distributions. Outliers are extremely common in social and behavioral science research (Mosteller & Tukey, 1968; Rosenthal,1978).

Given this observation there is a pressing need for more systematic investigation of the impact of outliers on the outcome of various parametric tests. The objective of this thesis is to examine the effect of outlier contamination on the robustness of validity and efficiency of the one-sample t test. One can demonstrate mathematically that the one-sample t test is robust to violations of normality at infinitely large sample sizes (Bradley, 1980b). However, at some unknown finite sample size this perfect robustness begins to break down when the underlying population distribution is not perfectly normal. The one-sample t test was chosen because it is the simplest educational research case, the univariate one-sample location problem. This test is used in educational research to calculate the confidence interval for a mean or to test the significance of a mean.

Researchers attempting to determine the robustness of a statistical test under violation of the assumption of normality have generally used one of two strategies. Mathematicians have relied strongly on asymptotic theory and other analytical approaches to gain information about robustness. Social science methodologists have conducted computer simulations or Monte Carlo (MC) studies to ascertain the robustness of a procedure. The exclusive use of either theoretical investigation or MC studies would, no doubt, result in a less than comprehensive examination of robustness for a given test. In support of this argument, Zimmerman and Zumbo (1993) discuss the derivation of power curves from asymptotic theory. They state that it is possible to derive power curves purely from theory and without need of computer simulation in some situations. However, when nonnormal populations are combined with small sample sizes, or when populations that do not have well-known, standard probability densities are being used computer simulation is needed. The authors stress that these two situations occur frequently in social science research. While these statements are made with specific reference to the derivation of power curves, they can quite reasonably be extended to include robustness studies in general. Accordingly, Monte Carlo simulation techniques are used in this study.

This thesis represents an expansion into new territory on four fronts. First, the parameters of this study have been chosen so that a systematic range of nonnormal population distributions are generated using a contamination model. Previous researchers have explored a very limited range

of nonnormal distributions and have focused on either robustness of validity or robustness of efficiency. Both types of robustness are examined for a wide range of nonnormal population distributions. The second novel area presented in this thesis is the use of a contamination index proposed by Zumbo (1993) as a measure of outlier contamination in a data set. The use of this contamination index provides a means for the data analyst to determine the extent of contamination which exists in a given data set. The third front which this thesis expands upon is in the examination and expression of results. Previous Monte Carlo studies in this area have principally relied upon tabulation and narrative description in the presentation of results. The results of the robustness of validity portion of this MC study are examined using regression techniques as suggested by Harwell (1992a & b, 1993). The modeling techniques are designed so that the results will be useful both for data analysts confronted with nonnormal data and for social science methodologists seeking a better understanding of the robustness of the one-sample t test. The fourth new area explored in this study is the quantification of the concept of robustness of efficiency. A fairly stringent criterion for this type of robustness is suggested. Thus, it is hoped that this thesis contributes to the field in both a practical and an epistemological way.

The thesis begins with a review of the literature pertaining to this problem. The purpose and development of the one-sample t test are reviewed with particular emphasis on the need for the assumption of normality. Population distributions are examined to clarify what is meant by the expression 'nonnormal'. Relevant concepts and terminology from the robustness literature are explored. Subsequently, methods for generating data in Monte Carlo studies are reviewed and the concept of contamination models is introduced. The derivation of the contamination index proposed by Zumbo (1993) is explained. Finally, existing studies of the robustness of the one-sample t test are discussed. Following this review of the literature a detailed description of the methodology of the study is presented. The final sections of this paper present the results obtained in the study and discuss the implications of these results for educational researchers.

Review of the Literature

The One-Sample t Test and the Assumption of Normality

The development of the t test is attributed to W.S. Gosset. As a chemist at Guiness Breweries, Gosset was provided with small samples for use in tests of product quality. He wanted to find a probability distribution of the difference between means which would be accurate for small samples (Ott, Larson, & Mendenhall, 1983). In Gosset's day researchers were using a procedure equivalent to the z test to compare means for large samples. Using this procedure the probability that the observed difference in means is due to chance, given the null hypothesis, can readily be determined with reasonable accuracy. However, two problems are encountered when the z test procedure is used with small samples. First, the sample standard deviation is not an accurate estimate of the standard deviation in the population when the sample is small. More importantly, the central limit theorem does not apply to small samples. This means that the use of the standard normal distribution as a probability distribution is not justified. Mosteller and Tukey (1968) indicate that prior to the development of the t test, many researchers used the z test for smaller samples and accepted their results with limited degrees of confidence. Gosset was not satisfied with the precision of this procedure and he endeavored to find a more accurate probability distribution of the difference between means for small samples.

Ultimately, Gosset found the appropriate distribution for this situation and he called it the t distribution. He also developed tables of the tail end values of this distribution which could be used in setting confidence limits and making significance tests. Gosset published his results in 1908 under the pseudonym of 'Student' because of company policies restricting publication. Thus, the statistical test which resulted from his work has come to be known as the Student's t test. The purpose of the one-sample t test is to determine the probability that the mean of a sample is drawn from an hypothesized population with a specified mean and unknown variance.

The t distribution is a unimodal, symmetric and bell shaped distribution with a graphical form similar to the normal distribution. Like the standard normal distribution, the t distribution has a mean of 0. However, there are some important differences between these two distributions

which should not be overlooked (Hays, 1973). Most of the differences between these distributions stem from the fact that the t distribution is actually a family of distributions with a different member for each sample size. At infinite or large sample sizes the t distribution yields the same probabilities as the normal distribution. In fact, the tables in most statistics textbooks indicate that the probabilities for these two distributions are essentially equal for sample sizes greater than 120 (see for example Shavelson, 1988). However, for any finite sample size the researcher must identify the exact t distribution to be used. This process is made simple by the use of degrees of freedom. For applications of the t distribution involving a single group the degrees of freedom is equal to N-1, where N denotes the sample size. In consulting tables of probabilities for the t distribution it becomes apparent that as the sample size decreases the t distribution differs considerably from the normal distribution. For smaller samples, the t distribution is flatter in the central region, has a wider spread and has more values occurring in the tails of the distribution (Hays, 1973).

Gosset's development of the t distribution represents an important contribution to statistics because it has enabled researchers to set confidence limits and make statistical significance decisions with greater accuracy. However, the use of the t test is limited to situations in which the underlying assumptions of this procedure are satisfied. The properties of the t test which necessitate the normality assumption are described in greater detail to emphasize the danger inherent in ignoring this assumption. The mathematical derivation of the t value is intrinsically linked to the normality assumption. The t value is calculated using

$$t = \frac{(M - \mu)}{s/\sqrt{N-1}},$$

where: M denotes the sample mean, μ denotes an hypothesized population mean, and s denotes the standard deviation in the sample. The numerator of this ratio is a random variable which indicates the difference between an observed mean and an hypothesized mean. The denominator is also a random variable. It is an estimate of the standard error of the mean based on the standard

deviation in the sample. Clearly, a t value is a random variable. However, t values are not estimates of a population value. Accordingly, they are referred to as test statistics rather than sample statistics (Hays, 1973). Like sample statistics, test statistics have sampling distributions which can be specified exactly using a density function. The sampling distribution of t has already been described in the preceding section. However, the t distribution as derived by Gosset is very difficult to specify unless the random variables in the numerator and denominator are statistically independent. These random variables are statistically independent if the information contained in the sample mean does not dictate the value of the sample standard deviation and vice versa. Thus, the researcher must determine under what conditions the mean and standard deviation are statistically independent. The following principle describes those conditions: "given random and independent observations, the sample mean M and the sample variance s^2 are independent if and only if the population distribution is normal" (Hays, 1973, p.311). Since the sample standard deviation is the square root of the sample variance, this principle applies directly to the calculation of the t value. In summary, unless the population is normal the sample mean and standard deviation are not independent across samples and the exact t distribution is very difficult to specify. This is the main reason that the use of the t test rests on an assumption of normality. When this assumption is violated the researcher cannot be sure that the t distribution is applicable to the data.

Unfortunately, this assumption will always be violated to some extent in any real data set. Evidence of this statement can be found in introductory statistics textbooks. For example, Shavelson states "the normal distribution does not actually exist. It is not a fact of nature. Rather, it is a mathematical model - an idealization - which can be used to represent data collected in behavioral research" (1988, p.109). Similarly, Mosteller and Tukey comment "so far as we know distributions that exactly fit this formula never occur in practice - not for individual observations, not for sample means, not for other derived quantities" (1968, p.87). Violations of the assumption of normality occur routinely in the practical use of the t test because precisely normal distributions exist only in theory. If the researcher wishes to continue to use parametric tests to generate accurate results then the extent and nature of the violation must be defined and the effect on the

outcome of the test must also be determined.

Two Types of Nonnormality

The assumption of normality refers to the underlying population distribution from which the sample is taken and not to the distribution observed in the sample. In some cases, particularly in the physical sciences, the shape of the underlying population distribution is known. If it is normal then the researcher proceeds with parametric tests. If the population is distinctly nonnormal the researcher would generally proceed with the use of a nonparametric test. Alternatively, a specific test procedure may be developed to deal with a known distinctly nonnormal population distribution. For the vast majority of cases in the social sciences the researcher is not able to determine the shape of the underlying population with any degree of certainty. In these situations the researcher must make a judgment as to the likelihood that the population is distributed normally and then proceed with the appropriate statistical test based on this judgment.

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 thesis 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. At present, the individual researcher must decide if the assumption of a normal underlying population distribution is tenable for a given data set. One method frequently used to shed light on the shape of underlying populations is to collect numerous samples of data and compare the distributions observed in these samples to the normal distribution. Two collections of data are discussed to illustrate this approach.

Stigler (1977) collected 20 sets of data from 1761 determinations of the parallax of the sun, from 1798 measurements of the mean density of the earth, and from circa 1880 measurements of the speed of light. As might be expected, the distributions of these real data sets did not precisely mirror the normal distribution. In general, Stigler found the sample distributions exhibited slightly

heavier tails than the normal distribution and most data sets contained a small number of outliers. He concluded that the data sets were not so extremely nonnormal as to warrant the use of nonparametric techniques and he proceeded to use these data sets to compare robust estimates of the mean. Thus, Stigler's data may be described as essentially normal with outliers. A researcher making a judgment about the shape of the underlying population distribution from these samples would likely conclude that the normality assumption is tenable.

Micceri (1989) also conducted a study which used collections of sample data to shed light on the shape of population distributions. This study has been cited frequently in recent educational and psychological research because it represents the first major attempt to describe the underlying population distribution of variables relevant to educational and psychological research. In this study, 440 distributions of data containing achievement and psychometric measures were collected. These data sets came from test results at the district, state and national level in the United States and from a number of research studies. Almost 70% of the data sets in this study contained 1,000 or more subjects. Micceri examined these data sets to determine the extent to which they differed from the normal. Specifically, he applied three measures of symmetry and two measures of tail weight to the data sets and compared these results to the normal distribution. He concluded that only 4.3% of the distributions could be considered even reasonable approximations to the Gaussian (p.164). Micceri particularly identified the lumpiness, multimodality and asymmetry of these distributions as problems. Overall, approximately 70% of the distributions were asymmetric, 50% exhibited lumpiness and 30% were classified as either bimodal or multimodal. Micceri concluded that "the current inquiry shows that even among the bounded measures of psychometry and achievement, extremes of asymmetry and lumpiness are more the rule than the exception" (p.161).

One potential difficulty with the approach used by Micceri to categorize these distributions should be identified. Horswell and Looney (1993) investigated the use of skewness coefficients in assessing departures from normality They conclude that the use of skewness and kurtosis coefficients, jointly, may provide a useful method of assessing normality. However, the use of

skewness tests alone is problematic. The authors demonstrate that these tests do not possess good specific diagnostic properties and therefore these tests do not reliably identify how a distribution departs from normality. They summarize research from a number of sources which demonstrate that some skewness coefficients have a high probability of misdiagnosing non-skewed distribution as skewed. Micceri's (1989) results must be examined with caution in light of these observations.

Aside from this potential problem, what is not made clear in Micceri's conclusions, is the manner in which the data analyst should proceed given data sets of the type he identified. Are the sample distributions essentially normal with some aberrant points or do they reflect inherently nonnormal underlying population distributions? In the former case, the researcher would most likely proceed with parametric techniques. In the latter case, nonparametric procedures would be required. In either case, a judgment is required and Micceri has not made his judgment clear.

The difference between the Stigler (1977) data sets and the Micceri (1989) data sets is important. Stigler's data are essentially normal with heavy tails and some outliers. In contrast, Micceri's data descriptions suggest a more extreme nonnormality. Evidently, the meaning of the term 'nonnormal' varies among researchers. For the purposes of this thesis 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 described by Stigler.

Figure 1. Schema for nonnormality.

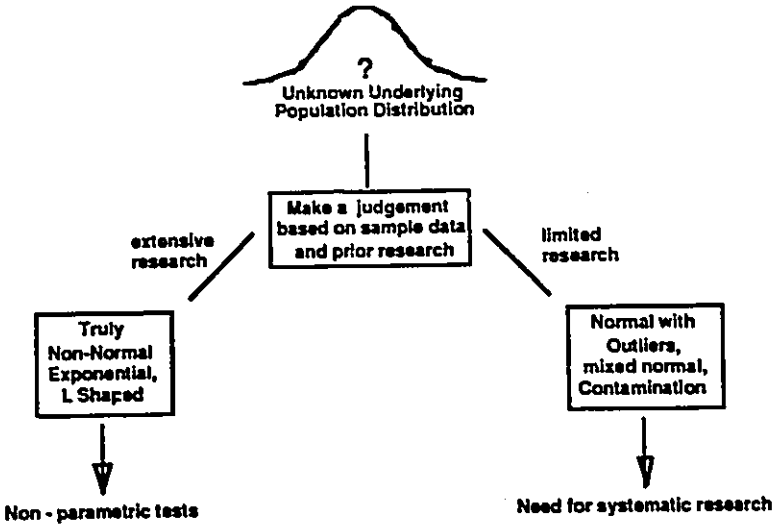


Figure 1 illustrates the two types of nonnormality discussed in this thesis. As stated earlier the researcher must make a judgment about the type of nonnormality contained in a given data set. This judgment is based on the available sample data and prior research in the field. The left side of Figure 1 illustrates one judgment that a researcher may make, a truly nonnormal underlying distribution exists. A clear example of this situation is the response time data collected by Bradley (1977). Bradley's sample distribution is L-shaped in form and prior response time research suggests that the distribution of this variable in the population may be truly nonnormal. Extensive research has been conducted on the effects of using parametric tests with truly nonnormal distributions. This research is summarized in this literature review. However, the effect of truly nonnormal population distributions on parametric tests is not the main concern of this thesis because a researcher confronted with truly nonnormal data is generally advised to use nonparametric tests.

The principle concern of this thesis is with the right side of the schema for nonnormality (Figure 1). These distributions are essentially normal but contain some outliers. Relatively few studies which investigate this type of nonnormality were located. Rosenthal (1978) has

demonstrated that outliers are extremely common in social and behavioral science research. The outliers identified by Rosenthal most commonly result from recording or data entry errors. Other sources of aberrance include misclassified individuals or atypical subject responses which result from factors such as fatigue, motivation or failure to understand the task or test item. The sampling distributions which result from these errors deviate somewhat from a precisely normal model. The tails of these distributions may be heavier than the normal distribution or in some cases where most of the errors are located on one side of the mean, the tails may be asymmetric. Indeed, scientists such as Simon Newcomb, Francis Galton, and Karl Pearson found a prevalence of these heavy tailed distributions as early as the 1850's (Stigler, 1973). In addition, Mosteller and Tukey (1968) identify this type of nonnormality as the most important because it is hard to detect, frequently ignored and yet drastically effects the sample mean and variance. Therefore, the term 'nonnormal' is used throughout this thesis to refer to these 'essentially normal with outlier' distributions. Given the prevalence of these data sets in educational and psychological research settings and the limited number of studies in this area, there is a pressing need for more systematic study of the impact of these normal with outlier distributions on the performance of parametric tests.

In addition to the presence of outliers in a data set, Wainer (1982) suggests that data analysts should also be concerned with 'fringeliers'. He defines fringeliers as data points that lie at or about three standard deviations from the center of the distribution. These fringeliers can have a strong influence on statistical procedures but are not as easily detected as outliers. Wainer states "in education and social science applications, we are primarily concerned with fringeliers in quantity or outliers in very small amounts" (1982, p.189). For the purposes of this study, fringeliers are included in the essentially normal with outlier distributions being investigated. A separate investigation of the specific impact of fringeliers is an interesting but wholly separate research problem.

Key Concepts in Robustness

The term 'robust' was introduced by Box in 1953 to refer to parametric tests which are not

greatly affected by violations of their assumptions (Boneau, 1960). This basic concept has been expanded in the past few decades. Researchers currently refer to two types of robustness, robustness of validity and robustness of efficiency. In determining the effect of violations of the assumption of normality on the one-sample t test both types of robustness must be considered. Zimmerman and Zumbo (1993) and Sawilowsky and Blair (1992) have stressed the importance of considering both types of robustness in order to comprehensively describe a statistical test. In addition, some 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 associated with a statistical procedure. These are the key concepts of robustness which must be explored in order to thoroughly examine the behavior of the one-sample t test in the presence of outliers.

Robustness of validity is said to occur if the accuracy or validity of a statement made from a statistical procedure is not highly dependent on the assumptions of the underlying model being perfectly met (Wainer,1982). The statements which are made based on a one-sample t test are statements concerning the probability that the sample mean is drawn from a specified population with a given mean. Thus, for the one-sample t test robustness of validity exists if the probability statements made under violation of the assumption of normality are as accurate as those statements made when samples are drawn from the normal population. This probability statement is made in terms of the Type I error rate. To determine if robustness of validity exists for the one-sample t test the Type I error rate obtained under violation of the assumption of normality is compared with the previously specified Type I error rate. The researcher selects an alpha value to indicate the Type I error rate considered acceptable. The robustness of validity issue can be stated in terms of the alpha value as follows: Is the alpha value obtained from a one-sample t test when the assumption of normality is violated the same as that obtained when this test is conducted from normal distributions? The symbol used to represent the probability obtained under the nonnormal distribution is p. Thus, the robustness of validity issue usually involves a comparison of the alpha value with the p-value.

A measure of the degree of robustness of validity must also be explored. Bradley (1978)

discusses this issue at length and concludes that a quantitative definition of robustness of validity can be achieved by stating, for a given alpha value, the range of p-values 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 ability of a statistical procedure to find significant differences when the underlying assumptions are violated. Essentially, 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 one-sample 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. In order to determine power under these conditions the researcher must begin by calculating the Type II error of the test. Type II error is defined as failure to reject the null hypothesis when it is false, denoted as BETA. The value of 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 can easily be specified in a Monte Carlo study. However, the choice of an appropriate effect size is more problematic. In most research studies substantive

theory is used in determining the ES. However, in a Monte Carlo study the choice of ES is more abstract. Useful guidelines for selecting effects sizes are provided by Cohen (1992). Small, medium and large ES indices have been operationally defined 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 one-sample t test using the sample mean and the hypothesized population mean. The resulting values are then classified as small (d=.20), medium (d=.50) and large (d=.80) effect sizes. The medium ES for the t test is equivalent to one half of a standard deviation. The small and large ESs are equal distances above and below the medium ES. In this study the power of the one-sample t test for all three effect sizes is compared for samples from normal and nonnormal population distributions.

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.

Methods of Data Generation in Monte Carlo Studies

Three forms of data generation have been employed in existing Monte Carlo (MC) studies of robustness. Many early MC studies generated simulated data sets using specific mathematical functions such as the rectangular, logistic, exponential and Cauchy distributions (see for example Boneau, 1960). Recently, researchers have advocated the use of real data in MC studies (Bradley, 1977; Micceri, 1989; Stigler, 1977). Still more recent is the suggestion that the results from existing MC studies be collected and used as the data set in a meta analysis examination of robustness (Harwell, 1992b; Harwell, Rubinstein, Hayes & Olds, 1992). Each of these three forms of data generation has advantages and disadvantages. The early MC studies using specific mathematical functions are useful in describing how parametric procedures perform under these conditions. Unfortunately, this approach to data generation may not reflect real data to any

reasonable extent. For example, the data sets collected by Micceri (1989) which were discussed earlier demonstrated a distinct prevalence of asymmetric distributions. Many of the specific mathematical functions selected for MC studies generate symmetric distributions of data. The information in these studies may be of considerable mathematical interest but these studies have limited practical application for the data analyst. In addition, the use of specific mathematical functions to generate data is clearly better suited to investigating the impact of truly nonnormal distributions than to the study of the impact of outliers on a parametric test. Accordingly, this method of data generation is not used in this study.

The meta analysis approach to robustness studies is perhaps most useful as a method of summarizing existing research. However, there are some serious limitations to using this approach as an alternative to other methods of data generation. Perhaps the most serious limitation is that the meta analysis approach limits any new robustness inquiry to the range of conditions examined in previous studies. In addition, the meta analyst must be concerned with missing values, potentially unequal cell sizes in the experimental design and a limited degree of information describing the manner in which the original data were generated. Thus, while the meta analysis approach provides information about the range of population distributions and sample sizes already studied for a given statistical test, it is not conducive to providing a comprehensive examination of the one-sample t test.

Given the limitations of generating data using specific mathematical functions and the problems inherent in the meta analysis approach, it would seem that the use of real data in MC studies is advisable. As discussed earlier, Stigler (1977) used real data sets of astronomical measures in his study of robust estimates of the mean. Similarly, Bradley's (1977) response time data were collected by conducting 2,520 trials of the time required to reach up from a fixed position and operate a push button. He described the distribution for this data set as L-shaped and he used this distribution in a number of robustness studies which are discussed later in this literature review. Bradley also identified a number of other practical research settings which generate a similar L-shaped distribution. Clearly, there was some support for the use of real data in MC

studies even before the publication of Micceri's 1989 article. Since the publication of this article an increasing number of researchers have adopted this approach. For example, Sawilowsky and Blair (1992) conducted a MC study of the robustness of the independent samples t test using eight distributions of real data identified by Micceri as commonly occurring. Similarly, Sawilowsky and Hillman (1992) investigated the power and Type I error of the independent samples t test under a distribution described as discrete mass at zero with gap. This distribution frequently occurs with first use or onset variables such as the age when a subject first began smoking. In both of these studies the authors have defended their choice of distributions as being more representative of the types of distributions encountered in educational and psychological research settings.

While the use of real data sets arguably provides more meaningful information than the mathematical functions applied in earlier MC studies, there are some drawbacks to the real data strategy. Essentially, the use of a real data set in a MC study provides only a snapshot of the robustness of the test under any given distribution. To illustrate this point consider the Sawilowsky and Blair (1992) study cited above. While eight different distributions are examined, only a snapshot of each condition is provided. With reference to asymmetry the authors include one real data set which is positively skewed and extremely asymmetric and another distribution which is negatively skewed and extremely asymmetric. Thus, the results indicate how the independent samples t test performs under two conditions of extreme asymmetry. Indeed, this degree of asymmetry might prompt a researcher to conclude that the normality assumption is not tenable for this data set and to abandon the use of parametric methods. A more comprehensive approach is to provide a panoramic view of how the robustness of the t test changes as a function of varying degrees of asymmetry. Thus, extreme asymmetry would be investigated as well as the less extreme asymmetric tails which may result from the presence of outliers or aberrant points. This panoramic view is difficult to obtain using real data sets.

Contamination Models

A viable alternative to the use of real data sets is to generate artificial data using

contamination models. A contamination model, also known as a mixed normal distribution, 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. P_p is 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. Alternatively, P_c may have a different mean and standard deviation than P_p. In this case the contamination is asymmetric. Increasing the difference between the mean of P_p and P_c creates increasing degrees of asymmetric outlier 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 simply 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 the design of 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. A note of caution should be exercised when reading this notation in published studies. 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 or mixed normal distributions as population models in educational research, as well as a number of other research domains, is discussed by Blair and Higgins (1980). These authors also 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 thesis. 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 discussed above.

Previous Evidence of Robustness for the t Test

While a large number of studies have been published for the independent samples t test, there are relatively few studies in which the one-sample t test is examined. Those studies which could be located are divided into two groups in this literature review; those in which robustness to truly nonnormal distributions is explored and those in which robustness to outliers is explored. By far the majority of studies belong to the former group. Studies of robustness to truly nonnormal distributions are reviewed because they provide some insight into the factors which should be included in the design of a study of robustness to outliers.

The most extensive series of studies in this first group is the work of Bradley (1978, 1980a, 1980b, 1980c) using the response time data described earlier in this thesis. Bradley conducted a number of different simulations in which the performance of both the one-sample and the independent samples t test is investigated. In these simulations Bradley compared the performance of the t test when sampling from his 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). All of these factors are manipulated in each of the studies Bradley conducted and a summary of the results associated with each factor is provided.

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. However, it is important to note that when a distribution is markedly skewed the location of the rejection region may be an important factor in establishing the robustness of the t test.

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 p-value 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 p-value 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 p-values 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 p-values were sometimes far greater than the alpha value and sometimes far smaller.

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 has 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 largely the result of the highly skewed nature of the L-shaped distribution. In support of this contention, Sawilowsky and Blair (1992) have 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). This conclusion can only be applied to the L-shaped population which Bradley explored. The second general conclusion made by Bradley is 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 one-sample t test, there are still 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 generally be advised to use nonparametric procedures. The second limitation of this study is that Bradley has examined only the Type I error rate for each of the samples. 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 under various conditions of violation.

The second group of studies, those exploring robustness of the t test to the presence of outliers is of greater relevance to this thesis. Unfortunately, no systematic studies of this type of nonnormality were located for the one-sample t test. Indeed, the absence of studies of the one-sample location problem is further testimony to the need for this study. However, a number of

simulation studies of the robustness of the independent samples t test to the presence of outliers have been located and are summarized. These studies are reviewed because they provide some insight into the factors which should be considered in the design of a study of robustness to outliers.

Rasmussen (1985) conducted a simulation study in which sampling was 5% from a contamination model with a mean of 33 and a standard deviation of 10 (given a normal parent distribution with mean of 0 and standard deviation of 1). The parameters of this contamination model indicate that the robustness of the independent samples t test to asymmetric outlier contamination is being investigated. He found the Type I error rate to be 'in line' with the nominal values and examination of his tabled results indicates that the t test functioned in a conservative manner for most conditions. This means that the independent samples t test demonstrates remarkable robustness of validity given this degree of outlier contamination. In terms of robustness of efficiency, Rasmussen found that the power of the independent samples t test is greatly diminished when extreme outlier contamination is present in the underlying population distribution. Two general conclusions can be taken from Rasmussen's study; robustness of validity to a rather extreme degree of outlier contamination is demonstrated but robustness of efficiency is severely compromised.

Zimmerman and Zumbo (1993) investigated the power of the independent samples t test using a number of specific mathematical functions as well as a mixed normal model. Of particular interest to this study is the mixed normal model they investigated. In this model 15% of the sampled values were from a contamination distribution with a mean of 0 and a standard deviation of 25 (given a normal parent with a mean of 0 and standard deviation of 1). The parameters of this mixed normal model indicate that the power of the independent samples t test is being explored given the presence of symmetric outlier contamination. The results of this study indicate that the power of the t test is improved if outliers are removed from the data prior to the analysis. These results are in keeping with those obtained by Rasmussen (1985) for asymmetric contamination.

Finally, Blair and Higgins (1980,1981) conducted a series of studies which examined the

power of the independent samples t test under a systematic range of contamination distributions. The percentages of contamination included in their design were 5%, 10%, 30% and 50%. Asymmetric contamination was investigated in varying degrees by increasing the mean shift between P_p and P_c. Mean shifts of 1,2,3,5, and 10 are reported. The standard deviation of P_c is varied to include 0.5, 1.0, 2.0, 4.0, and 8.0 (given P_p with a standard deviation of 1). While these studies are quite comprehensive and systematic, their purpose was to determine the asymptotic relative efficiency of the independent samples t test compared to nonparametric procedures. Thus, the results are expressed as asymptotic relative efficiencies and are of no use for this study. However, the systematic range of contamination models devised by Blair and Higgins (1981) is the basis for the experimental design used in this thesis.

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 relevance for data analysts.

A solution to this situation may be found in the contamination index proposed by Zumbo (1993). This index is based on a general procedure for using robust statistics in practical applications which is outlined by Hogg (1977) and Tukey (1977). This procedure is summarized by Lind and Zumbo (1993) and involves four steps. First, the usual analysis of the data is performed using classical statistical methods. This analysis is followed by an analysis of the data using robust methods. If the classical analysis results agree with the robust results then the classical analysis results are reported in the usual manner. However, if the robust results fail to

agree with the classical methods then the data should be re-examined for the presence of errors. A decision can then be made as to the treatment of these errors. The proposed contamination index is based upon the same general principle of examining the difference between robust and classical statistics.

The contamination index (CI) is calculated using

$$CI = \frac{|(Mean_c - Mean_r)|}{S_{med}}, \qquad (1)$$

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. The robust estimate of the mean which is suggested by Zumbo (1993) is the biweight. The denominator of this formula is the median absolute deviation. The median absolute deviation is a robust estimate of the standard deviation which is calculated using

$$S_{med} = \frac{med_i|x_i - med_i x_i|}{0.6745}, \qquad (2)$$

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). Thus, the median absolute deviation is the median value of the deviation of each score from the median of the sample. The median absolute deviation is frequently used as a robust estimate of the standard deviation.

The contamination index provides a measure of the extent to which the classical mean of a sample deviates from a robust estimate of the mean. This difference is standardized by taking into account the variability in the sample. As shown in the above formula, the measure of variability used in the contamination index is the median absolute deviation. Because the contamination index is standardized in this manner, a researcher can compare the index from one sample of data to another. The larger the magnitude of the index of contamination the greater the degree of outlier

contamination.

The index of contamination may appear somewhat complex to calculate at first glance. However, it can easily be calculated using a standard statistics package such as SPSS or SAS. Both the classical mean and the biweight are available options in these packages. The median absolute deviation is easily calculated by applying a few simple mathematical procedures to the median which is already generated by SPSS or SAS. Therefore, the index of contamination provides the data analyst with a single number representing the degree of contamination present in a given data set. 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. A complete discussion of the rationale and derivation of the index of contamination is provided in Zumbo (1993).

Research Questions

The specific research questions which are addressed in this thesis emanate from the literature review and are described in terms of the independent and dependent variables in the design. There are four dependent variables in the study: 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 effect sizes discussed earlier. There are four independent variables in the study. The first three independent variables are parameters of the contamination model. They were described in the literature review and are known as 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.

From the research design, three research questions are addressed. (1) How is the robustness of validity value (Type I error) affected by variations in the parameters of the contamination distribution? (2) What variables are best suited to build a useful model of Type I

error for data analysts confronted with outlier contaminated data? (3) How are the robustness of efficiency (power) values affected by variations in the parameters of the contamination distribution?

The answers to the first two research questions are expressed using Bradley's criterion for robustness of validity followed by response curve modeling (Rawlings, 1988) which in our context is equivalent to fixed effects regression modeling. Harwell (1992a, 1992b) suggests that the results of MC studies can be more readily understood by the use of regression techniques. He points out that numerous tables of values are difficult to synthesize in a meaningful manner. In addition, the use of narrative description can result in vague or ambiguous conclusions and misinterpretation of the results. He 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). The use of logistic regression techniques has also been suggested for the analysis of MC study results. However, Harwell (1992a, 1992b) has shown that the use of logistic regression techniques provides very similar results to the fixed effects regression models. Since fixed effects regression techniques are more widely understood than logistic regression techniques, these methods are used in this study. The use of fixed effects regression models for the analysis of power (research question three) is problematic and is discussed in greater detail when the results are presented. Answers to the third research question are obtained more clearly through the derivation of power curves and by using a fairly stringent criterion for robustness of efficiency. In the next section of this thesis the methodology used to obtain the Type I error and power values for these models is explained.

Methodology

Selecting Parameter Values

The first phase of the design of this study was the selection of the values for the parameters of the contamination distribution. The goal was to select these values so that a range of outlier contamination in the sample would result. The selection of the values was guided by previous

studies in which contamination models were used. 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 reasonable practical application in educational research settings. 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 example of 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.

As a result of this process 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. 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 the 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. Therefore, a total of 27 different degrees of outlier contamination have been generated in the design

of this study. Each contaminated 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. The degree of contamination increases from population 1 through to population 27 which contains the most extreme degree of outlier contamination explored in this study. For each of the 27 populations, five cells are included. Each cell represents one of the five sample sizes in the design.

In addition to these 27 contaminated populations, a normally distributed population is included in the design. The performance of each of the 27 contaminated populations is compared to the performance of the normal population throughout the design.

Table 1 Parameter Values and Resulting Population Distributions Used in the Study

Proportion	Mean Shift	n	Standard Deviation of P_c		
			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

The logical foundation to a Monte Carlo study is the generation of random numbers. Strictly speaking, computers do not generate random numbers but they can generate a series of numbers which meet accepted requirements for randomness. These numbers are more accurately called pseudo random numbers. One of the commonly available methods for generating these pseudo random numbers is termed a multiplicative congruential generator and this is the method employed in this thesis. The basic process is outlined in easily understood terms by Lehman (1977) and is based on principles first proposed by Lehmer in 1951. The process can be summarized as follows: the generator takes a starting value or seed from the computer's real time clock; this seed value is then multiplied by a constant and a modulo operation is performed using the word length of the computer in bits. This process generates a series of numbers in the range of 0-1, known as the unit rectangular distribution. The pseudo random generator has been tested to ensure that the series of numbers is of sufficient length so that sampling does not cycle through the series. The testing process has also determined that the distribution originally generated is indeed rectangular, that sequences within the series are independent and that runs and gaps of an unacceptable nature do not exist in the series.

The unit rectangular distribution created by the pseudo random number generator was transformed to a normal distribution using the Box-Muller method (1958). Each of the 27 contaminated populations was created by applying a transformation to the normal distribution. This transformation has the effect of applying 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 27 contaminated distributions and for the normal distribution. The mean, skewness and kurtosis 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 Box-Muller transformation is functioning as expected can be found in the values obtained for the normal distribution. 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 (arbitrarily set at points beyond 2.7 standard deviations) for a sample size of 15,000 would be about 104. The number of outliers for the normal population in the simulation was 99.

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 of the parameters except proportion of contamination. Thus, the number of outliers increases as the proportion of contamination increases. However, it must be noted that the number of outliers contained in contaminated distributions can only be used as 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. The standard deviation is positively biased (i.e. inflated) when contaminated populations are being explored; therefore, the number of outliers is underestimated.

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 can only be used as a crude 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 of contamination. Having established that the method of data generation is sound the next step in the methodology is to determine the Type I error rates.

Determining Type I Error Rates

In order to determine the Type I error rates of the one-sample t test for the normal distribution and the 27 contaminated distributions, the value of the hypothesized population mean, μ, is set at zero. In this case any differences which are found to be significant represent Type I errors. To clarify, the purpose of the one-sample t test is to determine if an observed sample mean, M, is different from an hypothesized population mean, μ. In this simulation study this difference is set at zero. If the results of the t test indicate a significant difference then a Type I error has been made. The actual mean of each of the 27 contaminated distributions was calculated from a sample of 30,000 observations for each population. The mean of each population was set to zero. Following this step the t test was performed.

At the beginning of the computer program which conducts the t test a counter variable is created and set at zero. This counter keeps track of the number of Type I errors. 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 because the use of such a test allows the

researcher in a practical application to identify a significant result in either direction. When a one-tailed test is used the results indicate either a significant result in the expected direction or in the case of non significance a lack of support for the research hypothesis. 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 this simulation study. The arguments in support of the use of two-tailed hypothesis testing for educational and psychological research settings are clearly outlined by Pillemer (1991).

The simplest way to explain the program which conducts the t test is to describe the process for one cell in the design. Let's begin with a sample size of eight and population 1. 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 (set at 0). If the sample mean differs significantly then a Type I error has occurred and the counter is incremented by one. 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. A total of 15 data points for each cell in the design are computed in the program. 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 15 data points were created for each cell in order to facilitate regression modeling of the results.

The accuracy of the t test 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 t test program functioned as intended for the simulation.

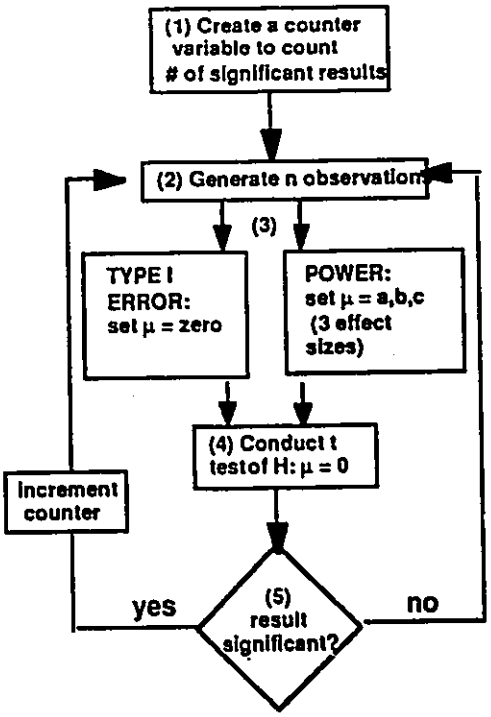
Determining Power Values

Power values were calculated for three effect sizes: small (d=.20), medium (d=.50), and large (d=.80). The effect sizes were introduced into the program by offsetting each sample value by the amount of the effect size. Conceptually, a difference between the sample mean and the hypothesized mean is created. The size of this difference is equal to the effect size being investigated. 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 effect size 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. In fact, the Type I error program and the three effect size programs were combined into one simulation algorithm to increase efficiency.

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 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 effect sizes 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 for the simulation.

The simulation algorithm is expressed as a flow chart in Figure 2. Each step of the algorithm has been described in the preceding section and the flow chart is included as a method of summarizing the program used in this study.

Figure 2. Flow chart of the simulation algorithm.



Obtaining Population Values for the Contamination Index

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 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 (2) 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. These values were then entered into the formula for the index of contamination, Equation (1). The result of this process is a single number, the contamination index, which indicates the degree of outlier contamination present in each of the populations. These values are provided as Table 2.

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

The skewness and kurtosis values obtained for samples of 15,000 values are also shown in Table 2. The accuracy of the method for calculating the contamination index was further tested in two 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 contamination index 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. This provides evidence that the program for calculating the CI values functioned as intended in the study.

Results and Conclusions

For clarity, the results of this study are presented according to each research question.

Research Question (1)

How is the robustness of validity value (Type I error) affected by variations in the parameters of the contamination distribution?

The Type I error rates are shown in Table 3. Each tabled value is the mean of the 15 data points collected for that cell. Therefore, each value in the table represents a total of 15 batches of 2000 replications of the t test (i.e. a total of 30,000 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 which are in this range are indicated in Table 3 in plain type. The moderate criterion requires values to fall between .04 and .06. Values which are in this range but fail to satisfy the fairly stringent criterion are indicated in Table 3 using the symbol †. The very liberal criterion requires values to lie between .025 and .075. Values which are in this range 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 established by Bradley. 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 one sample t test does not begin to deviate from the fairly strict criterion until population 17 in the design. This population has 8% asymmetric contamination with 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

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†**

plain type-fairly stringent †-moderate *-very liberal **- beyond very liberal

populations have asymmetric contamination at a rate of 15%. This 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 the first research question with reference to variations in the parameters of the contamination model regression modeling techniques are used.

In the first set of 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:

$$\text{TYPE I} = \%\text{CONTAM} + \text{MEAN SHIFT} + \text{STD DEV } P_C + \text{N}.$$

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). An examination of the correlation matrix (Table 4) allows for the ordering of the predictor variables in terms of their influence on Type I error rates. In order to create more parsimonious models, 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.

Table 4

Correlation Matrix for Type I Error and the Parameters of the Contamination Model

Variable	STD DEV P_C	% CONTAM	MEAN SHIFT	N
TYPE I ERROR	.139	.282	.377	-.178
STD DEV P_C	-	.000	.000	.000
% CONTAM		-	.000	.000
MEAN SHIFT			-	.000
N				-

Clearly, the variable which correlates most highly with Type I error rate is MEAN SHIFT. The correlation is positive which 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. This correlation (.282) 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. 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.

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. The four-way interaction resulted in a small increase (1%) in the variance explained. Therefore, the model including the three-way interaction terms is preferred over the model including the four-way term and accounts for 55.2% of the variance in Type I error rates.

Striving for the most parsimonious model, the three-way interactions were examined to determine if any were statistically non-significant. By examining the t values which test the significance of each variable in the model, the interaction of N*% CONTAM*STD DEV P_C was not statistically significant and was not included in the final model. The b-weights and their

associated standard errors for this final model are shown in Table 5.

Table 5

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

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. This means that the units of each variable must be considered when interpreting these values (Darlington, 1990).

The values of the b-weights must be interpreted carefully when interaction terms are included in the regression model. According to Darlington (1990) a two-way interaction means that the size of a conditional effect changes with another variable. 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 the 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 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. This effect can be seen clearly in the 8% and 15% proportions of contamination. 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 obvious 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 over batches of 2000 replications, there is some concern that a transformation of the data values might result in more meaningful results. The most appropriate transformation for this situation as suggested by Darlington (1990) is the logit transformation. 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 sets of regression models discussed above. The results obtained using the logit transformation are very close to the results obtained without the transformation. The R^2 values tended to be slightly lower (approximately 1-2%) using the logit values. Since all of the results were very similar there was no real evidence that the logit transformation was advantageous. Given that the transformed values are more difficult to interpret the remainder of the results have been expressed using only the untransformed values. Some specific comments concerning the application of logit transformations in the regression models for power are made later in the thesis.

Research Question (2)

What variables are best suited to build a useful model of Type I error for data analysts confronted with outlier contaminated data?

The set of regression models reported in the previous section does not address the second research question because a data analyst confronted with data containing outliers would have no means of determining the parameters of contamination for the data set. For this reason two additional sets of regression models have been created. The second set of regression models uses skewness and kurtosis values as the predictor variables. Skewness and kurtosis values are readily available on statistics packages and provide one method for the data analyst to characterize outlier contaminated values. Sample size was also used in this set of models. Since skewness and kurtosis are correlated the correlation matrix for these variables does not provide a good indication of the ordering of these variables in terms of importance. For this reason, all possible subsets of this regression model were calculated. Table 6 contains those models which are of interest to the data analyst. Only those models from the all possible subsets regression which include sample size as a variable have been reported. Sample size is always included in the model because this variable has conceptual importance to the data analyst in addition to the demonstrated correlation between this variable and the Type I error rate (as shown in Table 4).

Table 6

R-squared Values for the Regression of Skewness and Kurtosis on Type I Error

Model	R^2
Type I = n+ constant	.0318
Type I = n + skew + constant	.3095
Type I = n + kurt + constant	.0865
Type I = n + skew + kurt + constant	.3593

The R^2 values for these models can be used to determine both variable ordering and the model which would be most useful for the data analyst. The single most important variable accounting for variance in Type I error in these models is skewness. Kurtosis contributes little to the overall variance explained in the model. A data analyst could potentially account for about 30% of the variance in Type I error rate using only skewness and sample size. However, given the difficulties in using skewness coefficients which were identified by Horswell and Looney (1993) and discussed earlier in this thesis, this model is not recommended. These authors show that skewness and kurtosis coefficients when used jointly may provide a better method of assessing normality. In any case kurtosis values are available at no additional cost and without additional effort on the part of the data analyst. Accordingly, while there is no cost for including kurtosis in the model there may be a considerable cost for omitting this variable. For this reason, the full model accounting for 35.93% of the variance in Type I error rates is the preferred choice from this set of regression models. This model 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 set of models 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 complete 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. Furthermore, this model with the two-way interaction term accounts for approximately 41% of the variance and may be of use to the data analyst.

A third set of models was computed for the Type I error values. This set of models 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 b-weights could not be used for the skewness and kurtosis values because these variables are correlated. 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 CI are functioning as intended. The use of the contamination index accounts for a greater amount of variance in Type I error rates than does the use of skewness and kurtosis in the previous set of models.

As with the previous two sets of models, 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 b-weights and associated standard errors for this model are shown as follows,

$$\underset{\quad}{\text{TYPE I}} = \underset{(.0003)}{.0489} + \underset{(.000004)}{.00001}*N + \underset{(.0021)}{.0813}*CI - \underset{(.00003)}{.00057}(N*CI)$$

These b-weights are interpreted in the same manner as the previous sets of models. 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.

Research Question (3)

How are the robustness of efficiency (power) values affected by variations in the parameters of the contamination distribution?

The results from the power portion of the simulation are recorded in Tables 7, 8, and 9. 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 values for these cells are shown in Appendix A but should be interpreted with caution. 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 have been very difficult to interpret because of the empty cells in the data set. In addition, odd interactions may have shown up as a result of the pattern of empty cells. These interactions would be difficult to interpret. For these reasons, the first two sets of regression models which were created for Type I error (the parameters of contamination model and the skewness and kurtosis model) were not computed for the power results.

Table 7 Power Values for Each Population Distribution Under Study (Small Effect Size)

			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
.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%

Table 8 Power Values for Each Population Distribution Under Study (Large Effect Size)

			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	-

* - power is above normal by > 10% † - power is below normal by >10%

Table 9 Power Values for Each Population Distribution Under Study (Large Effect Size)

			Standard Deviation of P_c		
Proportion	Mean Shift	n	0.5	1.75	3.0
.0[illegible]	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	-

* - power is above normal by > 10% † - power is below normal by >10%

Table 10

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

However, one set of regression models was investigated for the power results. This set of models used the contamination index and sample size as the predictor variables. To avoid the difficulties created by inflated Type I error rates, this set of models included power values from populations with a CI value less than 0.20. Therefore, populations 25, 26 and 27 were not included in the regression models. The CI value of 0.20 was chosen by examining Tables 2 and 3 concurrently. It was noted that Type I error rates are inflated beyond Bradley's very liberal criterion for populations with a CI greater than 0.20. Interestingly, populations 16, 17, and 18 have CI values close to 0.20 and the Type I error rates for these populations satisfy only the very liberal criterion for robustness of validity. The skewness and kurtosis values did not lend themselves to such a clear demarcation. Therefore, power models were not developed using skewness and kurtosis values.

One set of models was created for each effect size value investigated in the study (i.e. small, medium, and large). The R^2 values are summarized in Table 11.

Table 11

R-squared Values for the Regression of the Contamination Index on Power at Small, Medium, and Large Effect Sizes

Model	R^2
Power Small ES = Constant + CI + N	.9885
Power Small ES = Constant + N	.9859
Power Medium ES = Constant + CI + N	.6579
Power Medium ES = Constant + N	.6578
Power Large ES = Constant + CI + N	.3744
Power Large ES = Constant + N	.3726

Two interesting results are found in Table 11. First, it is clear from the comparison of R^2 values for the models which contain only the constant and N, that sample size is the variable which accounts for almost all of the variance in power in these models. The CI variable was insignificant when included in the model on its own. This is true for all three effect sizes. The CI accounts for very little variance in the power values when Type I error is protected. This regression model is of no practical use since the influence of sample size on power is well known and documented (e.g., Cohen, 1977). The second interesting result shown in Table 11 is the decrease in R^2 values as the effect size value increases. This result is easily explained by a quick examination of the tabled power values. As the effect size increases, the number of tabled power values which approach or actually reach 1.0 increases. This trend is logically consistent with the nature of power and effect sizes. That is, the power to detect a large effect size should be considerably greater than the power to detect smaller effect sizes. The effect of this trend on the regression model is to reduce the R^2 value because the amount of variability in the results is considerably reduced. This is why the R^2 values for the regression model decrease as the effect size increases.

As with the earlier regression models for Type I error, a logit transformation of the power values was conducted and regression models for these transformed values were calculated. The trend of decreasing R^2 values as effect size increases was not observed for the logit power models.

The reason for this is clear. The logit transformation process effectively stretches the scale of the power values. The power values in Tables 7, 8, and 9 are all in the interval of 0 to 1. As the effect size increases, many of the tabled values approach or equal 1.0. As stated earlier this effectively reduces the variability in the results and lower R^2 values are found for the regression models. This problem does not occur for the logit transformed values. The scale is stretched beyond 1.0 and greater variability in the results is permitted. The observed R^2 values for the transformed power results were all .95 or greater. However, as noted for the untransformed power models, sample size accounted for almost all of the variance in power values at all three effect sizes.

In conclusion, regression modeling of the power values is not a satisfactory method of analysis. The only variable which is shown to be important is sample size. The relationship between sample size and power is already well known and documented. For this reason, the analysis of power results was undertaken using a method similar to Bradley's robustness of validity criterion and then using power curves.

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 7, 8, and 9. An asterisk (*) indicates that the power value for the contaminated population exceeded the normal value by more than 10%. A (†) symbol indicates the power value for the contaminated population was below the normal value by more than 10%.

An examination of Tables 7, 8, and 9 reveals a number of interesting trends. For the small effect size (Table 7) 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 10). 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 effect size (Table 8) only two cells are beyond the fairly stringent criterion for robustness of efficiency. This means that robustness of efficiency is greater for medium effect sizes than for small effect sizes. Both of the cells in the medium effect size 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 effect size (Table 9) 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.

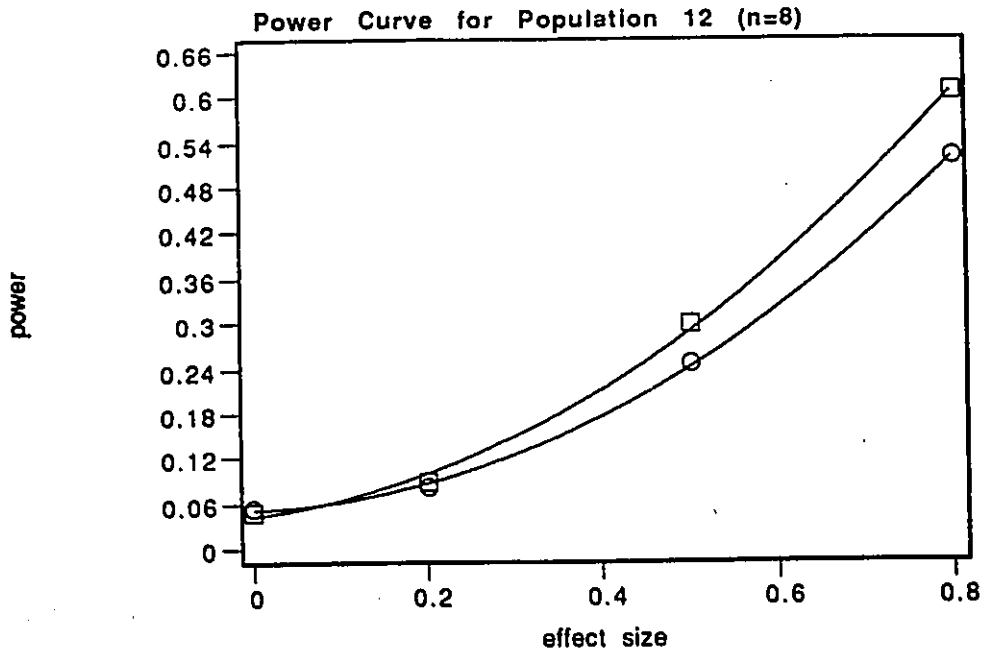
The results is Tables 7, 8, and 9 can be summarized with a few general statements. The robustness of efficiency of the one-sample t test decreases as the effect size becomes smaller. At small effect sizes asymmetric contamination results in a power loss. Paradoxically, symmetric contamination results in a power advantage. For the medium and large effect sizes 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.

Power Curves

A third form of analysis was applied to the power data in order to better understand the results. The power values have been converted into power curves using the program MacCurveFit (Raner, 1993). Only those cells in the design which satisfied the fairly stringent criterion for Type I error rate were included in the calculation of power curves. The reason for this decision, as stated earlier, is that when the Type I error rate is not maintained the obtained values are not actually power values. The first step in calculating these power curves was to enter the Type I error rate and the power values for the small, medium and large effect sizes (from Tables 3, 7, 8, & 9) into the MacCurvefit program. These values were plotted on the Y axis with the effect size plotted on the X axis. A second order polynomial curve was fit to the data values using the equation $Y = ax^2 + bx + c$. The program calculates an R^2 value to show the degree of model fit between the second order polynomial and the data values. In all cases this value exceeded .94 and in the majority of cases the R^2 value was .99. These R^2 values provide evidence that a second order polynomial is the appropriate function to describe the power curves in this study.

A power curve for each cell was plotted on one graph together with the normal power curve for the same sample size. Figure 3 is an example of one pair of power curves for population 12 at a sample size of 8.

Figure 3. Normal and contaminated power curves for population 12 (n=8).



The curve denoted by squares is for the contaminated distribution. The curve denoted by circles is for the normal distribution at a sample size of 8. The advantage of plotting the two curves together on one graph is that it permits an immediate comparison of the power under normal and contaminated populations. For this example the contaminated distribution has a power advantage at every effect size beyond approximately 0.1.

A visual inspection of these graphs provides some insight into how the power values for the contaminated distributions compare to the normal. However, it is more useful to find a means of quantifying this difference. The polynomial expression $Y = ax^2 + bx + c$ can be used to quantify the power difference. The a, b, and c values from the polynomial expression were recorded for each contaminated cell in the design. The values for the normal distribution are referred to as d, e, and f, respectively. The area between the power curve for the contaminated population (a,b,c) and

the power curve for the normal population (d,e,f) for a given sample size is then calculated. The calculation of this area is achieved by determining the integral of the two functions using formula (3).

$$q(x) = \int_0^{0.8} r(x) - n(x)dx$$
$$= \int_0^{0.8} (ax^2 + bx + c) - (dx^2 + ex + f)dx$$
$$= \frac{(a-d)}{3}0.8^3 + \frac{(b-e)}{2}0.8^2 + (c-f)0.8 \quad (3)$$

This integral is evaluated for the interval from 0 to 0.8 because this is the range of effect sizes for which power values were obtained in the study. For example, for population 12 at a sample size of 8 the values for a, b, and c obtained from the MacCurveFit program are .705536, .139940, and .0446441 respectively. The values for d, e, and f refer to the obtained values for the normal distribution at a sample size of 8 and are .681490, .0392235, and .0527365 respectively. By substituting these values into equation (3) the area between the two power curves is determined. For this example the area is .02985921.

This method is demonstrated by Keller-McNulty and Higgins (1987) using values from a study of robust permutation tests for location. These authors report the area between the two power curves directly. However, the magnitude of these area values is not easily interpreted. For this reason the ratio of the area between the normal and contaminated power curves is compared to the total area below the normal power curve and then expressed as a percent. These values are shown in Table 12. Positive values represent a gain in power for the contaminated population over the normal distribution. Negative values represent a power loss.

Table 12 Power Difference Relative to the Normal Distribution Expressed as a Percentage

			Standard Deviation of P_c		
Proportion	Mean Shift	n	0.5	1.75	3.0
.01	0	8	1.13	2.46	-
		16	-0.72	1.96	3.35
		32	0.53	2.33	2.46
		64	0.20	1.50	1.29
		128	0.38	1.68	1.49
	1.5	8	-0.11	-2.57	2.24
		16	-1.28	-2.45	1.65
		32	-0.05	-0.80	2.08
		64	-0.19	-0.98	0.94
		128	-0.14	-1.03	0.81
	3.0	8	-1.86	-3.52	3.17
		16	1.84	-3.33	3.77
		32	1.22	-1.69	3.30
		64	1.21	-2.10	1.52
		128	1.22	-2.36	1.74
.08	0	8	0.12	8.00	**17.46**
		16	-2.38	4.78	7.49
		32	-1.56	3.34	1.17
		64	-1.24	2.66	0.52
		128	-1.30	2.39	0.02
	1.5	8	-	-4.40	2.59
		16	-2.85	-0.58	3.42
		32	-0.97	1.67	1.90
		64	-0.57	0.51	-0.43
		128	-0.87	1.01	-0.33
	3.0	8	-	-	-
		16	-	-	-
		32	1.93	-	-
		64	0.85	-	-
		128	1.09	-	1.20
.15	0	8	4.30	2.70	-
		16	0.19	0.11	-
		32	0.84	-0.85	-1.06
		64	0.07	-0.58	-1.68
		128	0.02	-0.70	-1.79
	1.5	8	-2.25	-	2.59
		16	-2.19	-	-
		32	-0.31	-0.65	-
		64	-0.15	-1.82	-
		128	-0.36	-	-
	3.0	8	-	-	-
		16	-	-	-
		32	0.10	-	-
		64	-0.57	-	-
		128	-0.71	-0.28	-

+ values indicate power advantage - values indicate a power loss

One minor difficulty was experienced in the calculation of the integrals whenever the two power curves crossed. The area between the curves which is outside the point of crossing is subtracted from the area between the curves. This may alter the accuracy of the power difference being calculated. Since the power curves in this study cross in less than 10% of the cases and the area beyond the point of crossing is typically very small, this difficulty is noted but is of no major conce.... Keller-McNulty and Higgins (1987) also show crossed power curves in their study. However, they do not discuss what procedure, if any, was used to address this problem.

An examination of Table 12 can lead to some general conclusions. In only one cell is a power advantage greater than 10% noticed. This cell is population 12 at a sample size of 8. Most of the remaining cells in the table reflect 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. Table 12 does not indicate the effect sizes. The examination of power curves such as Figure 3 allows only a rough visual examination of the power differences at each effect size. This is unsatisfactory given the importance of considering effect size 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 in determining how power values are affected by variations in the parameters of contamination.

Discussion

In the introduction to this thesis four aspects of the present Monte Carlo study were described as expansions into new territory. The results concerning the robustness of validity and efficiency of the one-sample t test are discussed within the framework of these four novel areas. The first novel area introduced in this thesis is the systematic range of nonnormal populations which were generated using the contamination model. Previous researchers explored a very limited range of nonnormal distributions and focused on either robustness of validity or robustness of efficiency. These researchers typically used methods of data generation which are more relevant to investigations of truly nonnormal underlying population distributions. The use of contamination models in the present study allowed for a much more panoramic view of the factors which influence the robustness of the one-sample t test to outlier contamination.

Specifically, the results indicated that the 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 one-sample 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 one-sample t test cannot be determined. If robustness of validity is intact then power values are maintained when medium and large effect sizes 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 effect sizes. For medium and

large effect sizes power differences are noted only for sample sizes of 8 and 16. However, when small effect sizes are being investigated, the power values are not as expected. Specifically, at small effect sizes, asymmetric contamination results in a power loss. Again, paradoxically, symmetric contamination results in a power advantage over the normal distribution for these small effect sizes. 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 Monte Carlo studies of robustness of efficiency is an important factor in developing a full understanding of these relationships.

Contamination models provide an excellent method for investigating robustness in Monte Carlo studies. Both symmetric and asymmetric contamination of varying degrees can be readily simulated. This is an important asset of this methodology in light of the results observed for the one-sample t test. That is, both robustness of validity and robustness of efficiency functioned differently under conditions of symmetric versus asymmetric contamination. In addition, the identification of complex interactions is made possible using the parameters of the contamination model. This advantage is more thoroughly examined in the discussion of the fixed effects regression modeling techniques. One final advantage to the use of contamination models should be mentioned; these models facilitate the replication of Monte Carlo studies and also 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.

The second novel area presented in this thesis is the use of a contamination index proposed by Zumbo (1993). The use of the 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 one-sample 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 CI. Second, since CI and sample size are uncorrelated the model can be more easily interpreted than the skewness and kurtosis model. The third advantage is that 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 application of this proposed method for quantifying contamination.

For the educational 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.

The third front which this thesis expands upon is in the examination and expression of results through regression modeling (i.e. response curve modeling). Previous Monte Carlo studies in this area have principally relied upon tabulation and narrative description in the presentation of results. The results of this thesis were analyzed extensively using regression. Regression

modeling is of limited use in the analysis of robustness of efficiency for two reasons. First, regression modeling could not be used with the parameters of the contamination model or with the skewness and kurtosis values 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. In an attempt to overcome this difficulty, regression models for power were examined using the contamination index. The empty cells were removed from the design by including only those populations with a CI value less than 0.20. The results of these regression models demonstrated that sample size accounted for nearly all of variance in power values. If an impact of contamination on power values exists, it could not be discerned. This is the second problem encountered when regression techniques are used for the analysis of power values. The relationship between sample size and power is already well known and documented. Thus, the use of regression modeling for power results does not contribute any new information to the field of robustness.

Regression modeling was much more useful in the examination of 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 Monte Carlo study is that the models guided our interpretation of Type I error rates. Clearly, the three-way interactions which became evident through modeling were not readily discernible from the tabled values and their narrative description.

The regression model which resulted from this process is informative for social science methodologists seeking a better understanding of the performance of the one-sample t test. However, this model is of little use to data analysts confronted with outlier contaminated data

because the parameters of contamination cannot be determined. Two sets of models of practical relevance to data analysts were explored: the skewness and kurtosis models and the contamination index models. The use of the CI model is recommended and the reasons for this choice have already been outlined.

One limitation 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 not much variance in the results of this Monte Carlo study. This may be true in other Monte Carlo 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. For example the CI model including interaction terms accounts for about 48% of the variance observed. It is difficult to determine the meaningfulness of this amount of variance.

The fourth new area explored in this study is the quantification of the concept of robustness of efficiency. 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 thesis. This criterion provided the most useful method of summarizing the power results in this study. The limitations of regression modeling of the results for power have already been discussed. Power curves have been used in some of the published Monte Carlo studies reviewed in this thesis. However, the use of power curves as a method of summarizing results is problematic because the importance of effect size in assessing power cannot be discerned from these curves. The fairly stringent criterion for robustness of efficiency proposed in this thesis requires the power difference between the contaminated population and the normal population to be within + or - 10% of the normal value.

This study provides further support for the apparent sensitivity of normal theory tests to the

asymmetry of the distribution. Harwell and Serlin (1989) state that there is a difficulty in checking the normality assumption of normal theory tests. However, herein is presented a possible route to alleviate this difficulty, the contamination index. The challenge is to conduct further research investigating the performance of this index.

References

Blair, R.C. & Higgins, J.J. (1980). The power of t and Wilcoxon statistics: A comparison. Evaluation Review, 4(5), 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(4), 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(5), 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(4), 275-278.

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.

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

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155-159.

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

Harwell, M.R. (1992a). Analyzing and reporting Monte Carlo results in methodological research

in education and psychology . Manuscript submitted for publication.

Harwell, M.R. (1992b). Summarizing Monte Carlo results in methodological research. Journal of Educational Statistics, 17(4), 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., Rubinstein, E.N., Hayes, W.S. & Olds, C.C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. Journal of Educational Statistics, 17(4), 315-339.

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

Hays, W.L. (1973). Statistics for the social sciences. New York, NY: Holt, Rinehart and Winston.

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(2), 437-459.

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

Keller-McNulty, S. & Higgins, J.J. (1987). Effect of tail weight and outliers on power and Type I error of robust permutation tests for location. Communications in Statistics :Simulation and Computation . 16(1), 17-35.

Lehman, R.S. (1977). Computer simulation and modeling. Hillsdale, NJ: Lawrence Erlbaum.

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.

Ott, L., Larson, R.F. & Mendenhall, W. (1983). Statistics: a tool for the social sciences. Boston, MA: Duxbury Press.

Pillemer, D.B. (1991). One- versus two-tailed hypothesis tests in contemporary educational research. Educational Researcher, 20(9), 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(4), 505-510.

Rawlings, J.O. (1988). Applied regression analysis: A research tool. Belmont, CA: Wadsworth & Brooks.

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(2), 352-360.

Sawilowsky, S.S., & Hillman, S.B. (1992). Power of the independent samples t test under a prevalent psychometric measure distribution. Journal of Consulting and Clinical Psychology, 60(2), 240-243.

Shavelson, R.J. (1988). Statistical reasoning for the behavioral sciences. Needham Heights, MA: Allyn and Bacon.

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(6), 1055-1098.

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.

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. (1993). Development and testing of guidelines for the implementation of the Hogg-Tukey procedure: The one-sample location problem. Working Paper 93-01 of the Edumetrics Research Group, University of Ottawa.

Zumbo, B.D., & Zimmerman, D.W. (1993). Alternatives to classical statistical procedures. Canadian Psychology, 34, 365-367.

Appendix A

Table 13 Complete Power Table for Small Effect Size

			Standard Deviation of P_C		
Proportion	Mean Shift	n	0.5	1.75	3.0
.01	0	8	.085333	.087367	*.092833*
		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	*.082600*	**.067733†**	**.064667†**
		16	.112367	**.105533†**	**.094800†**
		32	.186633	.187267	**.169667†**
		64	.340833	.352600	.329533
		128	.594733	.637333	.605533
	3.0	8	*.063533*	*.050533*	*.049500*
		16	*.097400*	*.079700*	*.072533*
		32	.177833	*.148367*	*.142433*
		64	.355467	*.317167*	*.328900*
		128	.638867	*.599900*	.640433
.15	0	8	.086700	.082500	*.085433*
		16	.121567	.120667	*.129167*
		32	.200667	.189567	.199133
		64	.351567	.344000	.335300
		128	.614700	.598000	.573133
	1.5	8	.085133	*.060100*	**.055300†**
		16	.118033	*.086200*	*.084300*
		32	.194467	**.158667†**	*.161667*
		64	.348900	**.304333†**	*.330967*
		128	.605233	*.570767*	*.611033*
	3.0	8	*.059367*	*.051033*	*.040933*
		16	*.089100*	*.072267*	*.059033*
		32	**.164167†**	*.145900*	*.129567*
		64	.326433	*.312433*	*.313800*
		128	.597233	.606600	*.628067*

* - power is above normal by > 10% † - power is below normal by >10%

Table 14 Complete Power Table for Medium Effect Size

			Standard Deviation of P_c		
Proportion	Mean Shift	n	0.5	1.75	3.0
.01	0	8	.249500	.252867	*.262833*
		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	*.238700*	.234433	.251500
		16	.456733	.473067	.508833
		32	.774133	.803967	.822600
		64	.975100	.982967	.985800
		128	.999967	.999933	.999867
	3.0	8	*.211067*	*.208200*	*.234133*
		16	*.466133*	*.475633*	*.505333*
		32	.815867	*.831200*	*.864833*
		64	.985867	*.989400*	*.994467*
		128	1.00000	*1.00000*	1.00000
.15	0	8	.260867	.255667	*.315600*
		16	.474667	.475733	*.526933*
		32	.785067	.774233	.768133
		64	.976700	.972333	.961600
		128	.999900	.999800	.999633
	1.5	8	.238500	*.214933*	.251767
		16	.459200	*.456700*	*.516200*
		32	.776067	.798433	*.835367*
		64	.975167	.983600	*.986033*
		128	.999867	*.999900*	*1.00000*
	3.0	8	*.190633*	*.187433*	*.208200*
		16	*.446633*	*.461367*	*.503567*
		32	.803167	*.835333*	*.877667*
		64	.986033	*.991633*	*995833*
		128	.999967	1.00000	*1.00000*

* - power is above normal by > 10% † - power is below normal by >10%

Table 15 Complete Power Table for Large Effect Size

			Standard Deviation of P_c		
Proportion	Mean Shift	n	0.5	1.75	3.0
.01	0	8	.521733	.526967	*.542233*
		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	*.510267*	.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	*.533833*	*.549067*	*.604233*
		16	*.892233*	*.914733*	*.931933*
		32	.997300	*.999100*	*.999133*
		64	1.00000	*1.00000*	*1.00000*
		128	1.00000	*1.00000*	1.00000
.15	0	8	.531733	.535033	*.618300*
		16	.850533	.846800	*.838433*
		32	.991467	.990167	.978267
		64	1.00000	1.00000	.999767
		128	1.00000	1.00000	1.00000
	1.5	8	.504833	*.529867*	**.610100***
		16	.843500	*.884533*	*.899967*
		32	.992100	.996167	*.995033*
		64	1.00000	1.00000	*.999967*
		128	1.00000	1.00000	*1.00000*
	3.0	8	*.503900*	*.547567*	*.629967*
		16	*.891500*	*.928233*	*.952833*
		32	.998100	*.999433*	*.999767*
		64	1.00000	*1.00000*	*1.00000*
		128	1.00000	1.00000	*1.00000*

* - power is above normal by > 10% † - power is below normal by >10%