Journal of Modern Applied Statistical Methods

Volume 1 | Issue 1 Article 12

5-1-2002

Hotelling's T^2 VS. The Rank Transform With Real Likert Data

Michael J. Nanna
Arizona State University

Journal Of Modern Applied Statistical Methods
Winter 2002, Vol. 1, No. 1, 83-99

1538 - 9472/02/$30.00

Hotelling's T^2 VS. The Rank Transform With Real Likert Data

Michael J. Nanna
Department of Educational Leadership and Policy Studies
Arizona State University

Monte Carlo research has demonstrated that there are many applications of the rank transformation that result in an invalid procedure. Examples include the two dependent samples, the factorial analysis of variance, and the factorial analysis of covariance layouts. However, the rank transformation has been shown to be a valid and powerful test in the two independent samples layout. This study demonstrates that the rank transformation is also a robust and powerful alternative to the Hotellings T^2 test when the data are on a likert scale.

Keywords: Robustness, Statistical power, Parametric tests, Conditionally distribution-free tests, Multivariate tests

Introduction

Hotelling's T^2 is a specific case of multivariate analysis of variance (MANOVA) having one independent variable with two levels and multiple dependent variables. It is the multivariate analogue of the independent samples t-test. Instead of examining differences in means between two samples on one dependent variable, Hotelling's T^2 tests for the equality of centroids (mean vectors) between two groups. T^2 is a parametric test having the usual assumptions that theoretically must be met for its valid application. For example, T^2 assumes that the data of the underlying distributions is normally distributed, (i.e., assumes multivariate normality), that there is independence of observations, and that the variance/covariance matrices are equal.

The Rank Transform (RT) is a set of nonparametric-like (and sometimes referred to as conditionally distribution-free) procedures. The RT is performed by replacing original observations with their respective ranks, computing a parametric statistic on these ranks, and then referring the test statistic to the usual table of critical values (Conover & Iman, 1981). It has been suggested that RT procedures have robustness and power properties similar to their parametric counterparts when normal theory assumptions are met, and have superior robustness and power properties when assumptions are not met (Conover & Iman, 1981).

Purpose of Current Study

The purpose of this study is to examine the robustness and power properties of the two independent sample Hotelling's T^2 test and the RT using Monte Carlo techniques with samples drawn from real data sets that are based on ordinal level likert scaled type data. An example of a common five-point likert scale is 1 "Strongly Disagree", 2 "Disagree", 3 "Neutral", 4 "Agree", and 5 "Strongly Agree". The discrete nature of the response set precludes the normality distribution assumption from being met.

Relevance to Social and Behavioral Science Research

Using Monte Carlo methods, it is possible to observe the operating characteristics of a statistic under real situations. This is important because statistics frequently support and/or drive decisions and policy in real applied settings such as education, psychology, medicine, and other social and behavioral science disciplines. Using inappropriate (or less efficient) statistics can lead to "analyses that are less powerful, and potentially to inferences that are invalid" (Hunter & May, 1993, p. 386).

The implications are striking as examination of statistical power in clinical and applied research settings has consistently demonstrated low statistical power. The probability of committing a Type II error has been shown to reach .91 for detecting small effect sizes in applied data analysis (Cohen, 1962; Cohen, 1977; Sedlmeir & Gigerenzer, 1989; Matyas & Ottenbacher, 1993).

A discipline marked by studies with low statistical power means that researchers are unable to detect treatment effects that otherwise might provide the basis for developing more effective programs or interventions. Furthermore, strict reliance on statistical significance testing without consideration of robustness and power can lead to erroneous interpretation of research findings as researchers may decide to not follow up statistically insignificant results (Keppel, 1975). Lack of consensus and contradictory findings in the literature may also preclude a discipline from establishing a useful and dynamic body of knowledge (Ottenbacher, 1995). Issues of robustness and power are not theoretical problems, but rather pragmatic issues having important consequences in applied settings.

Michael Nanna is a faculty associate in the division of Social and Philosophical Foundations within the Dept. of Educational Leadership and Policy Studies at Arizona State University.

Hotelling's T^2

Many studies on Hotelling's T^2 indicated it has acceptable Type I error rates under normal and various non-normal conditions (Jensen, 1982; Mardia, 1975; Kariya, 1981; Everitt, 1979; Harwell & Serlin; 1995; Blair, Higgins, Karniski, & Kromrey; 1994; Hopkins & Clay; 1963; Algina & Oshima; 1990; Ito & Schull; 1964; Holloway & Dunn, 1967). Other studies, however, have shown T^2 displays inflated Type I error under asymmetry, even when sample sizes are relatively large (Chase & Bulgren, 1971; Serlin & Harwell, 1989; Utts & Hettsmanperger, 1980; and Everitt, 1979).

Hakstian, Roed & Lind (1979) offer the following summary concerning the robustness of T^2:

1. "The HT^2 procedure is generally robust with respect to violation of the homogeneity of covariance matrix assumption for equal sample sizes, even when the ratio of sample size to number of dependent variables is small.

2. "For $n_1 \neq n_2$, however, the test moves rapidly towards unacceptable Type I error rates as the degree of population covariance matrix heterogeneity is increased.

3. "The T^2 procedure is not robust in the face of covariance matrix heterogeneity coupled with unequal n's, even for relatively mild departures from equality of the covariance matrices, sample sizes, or both." (p. 1,261)

Rank Transformation

The RT is a set of procedures that substitutes the ranks of data for the raw data values and then calculates a usual parametric statistic on the ranked data. It has been suggested that rank tests provide a useful alternative method of analysis when the assumptions of parametric tests (i.e., t and F) are not met. Initial simulation results indicated that the RT's robustness and power properties are similar to its normal theory counterpart when assumptions are met, and are often superior when assumptions are not met (Conover & Iman, 1981).

Conover and Iman (1981) suggested the use of the RT as a bridge between parametric and nonparametric statistics for many different data analysis situations. Subsequently, however, considerable journal space has been given to the RT's lack of robustness and power for the two dependent samples, factorial analysis of variance, and factorial analysis of covariance layouts, so there is no need to review those results here. The RT has been shown, however, to be a robust and powerful alternative to its parametric counterpart in the context of both the two independent samples and the one-way analysis of variance layouts.

In terms of the multivariate two independent samples layout, Bhattacharyya, Johnson and Neave (1971), Tiku and Singh (1982), and Nath (1982) examined the robustness and comparative power properties of T^2 and the rank transformed T^2 (F_R) and found both statistics to be robust for data samples from a uniform distribution. However, Type I error rates for both statistics exceeded that of nominal alpha under the exponential and lognormal distributions (Everitt, 1979). Similar results were also found by Nath and Duran (1983). Zwick (1986) demonstrated that the RT is robust and often more powerful than the F test, but did not recommend its routine application due to its highly specific and volatile behavior. These studies, however, were limited to artificial distributions.

Scaling: Ordinal Measurement

The robustness and power of a statistic using data which is scaled at the ordinal level of measurement is important given the number of measures that exist in rehabilitation medicine, psychology, and education use ordinal data – particularly Likert scales. There are numerous studies that address the issue of scales of measurement and its impact on a statistics performance but are beyond the scope of this paper. For a review on this issue, consult Anderson (1961); Boneau (1961); Senders (1958); Siegal (1956); Stevens (1946, 1951); Hsu and Feldt (1969); Heeren and D'Agostino (1987); Nanna & Sawilowsky, (1998); Siegal, (1956); Stevens, (1946); Gaito, (1986); Lord, (1953); and Zumbo & Zimmerman, (1991).

Indeed, measurement issues have been debated in the statistics and measurement literature for decades in the context of the "weak measurement vs. strong statistics" controversy. On the basis of considerable simulation evidence (see, e.g., Sawilowsky, 1990; Hunter & May, 1993; Zumbo & Zimmermen, 1993; and Sawilowsky, 1993), additional discussion on this issue will be dismissed from consideration in choosing between parametric and nonparametric tests.

There is a paucity of research on the properties of statistics applied to likert scaled data. The independent samples t-test was shown to be robust with respect to Type I errors in simulation studies conducted by Heeren & D'Agostino (1987) and Hsu & Feldt (1969). These results were replicated, and extended in terms of statistical power, for both the t-test and its rank transformation counterpart by Nanna & Sawilowsky (1998) for likert scaled data.

The likert scaled data used for this study was obtained from the Functional Independence Measure (FIM), which is one of the most widely used assessment instruments in medical rehabilitation. In fact, "about 60% of rehabilitation facilities nationwide use the FIM" (Stineman, et al., 1996, p. 1101; see also Granger, et al., 1986). It was developed to provide uniform assessment of severity of

disability and to specify medical rehabilitation outcomes. (e.g., physical, cognitive and social variables associated with disability.) It is an 18 item assessment tool comprised of multiple, 7-point likert scales which consist of scores which range from complete dependence (1) to complete independence (7). The scale was originally designed so that ratings on all 18 items were summed into a single score that was then used to estimate overall burden of care (Stineman, et al. 1996). Total FIM scores range from 18 (complete dependence) to 126 (complete independence).

FIM scores are based on the observation of a patient meeting specific objective physical and behavioral criteria and are usually rated by clinical observation at the time of admission, and again prior to discharge from rehabilitation services. It is intended to measure levels of disability regardless of the underlying pathological condition and is considered independent of the rater's clinical background (Brynes & Powers, 1989; Keith, et al., 1987; Granger, et al., 1990). Previous studies have indicated high levels of instrument reliability (r = .95, Byrnes & Powers, 1989) and interrater agreement (.93 & .97, Hamilton, et al., 1991). Ottenbacher, et al., (1996), concluded that the FIM "provided good interrater reliability across a wide variety of raters with different professional backgrounds and levels of training. The median interrater reliability value was .95 and was based on a large cumulative sample of patients representing a wide variety of disability levels and medical conditions" (p. 1230).

Methodology

Monte Carlo techniques were used to independently sample from both admit and discharge data sets of several FIM score distributions. As suggested by Micceri (1989), real data sets were used to model admit and discharge populations (as opposed to using mathematically convenient distributions) associated with each FIM score.

The FIM scores used in this study were obtained by evaluating 903 geriatric patients admitted to a large Mid-Western rehabilitation facility from 1991 to 1995. Patients were evaluated using the FIM at the time of admission, and again at the time of discharge. Seven (1, 3, 4, 5, 6, 7, 13) of the eighteen individual admit and discharge FIM score distributions were selected for further study and are depicted in Figures 1-7. The histograms of the remaining FIM score distributions were similar to these seven, and for parsimony, are not addressed in the current study.

Differences obtained from independently sampling with replacement from the pre-test (admit) and then post-test (discharge) scores were used to represent gain due to treatment interventions (i.e., real treatment effects) as opposed to artificially modeled treatments (i.e., adding a constant to the initial distribution of scores to model a shift in location parameter) as has been done in previous studies. Although no formal effect sizes are calculated for this study, differences between the admit and discharge populations may serve as evidence of the presence of effect sizes. Because the distributions used in this study are meant to represent the actual populations, effect sizes are implicit in the difference between the means of the admit and discharge populations (see Table 1.). Therefore, sampling directly from the population distributions obviated the need to model synthetic treatment effects in this Monte Carlo study. Upon admission to the hospital, patients participated in multiple treatment regimens during their stay at the hospital before being discharged. The treatment interventions are presumed to account for the difference in means between admit and discharge distributions.

A subscale score was constructed for a separate study and, for this study, was used to model data scaled at a more continuous level of measurement. This subscale distribution, named OT, is a composite score comprised of seven FIM items that measure domains commonly assessed by occupational therapists. A list of descriptive statistics (e.g., mean, standard deviation, skew, and kurtosis) for each admit and discharge distribution used in this study can be found in Table 1.

A Fortran program was written for the IBM compatible Pentium PC accessing IMSL (1987) subroutines to sample with replacement for sample sizes $n_1 = n_2 = (5,15)$, (10,10), (10,20), (15,15), (15,45), (30,30), (25,75), and (50,50) for the number of dependent variable combinations of 2 and 5. In order to insure the validity of the Monte Carlo simulation, samples were initially taken from a multivariate normal distribution to examine actual Type I error rates produced under normality.

The first portion of this study consisted of examining the robustness of both T^2 and the RT in both two-dependent and five-dependent variable combinations. The robustness portion of the study was performed by independently sampling with replacement from each FIM admit distribution using different sample size and dependent variable combinations, calculating the appropriate statistic, and comparing obtained alpha with nominal alpha levels. A similar technique was employed for the five dependent variable layout. For five dependent variables, additional runs were performed when sampling from each individual distribution five times (e.g., FIM 7 distribution serving as each dependent variable). A subset of the results are presented in this paper and are representative of the remaining results. A copy of the remaining results for dependent variable combinations are available from the author.

The second portion of this study consisted of examining the comparative power of T^2 vs. the RT. This was accomplished by independently sampling, with replacement, from FIM admit distributions, and then comparing that with a sample taken from the related discharge

Table 1. Descriptive Statistics for Admit and Discharge Fim Distributions.

Distribution	Mean	Std. Dev.	Kurtosis	Skew
Fim 1 Admit	**5.37**	**1.50**	**.69**	**-1.00**
Fim 1 Disch.	5.93	1.30	3.63	-1.79
Fim 3 Admit	**3.64**	**1.33**	**-.22**	**-.05**
Fim 3 Disch.	4.81	1.47	-.02	-.84
Fim 4 Admit	**4.40**	**1.47**	**.09**	**-.49**
Fim 4 Disch.	5.25	1.48	.61	-.96
Fim 5 Admit	**3.19**	**1.44**	**-.51**	**.29**
Fim 5 Disch.	4.45	1.66	-.73	-.58
Fim 6 Admit	**3.87**	**1.61**	**-1.01**	**-.17**
Fim 6 Disch.	4.96	1.71	-.22	-.92
Fim 7 Admit	**4.32**	**1.88**	**-1.07**	**-.41**
Fim 7 Disch.	5.29	1.88	-.14	-1.05
Fim 13 Admit	**1.15**	**.65**	**22.38**	**4.73**
Fim 13 Disch.	2.25	1.62	-.37	1.02
OTFim Admit	**28.34**	**8.53**	**.67**	**-.05**
OTFim Disch.	35.54	9.02	.59	-.80

distributions. The T^2 test and the RT were calculated for the sample size and FIM admit/discharge distribution combinations outlined previously in the robustness portion of the simulation study. Each experiment was repeated at the .10, .05, and .01 alpha levels.

For the power portion of the study, group (admit vs. discharge) served as the primary independent variable (i.e., group one consisted of scores obtained from the admit distributions, and group two consisted of scores obtained from the associated discharge distributions). FIM scores served as the dependent variables for each group (i.e., combinations of either 2 or 5 FIM scores which served as the dependent variables). Differences in randomly selected FIM scores from the admit distributions were tested against randomly selected scores from their related discharge distributions. Random and independent samples with replacement were selected for each sample size and number of dependent variable combinations.

It is emphasized that independent samples tests are appropriate for this study as individual scores were randomly selected with replacement from each distribution and not pairs of scores. In a separate study using the same data distributions, Nanna & Sawilowsky (1998) took independent random samples from several of the admit distributions used in this study and using Monte Carlo methods correlated them with independent samples drawn from their corresponding discharge distributions. They repeated this experiment 1,000,000 times, and found the long run correlation to be $r = -.0083$.

Figure 1. Distribution of FIM 1 Scores

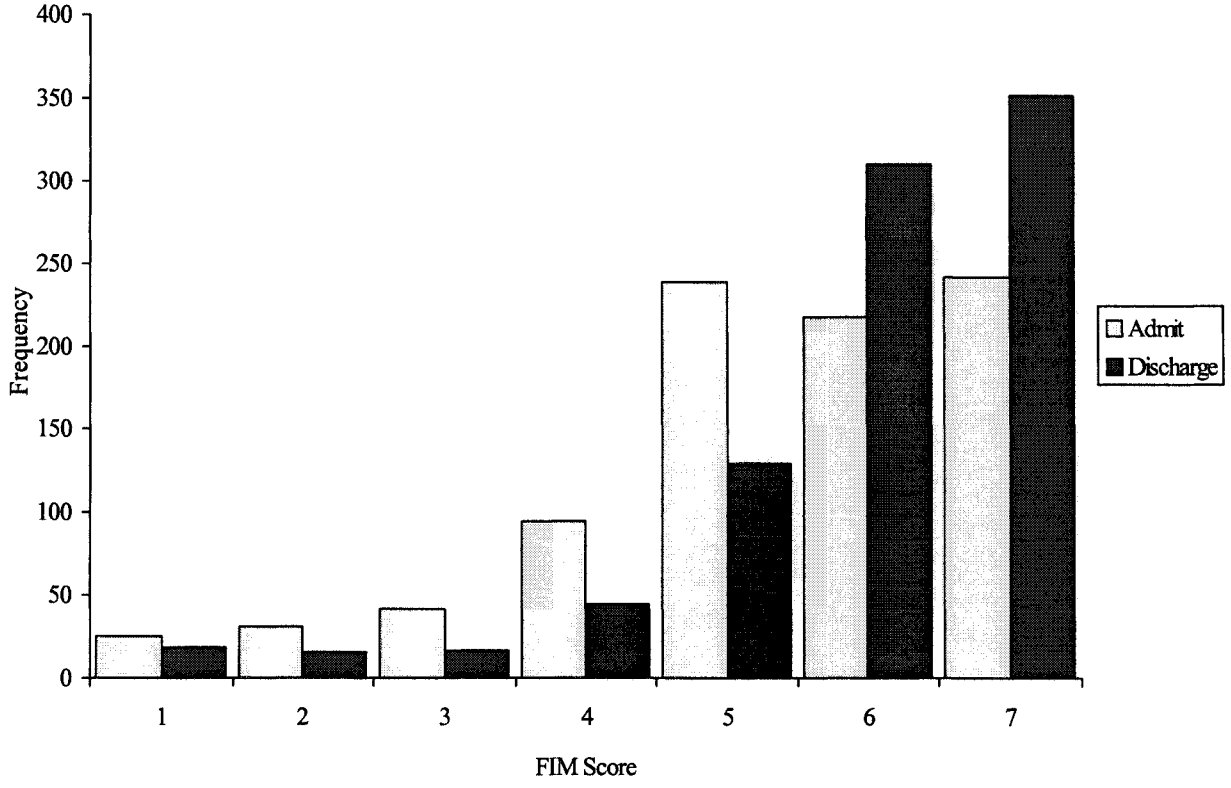
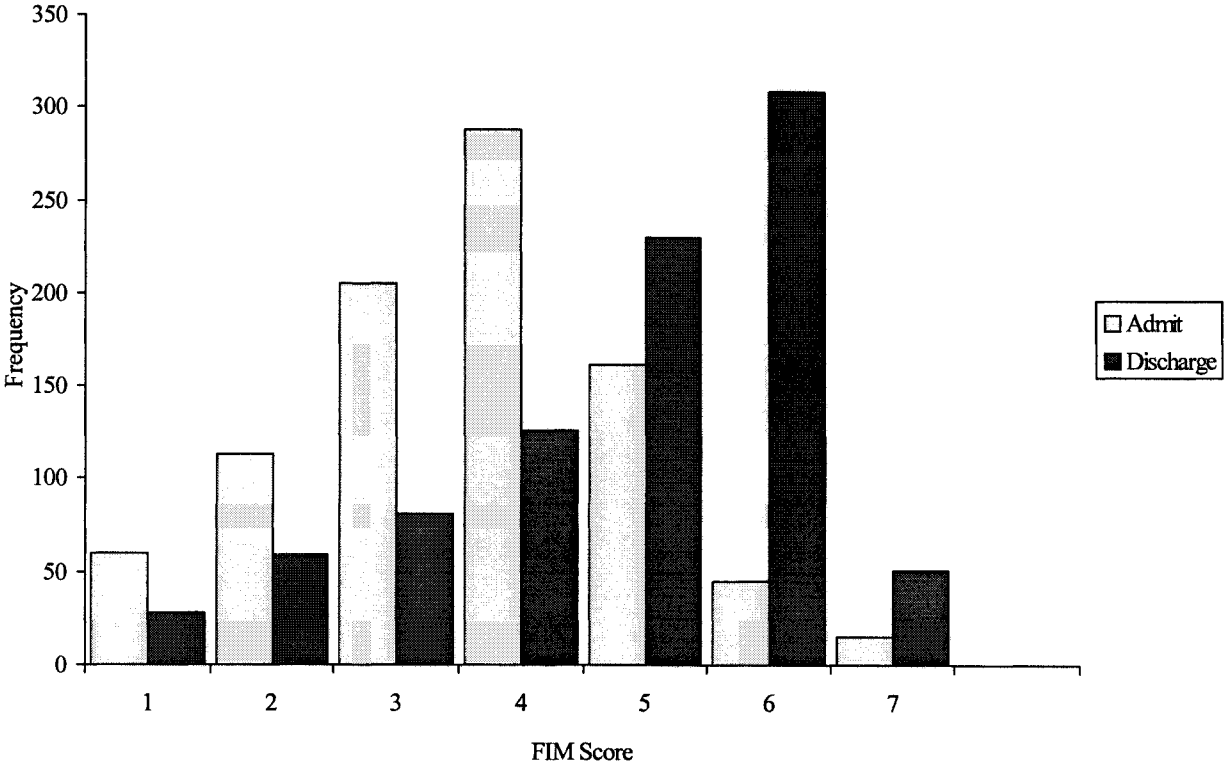


Figure 2. Distribution of FIM 3 Scores



Results

Robustness - Two Dependent Variables

Results of each test were recorded for both statistics. The number of replications per experiment was 10,000 and the proportion of rejections served as an indication of robustness or power for each statistic. The one-tailed power of T^2 and the RT test were compared at the 0.10, 0.05, and 0.01 alpha levels. Results for the .05 alpha level are presented in this paper with the remaining results (.10 and .01 alpha levels) available from the first author.

The ability to detect differences between the admit and discharge data distributions (i.e., differences in the centroids or mean vectors), served as an estimate of power for each statistic. Power results in this study are estimates only as effect sizes may be different between the T^2 and RT tests. However, the results do reflect realistic conditions.

It is important to note that the discharge distributions are comprised of scores from the same individuals as in the admit distributions after receiving treatment. The null hypothesis being tested is the independent hypothesis and not the matched-pairs hypothesis as samples are drawn independently and separately from both the admit and discharge distributions. An example of the null hypothesis being tested in the power portion of this study using three dependent variables (e.g., FIM 1, FIM 3, and FIM 5) is as follows:

$$H_o: \begin{pmatrix} \mu_{FIM1a} \\ \mu_{FIM3a} \\ \mu_{FIM5a} \end{pmatrix} = \begin{pmatrix} \mu_{FIM1d} \\ \mu_{FIM3d} \\ \mu_{FIM5d} \end{pmatrix}$$

Figure 3. Distribution of FIM 4 Scores

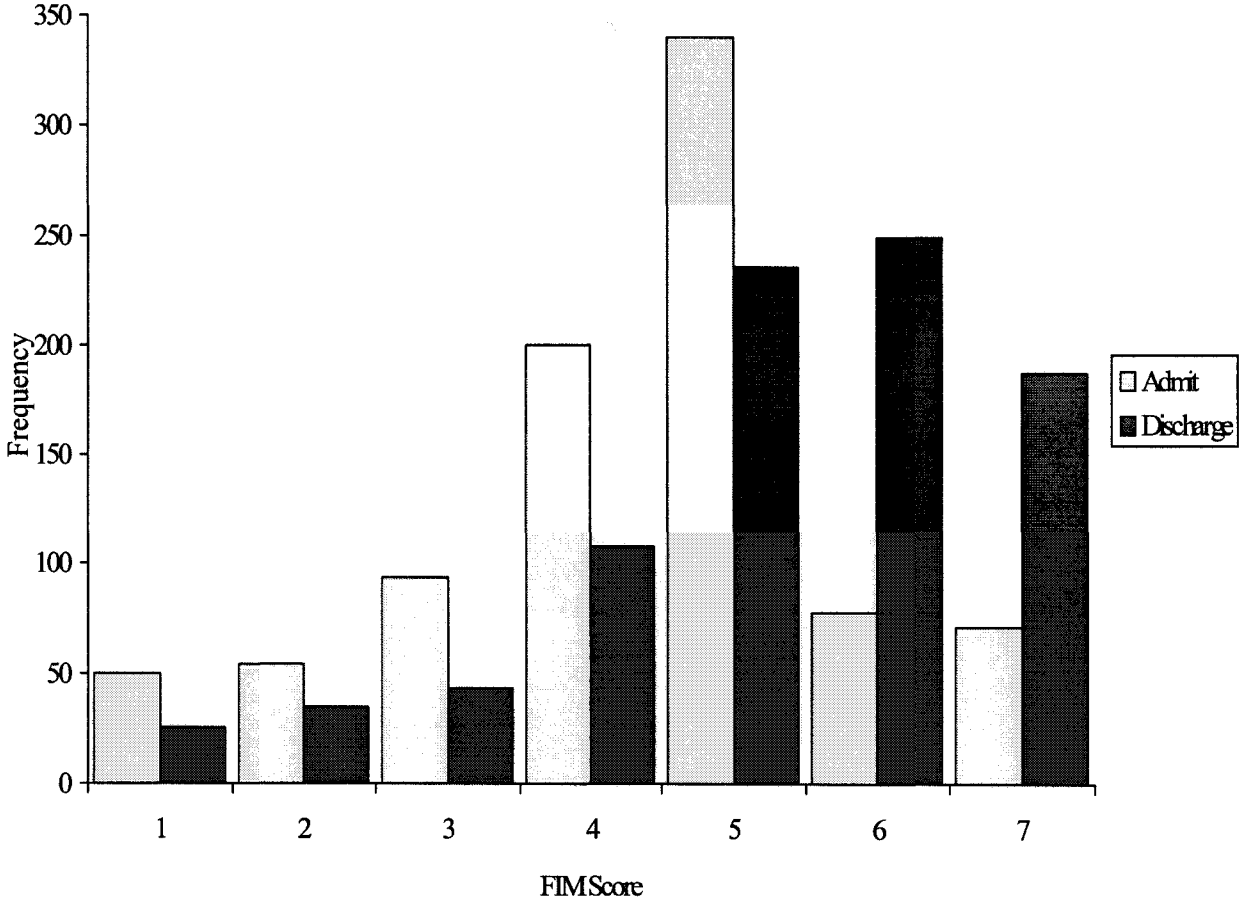


Figure 4. Distribution of FIM 5 Scores

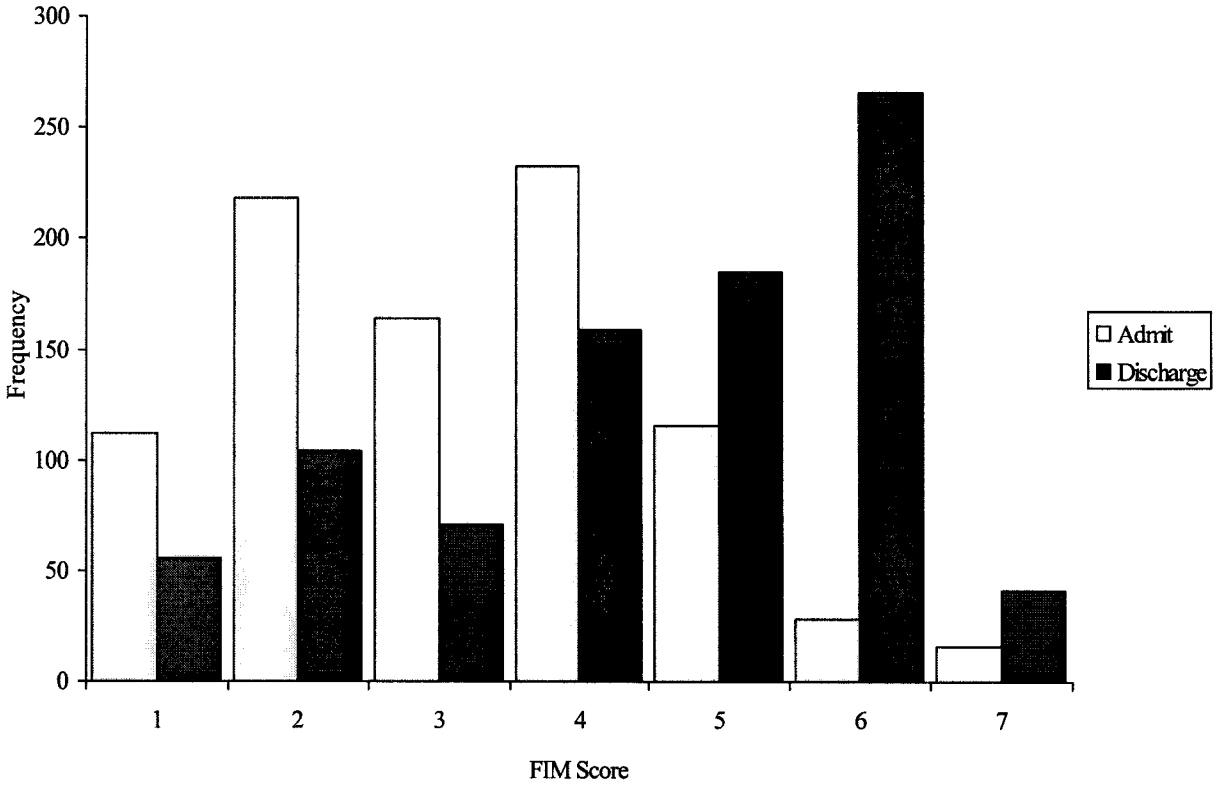


Figure 5. Distribution of FIM 6 Scores

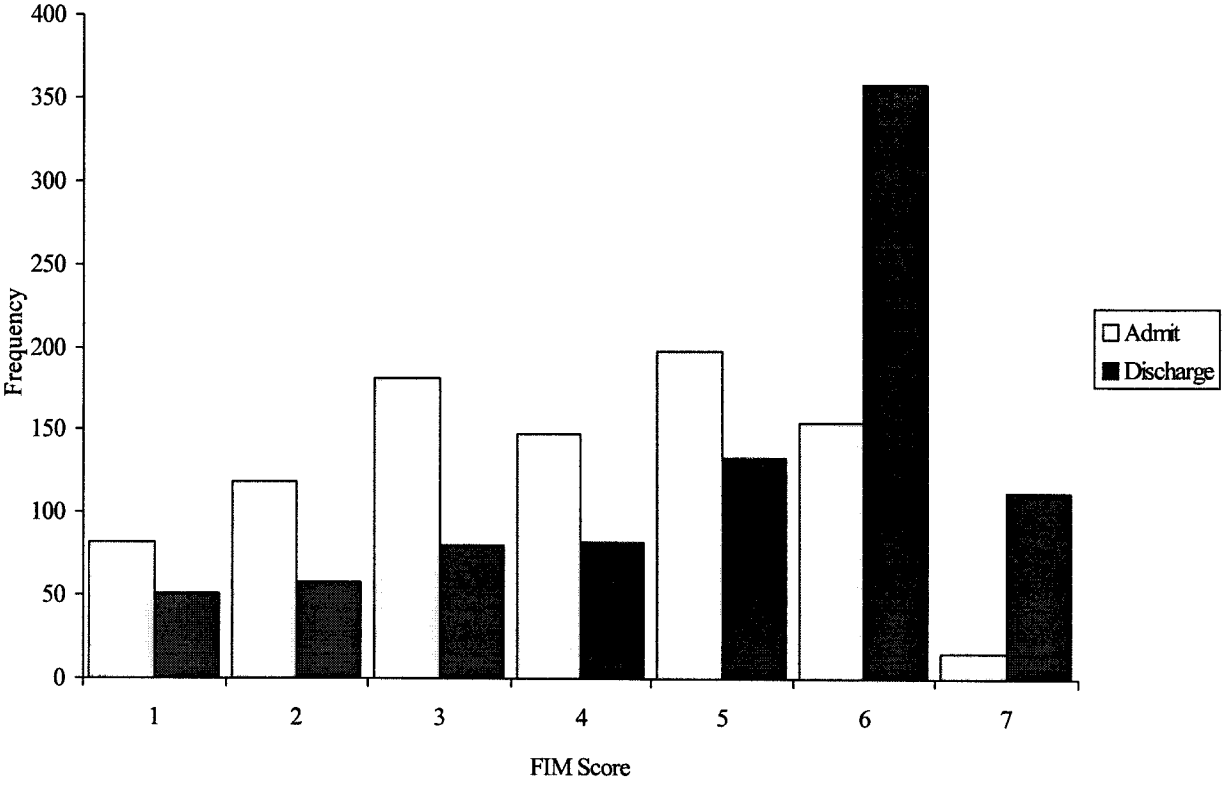


Figure 6. Distribution of FIM 7 Scores

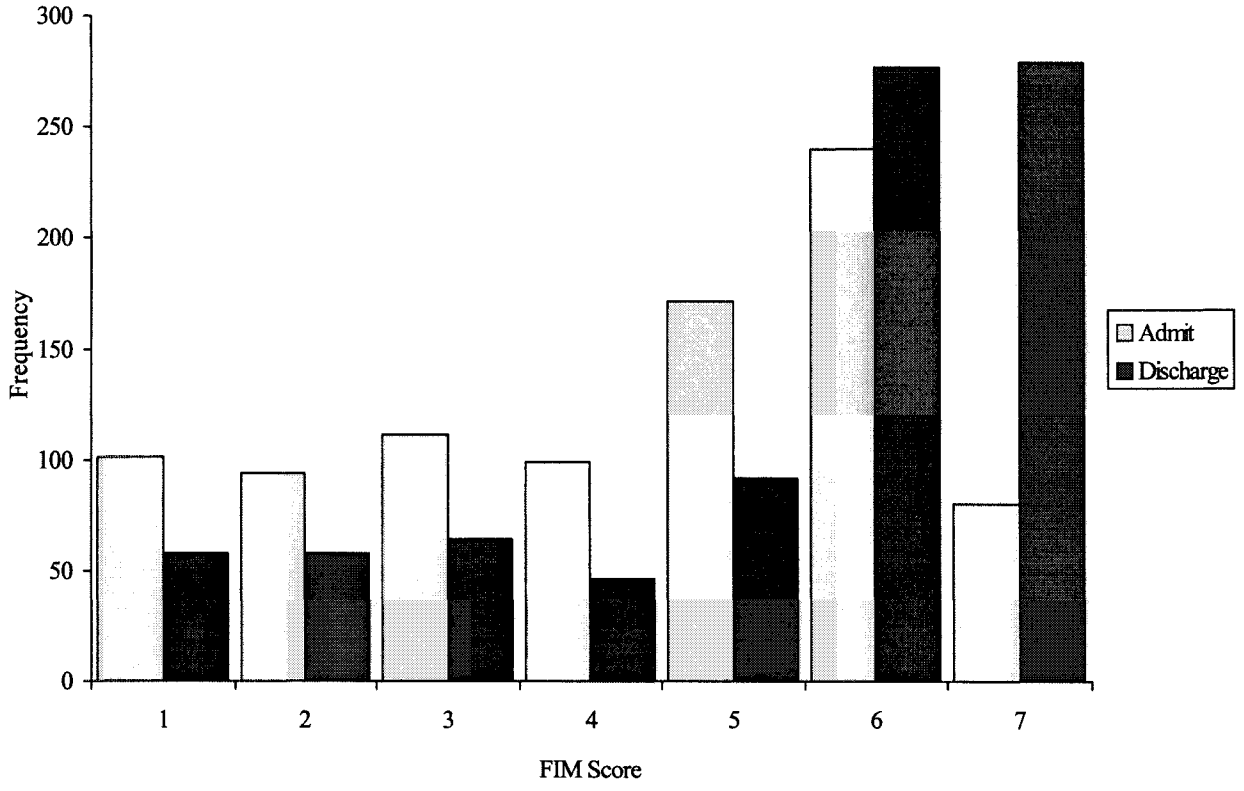
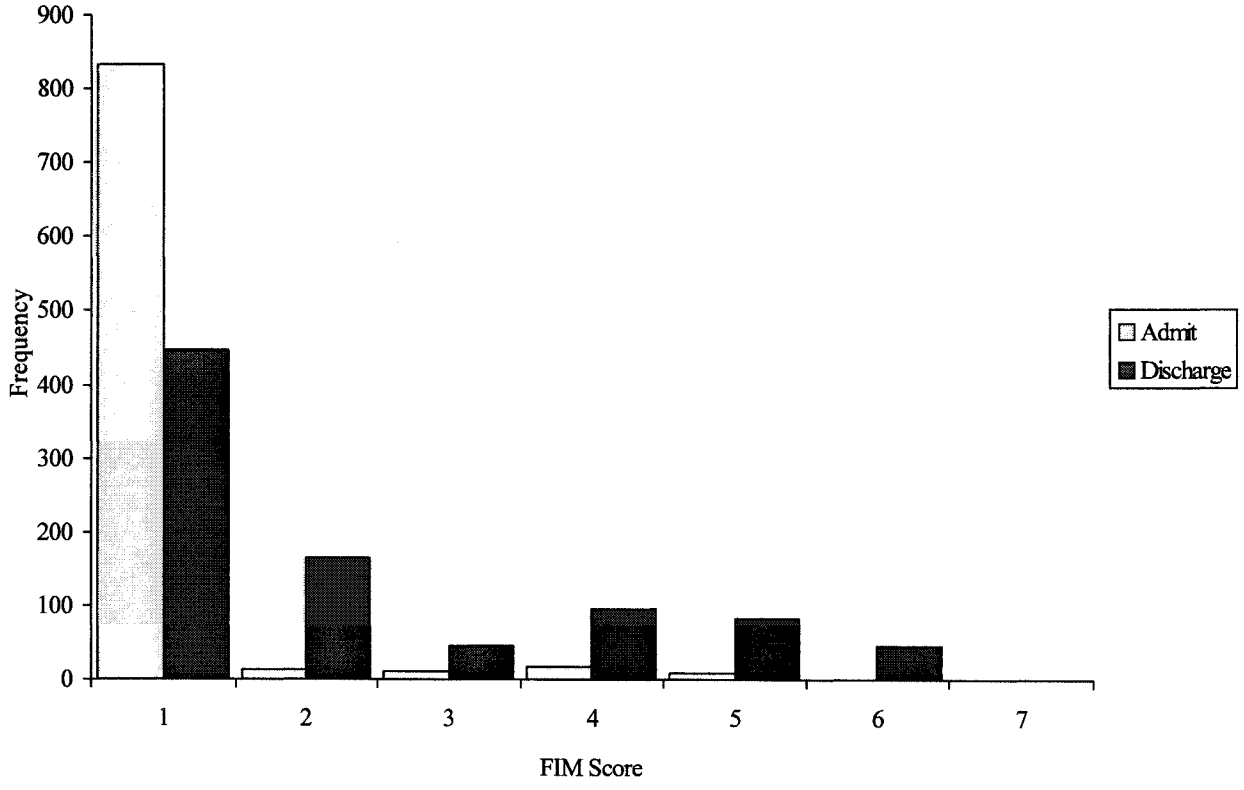


Figure 7. Distribution of FIM 13 Scores



In general, both T^2 and RT were robust with respect to Type I error using Bradley's (1978) "liberal" (p.146) criteria ($|\rho-\alpha| \leq \alpha/2$) but often non-robust using more stringent criteria in other reasonable conditions. This is particularly true the presence of a highly skewed distribution such as FIM 13. The presence of the highly skewed FIM 13 distribution caused both tests to become very conservative. This was true for each alpha level examined (i.e., .10, .05, and .01).

This distribution is roughly analogous to the L-shaped distribution investigated by Bradley (1977; 1980; 1982) which led to similar occurrences in the univariate case. Indeed, in situations where Fim 13 was sampled, the degree of non-robustness was quite pronounced with Type I error rates for both T^2 and RT becoming extremely conservative, reaching a low of .0177 (normal alpha = .05) at sample size (5,15) using the combination of FIM 1 and FIM 13.

Levels of non-robustness, as expected, were dependent on the interplay of several factors including sample size and alpha level. It is evident that, as in the univariate case, blanket statements concerning robustness of T^2 and RT cannot be made. Moreover, where non-robustness was present, the RT was frequently more conservative than T^2 and, in general, levels of robustness did not improve as sample size increased.

Robustness - Five Dependent Variables

Results for the five-dependent variable condition were similar to the two-dependent variable and are located in Table 4. Both T^2 and RT were relatively robust with respect to Type I error under most combinations but relatively non-robust at others. As in the two-dependent variable condition, when levels of non-robustness were found they tended to be conservative in the presence of extreme asymmetry such as that displayed in FIM 13. What is remarkable is that each test became conservative when FIM 13 was present, regardless of the combinations of other shapes. That is, the effect of a skewed distribution was not offset by the presence of more "tame" distributions. Unlike the two-dependent variable condition, however, both tests tended to improve as sample size increased. There were, of course, several noted exceptions.

The levels of robustness had little deviation from nominal alpha except when the distribution sampled was highly skewed (e.g., FIM 13). See Table 5.

Power - Two Dependent Variables

Results for this portion of the study are located in Table 3. In general, across all data sets, alpha levels, and sample size combinations, the RT was consistently more powerful than T^2. Although, the power of both tests, as expected, increased with increased sample size, the RT, on

Figure 8. Distribution of OT Composite Score

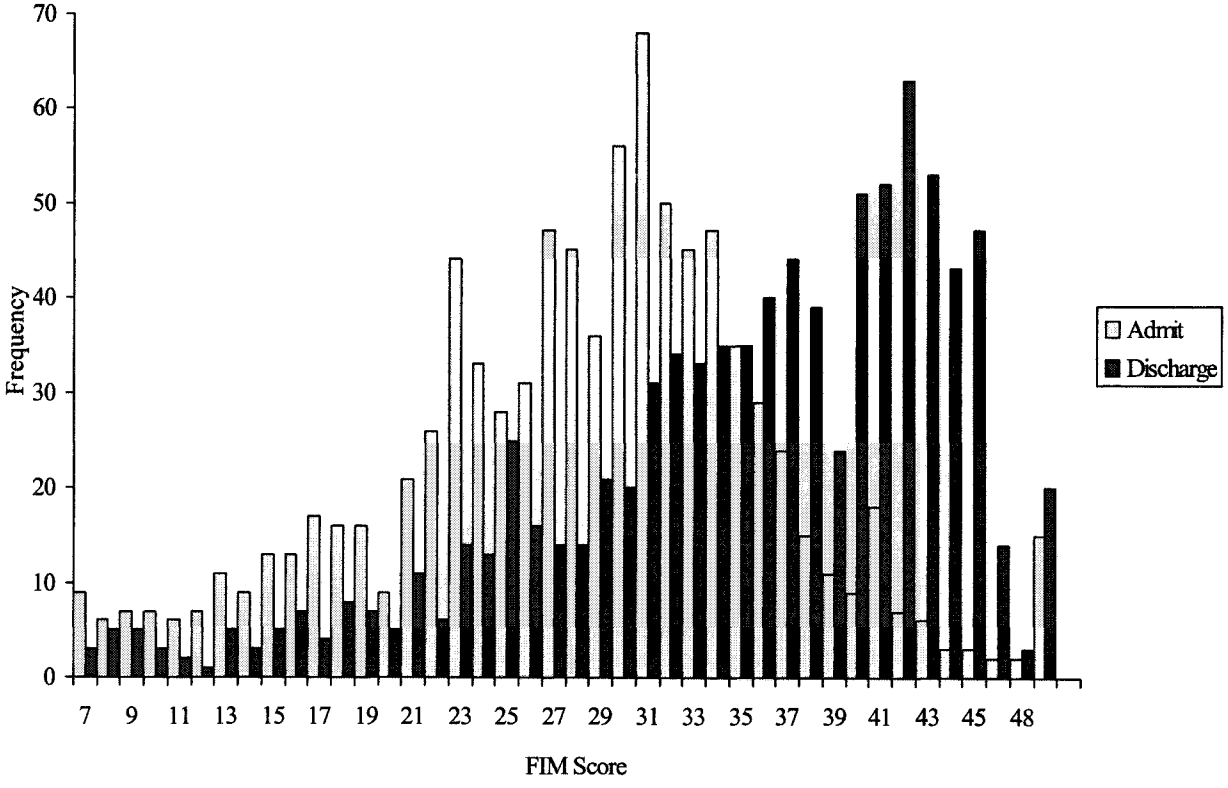


Table 2: Type I Error Rates for Two-dependent Variable Combinations at Alpha = .05.

Fim Distributions	Statistic	Sample Size 5,15	10,10	10,20	15,15	15,45	30,30	25,75	50,50
1, 3	**HT^2**	**.0468**	**.0468**	**.0371**	**.0513**	**.0487**	**.0469**	**.0507**	**.0496**
	RT	.0486	.0510	.0382	.0527	.0491	.0499	.0534	.0513
1, 7	**HT^2**	**.0456**	**.0529**	**.0503**	**.0493**	**.0427**	**.0520**	**.0501**	**.0468**
	RT	.0497	.0570	.0522	.0515	.0458	.0503	.0503	.0501
1,13	**HT^2**	**.0365**	**.0199**	**.0265**	**.0177**	**.0463**	**.0246**	**.0470**	**.0308**
	RT	.0329	.0211	.0253	.0199	.0426	.0233	.0467	.0309
1, Ot	**HT^2**	**.0496**	**.0469**	**.0484**	**.0477**	**.0464**	**.0467**	**.0532**	**.0449**
	RT	.0532	.0527	.0525	.0514	.0472	.0486	.0527	.0485
3,7	**HT^2**	**.0486**	**.0494**	**.0487**	**.0523**	**.0500**	**.0463**	**.0530**	**.0472**
	RT	.0485	.0506	.0511	.0522	.0479	.0457	.0511	.0484
3,13	**HT^2**	**.0368**	**.0200**	**.0279**	**.0204**	**.0413**	**.0244**	**.0425**	**.0312**
	RT	.0316	.0186	.0252	.0210	.0386	.0247	.0413	.0296
4,7	**HT^2**	**.0486**	**.0482**	**.0454**	**.0487**	**.0514**	**.0520**	**.0514**	**.0501**
	RT	.0518	.0491	.0489	.0512	.0498	.0532	.0507	.0508
4,13	**HT^2**	**.0392**	**.0192**	**.0249**	**.0201**	**.0396**	**.0233**	**.0477**	**.0315**
	RT	.0342	.0187	.0239	.0209	.0367	.0233	.0490	.0302
4, Ot	**HT^2**	**.0512**	**.0521**	**.0487**	**.0516**	**.0478**	**.0493**	**.0487**	**.0488**
	RT	.0513	.0564	.0503	.0541	.0486	.0511	.0504	.0500
6,7	**HT^2**	**.0496**	**.0537**	**.0517**	**.0510**	**.0502**	**.0537**	**.0500**	**.0532**
	RT	.0506	.0536	.0512	.0523	.0503	.0540	.0508	.0531
6,13	**HT^2**	**.0399**	**.0223**	**.0255**	**.0192**	**.0451**	**.0259**	**.0500**	**.0327**
	RT	.0319	.0203	.0225	.0192	.0412	.0246	.0444	.0307
6,Ot	**HT^2**	**.0512**	**.0509**	**.0482**	**.0513**	**.0456**	**.0496**	**.0510**	**.0525**
	RT	.0502	.0547	.0511	.0544	.0460	.0511	.0501	.0506
13,Ot	**HT2**	**.0375**	**.0208**	**.0256**	**.0203**	**.0489**	**.0215**	**.0447**	**.0290**
	RT	.0314	.0218	.0243	.0208	.0412	.0226	.0422	.0296

average, tended to reject the null hypothesis anywhere from 1% to 17% more often than T^2. And although differences were moderate at times, the power of RT was rarely less than T^2. The distribution of FIM 13 scores had a negative impact on power at small sample sizes compared with other distribution combinations, but each test generally rehabilitated itself as sample size increased. However, power in the presence of FIM 13 scores must be interpreted in the

Table 3: Type I Error Rates for Five-dependent Variable Combinations at Alpha = .05.

Fim		Sample Size							
Distributions	Statistic	5,15	10,10	10,20	15,15	15,45	30,30	25,75	50,50
1,4,5,6,7	**HT^2**	**.0499**	**.0501**	**.0529**	**.0480**	**.0515**	**.0508**	**.0501**	**.0494**
	RT	.0515	.0544	.0524	.0499	.0509	.0529	.0516	.0497
1, 5,6,7,13	**HT^2**	**.0439**	**.0350**	**.0359**	**.0322**	**.0445**	**.0343**	**.0478**	**.0408**
	RT	.0410	.0363	.0356	.0342	.0422	.0346	.0444	.0410
1,6,7,13,Ot	**HT^2**	**.0424**	**.0349**	**.0381**	**.0368**	**.0464**	**.0374**	**.0472**	**.0453**
	RT	.0371	.0362	.0362	.0370	.0451	.0361	.0491	.0427
3,4,5,6,7	**HT^2**	**.0487**	**.0535**	**.0508**	**.0509**	**.0456**	**.0527**	**.0511**	**.0465**
	RT	.0499	.0547	.0520	.0515	.0489	.0524	.0519	.0479
3,5,6,7,13	**HT^2**	**.0470**	**.0354**	**.0365**	**.0327**	**.0490**	**.0373**	**.0496**	**.0411**
	RT	.0435	.0359	.0352	.0330	.0463	.0365	.0466	.0401
3,6,7,13,Ot	**HT^2**	**.0443**	**.0374**	**.0380**	**.0362**	**.0445**	**.0351**	**.0505**	**.0439**
	RT	.0371	.0399	.0354	.0378	.0422	.0366	.0492	.0418
4,5,6,7,13	**HT^2**	**.0431**	**.0372**	**.0395**	**.0347**	**.0467**	**.0397**	**.0508**	**.0386**
	RT	.0418	.0361	.0385	.0343	.0461	.0376	.0463	.0394
4,6,7,13,Ot	**HT^2**	**.0485**	**.0346**	**.0362**	**.0365**	**.0470**	**.0373**	**.0463**	**.0446**
	RT	.0444	.0363	.0342	.0372	.0428	.0367	.0452	.0412
5,6,7,13,Ot	**HT^2**	**.0418**	**.0341**	**.0378**	**.0335**	**.0453**	**.0355**	**.0506**	**.0397**
	RT	.0392	.0333	.0379	.0355	.0438	.0368	.0474	.0375

context of its associated Type I error.

Power - Five Dependent Variables.

Results for this portion of the study are located in Table 6. Although the RT maintained power over T^2 in most situations, the difference in magnitude tended to be less than in the two-dependent variable condition. There are instances, however, where the RT is substantially more powerful than T^2. However, large differences in power were less frequent and both tests tended to perform equally well with greater frequency as sample size increased. An additional result seemed to be that the impact of FIM 13 scores was less profound. Under the conditions where each distribution was sampled multiple times (see Table 7), the only remarkable results were, again, when Fim 13 served as the sole dependent variable distribution (i.e., FIM 13 sampled five times). In general, RT worked better at small sample sizes, but T^2 recovered as sample size increased.

Summary of Results

Results suggest that both T^2 and the RT are robust under most non-normal situations in the independent samples case when data are scaled at the Ordinal level. These results hold for both the two-dependent variable and five-dependent variable conditions. The most noted exception for both conditions was in the presence of a highly skewed distribution such as the FIM 13 used in this study. With respect to power, T^2 recovers somewhat as sample size reaches (25,75) and (50,50), however, RT maintains a modest power advantage at all sample sizes and

Table 4: Type I Error Rates for Five-dependent Variables Sampling Using One Distribution for Each Variable at Alpha = .05.

Fim Distributions	Statistic	Sample Size 5,15	10,10	10,20	15,15	15,45	30,30	25,75	50,50
1	**HT^2**	**.0485**	**.0476**	**.0474**	**.0491**	**.0520**	**.0515**	**.0462**	**.0529**
	RT	.0460	.0546	.0512	.0564	.0524	.0527	.0461	.0522
4	**HT^2**	**.0495**	**.0487**	**.0469**	**.0515**	**.0488**	**.0481**	**.0531**	**.0509**
	RT	.0493	.0580	.0467	.0496	.0492	.0516	.0505	.0521
7	**HT^2**	**.0505**	**.0513**	**.0531**	**.0540**	**.0516**	**.0523**	**.0531**	**.0475**
	RT	.0506	.0560	.0526	.0546	.0509	.0545	.0529	.0489
13	**HT^2**	-	-	-	-	**.0382**	**.0015**	**.0422**	**.0079**
	RT	-	-	-	-	.0296	.0008	.0351	.0069
Ot	**HT^2**	**.0453**	**.0475**	**.0535**	**.0507**	**.0467**	**.0484**	**.0508**	**.0475**
	RT	.0478	.0529	.0562	.0559	.0507	.0505	.0483	.0493

* Results were not available for sample sizes less than (15, 45) for FIM 13. Due to the highly skewed and discrete nature of this distribution and given the restrictions on the number of repetitions in this simulation study, no variance could be calculated for small sample sizes.

distribution combinations. There were noted exceptions. For instance, T^2 displayed almost negligible power advantages over RT when distribution combinations (1,3), (1,7), and (1,13) were sampled for reasons unknown.

Although T^2 is more powerful under normality, the RT has been shown to be only slightly less powerful (Bhattacharyya, Johnson, & Neave, 1971; Tiku & Singh, 1982; Nath & Duran, 1983) and in non-normal situations is accepted as being more powerful in many cases (Nath, 1982; Nath & Duran, 1983; Zwick, 1986). In this study, the RT consistently outperformed T^2 at nearly every sample size, alpha level, and dependent variable combination. Unlike previous studies examining robustness and power which often employ the use of artificial distributions and/ or treatment effects, this study examined the operating characteristics of T^2 and the RT under real conditions. In fact, the results of this study have direct implications to the substantive field from which these distributions came, as well as other fields with similar types of data.

Conclusion

With respect to ordinal scaled data in the form of likert scaled data commonly obtained in applied data analysis situations and its inherent violation of normality when testing for equality of centroids, the Rank Transformation procedure provides an increase in power over Hotelling's T^2 while maintaining acceptable Type I error rates. This is particularly true with two-dependent variables and smaller sample sizes. The results must not be extrapolated beyond the context of this study to other data analysis layouts, such as the multivariate factorial analysis of variance. The robustness results of this study on the multivariate two independent samples layout with likert scaled data were in agreement with, and extend the results found by Heeren & D'Agostino (1987) and Hsu & Feldt (1969) on the univariate independent samples t-test; and the robustness and power results found by Nanna & Sawilowsky (1998) for the rank transformation analog to the univariate two independent samples layout with likert scaled data.

References

Akritas, M.G. (1990). The rank transform method in some two-factor designs. *Journal of the American Statistical Association, 85*, 73-78.

Akritas, M.G. (1991). Limitations of the rank transform procedure: A study of repeated measures designs, part 1. *Journal of the American Statistical Association, 86*, 457-460.

Table 5: Power for Two-dependent Variable Combinations at Alpha = .05.

Fim Distributions	Statistic	Sample Size 5,15	10,10	10,20	15,15	15,45	30,30	25,75	50,50
1, 3	**HT^2**	**.3225**	**.3953**	**.5205**	**.5693**	**.7683**	**.8755**	**.9429**	**.9837**
	RT	.3255	.4357	.5733	.6245	.8357	.9174	.9740	.9919
1, 7	**HT^2**	**.2155**	**.2211**	**.3096**	**.3323**	**.4857**	**.5924**	**.7095**	**.8315**
	RT	.2075	.2686	.3570	.4024	.5823	.7094	.8214	.9179
1,13	**HT^2**	**.1150**	**.2404**	**.3042**	**.5067**	**.6717**	**.9256**	**.9598**	**.9967**
	RT	.1058	.3371	.4203	.6584	.8473	.9795	.9922	.9997
1, Ot	**HT^2**	**.3176**	**.3967**	**.5224**	**.5754**	**.7614**	**.8698**	**.9408**	**.9817**
	RT	.3290	.4434	.5867	.6454	.8470	.9287	.9800	.9949
3,7	**HT^2**	**.3300**	**.4429**	**.5587**	**.6216**	**.8050**	**.9099**	**.9657**	**.9942**
	RT	.3599	.5018	.6445	.7009	.8916	.9525	.9898	.9984
3,13	**HT^2**	**.2411**	**.4723**	**.5792**	**.7417**	**.9115**	**.9897**	**.9981**	**.9997**
	RT	.2723	.5661	.7187	.8491	.9738	.9980	.9999	.9999
4,7	**HT^2**	**.2381**	**.2959**	**.3948**	**.4396**	**.6089**	**.7518**	**.8365**	**.9346**
	RT	.2549	.3635	.4781	.5396	.7363	.8555	.9317	.9787
4,13	**HT^2**	**.1432**	**.3221**	**.3914**	**.6033**	**.7843**	**.9633**	**.9849**	**.9988**
	RT	.1617	.4274	.5493	.7497	.9246	.9903	.9989	.9999
4, Ot	**HT^2**	**.3626**	**.4719**	**.5836**	**.6541**	**.8351**	**.9243**	**.9752**	**.9954**
	RT	.3872	.5349	.6670	.7334	.9156	.9688	.9938	.9992
6,7	**HT^2**	**.2636**	**.3325**	**.4326**	**.4832**	**.6637**	**.8083**	**.8878**	**.9635**
	RT	.2823	.3965	.5200	.5743	.7854	.8905	.9556	.9893
6,13	**HT^2**	**.1604**	**.3638**	**.4498**	**.6422**	**.8290**	**.9739**	**.9903**	**.9998**
	RT	.1803	.4690	.6042	.7822	.9476	.9946	.9986	1.000
6,Ot	**HT^2**	**.3818**	**.5024**	**.6238**	**.6875**	**.8624**	**.9416**	**.9804**	**.9967**
	RT	.4135	.5622	.7058	.7643	.9345	.9763	.9960	.9994
13,Ot	HT^2	.2392	.4762	.5869	.7454	.9070	.9883	.9966	.9998
	RT	.2686	.5872	.7294	.8612	.9778	.9987	1.000	1.000

Algina, J. & Oshima, T. (1990). Robustness of the independent samples Hotelling's T^2 to variance-covariance heteroscedasticity when sample sizes are unequal and in small ratios. *Psychological Bulletin, 108(2)*, 308-313.

Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin, 58,* 305-316.

Table 6: Power for Five-dependent Variable Combinations at Alpha = .05.

Fim Distributions	Statistic	Sample Size 5,15	10,10	10,20	15,15	15,45	30,30	25,75	50,50
1,4,5,6,7	**HT^2**	**.3811**	**.4998**	**.6751**	**.7382**	**.9282**	**.9831**	**.9960**	**.9998**
	RT	.3904	.5705	.7524	.8066	.9694	.9942	.9996	1.000
1,5,6,7,13	**HT^2**	**.3202**	**.4950**	**.6665**	**.7930**	**.9544**	**.9953**	**.9999**	**1.000**
	RT	.3302	.5882	.7694	.8798	.9881	.9995	1.000	1.000
1,6,7,13,Ot	**HT^2**	**.3312**	**.5105**	**.6891**	**.8065**	**.9600**	**.9964**	**.9999**	**.9999**
	RT	.3528	.6197	.8080	.9003	.9932	.9999	1.000	1.000
3,4,5,6,7	**HT^2**	**.4549**	**.6134**	**.7919**	**.8506**	**.9758**	**.9970**	**.9997**	**1.000**
	RT	.4853	.6824	.8626	.9020	.9937	.9992	1.000	1.000
3,5,6,7,13	**HT^2**	**.3920**	**.6135**	**.7923**	**.8892**	**.9873**	**.9991**	**.9999**	**1.000**
	RT	.4251	.7005	.8781	.9440	.9977	.9999	1.000	1.000
3,6,7,13,Ot	**HT^2**	**.4116**	**.6232**	**.7979**	**.8926**	**.9869**	**.9997**	**1.000**	**1.000**
	RT	.4474	.7194	.8929	.9547	.9989	.9999	1.000	1.000
4,5,6,7,13	**HT^2**	**.3416**	**.5370**	**.7097**	**.8235**	**.9720**	**.9995**	**.9998**	**1.000**
	RT	.3636	.6275	.8222	.9061	.9945	.9999	1.000	1.000
4,6,7,13,Ot	**HT^2**	**.3574**	**.5447**	**.7290**	**.8378**	**.9744**	**.9984**	**.9996**	**1.000**
	RT	.3886	.6595	.8460	.9254	.9963	.9997	1.000	1.000
5,6,7,13,Ot	**HT^2**	**.3949**	**.6108**	**.7912**	**.8870**	**.9869**	**.9996**	**1.000**	**1.000**
	RT	.4204	.7065	.8814	.9472	.9982	1.000	1.000	1.000

Bhattacharyya, G. K., Johnson, R.A., & Neave, H. R. (1971). A comparative power study of the bivariate rank sum test and T^2. *Technometrics, 13*(1), 191-198.

Blair, R. C., Sawilowsky, S. S., & Higgins, J. J. (1987). Limitations of the rank transform in tests for interaction. *Communications in Statistics: Computation and Simulation, B16*, 1133-1145.

Blair, R. C., Higgins, J. J., Karniski, W., & Kromrey, J. D. (1994). A study of multivariate permutation tests which may replace Hotelling's T^2 in prescribed circumstances. *Multivariate Behavioral Research, 29*(2), 141-163.

Boneau, C. A. (1961). A note on measurement scales and statistical tests. *American Psychologist, 16*, 260-261.

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

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician, 31*(4), 147-150

Bradley, J. V. (1980b). Nonrobustness in classical tests on means and variances: a large scale sampling study. *Bulletin of the Psychonomic Society, 15*(4), 275-298.

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

Byrnes, M. B. & Powers, F. F. (1989). FIM: Its use in identifying rehabilitation needs in the head injured patient. *Journal of Neuroscience Nursing, 21*, 61-63.

Chase, G. R. & Bulgren, W. G. (1971). A monte carlo investigation of the robustness of T^2. *Journal of the American Statistical Association, 66*, 499-502.

Table 7: Power for Five-dependent Variables Using the Same Distribution for All Sampled Variables at Alpha = .05.

Fim Distributions	Statistic	Sample Size 5,15	10,10	10,20	15,15	15,45	30,30	25,75	50,50
1	**HT^2**	**.2908**	**.2329**	**.4023**	**.3739**	**.6404**	**.7232**	**.8603**	**.9336**
	RT	.2383	.2654	.4047	.4254	.6699	.7970	.9036	.9685
4	**HT^2**	**.3633**	**.4633**	**.6475**	**.7083**	**.9123**	**.9732**	**.9951**	**.9997**
	RT	.3812	.5509	.7362	.8092	.9638	.9919	.9994	1.000
7	**HT^2**	**.3015**	**.3699**	**.5235**	**.5883**	**.8254**	**.9293**	**.9784**	**.9962**
	RT	.3240	.4899	.6610	.7296	.9372	.9834	.9979	.9999
13	**HT^2**	**.0250**	**.4770**	**.6342**	**.9263**	**.9967**	**1.000**	**1.000**	**1.000**
	RT	.0466	.6910	.8767	.9878	1.000	1.000	1.000	1.000
Ot	**HT^2**	**.6409**	**.7925**	**.9227**	**.9541**	**.9983**	**1.000**	**1.000**	**1.000**
	RT	.6865	.8520	.9649	.9792	.9998	1.000	1.000	1.000

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

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

Conover, W. J. & Iman, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs. *Communications in Statistics - Theoretical Methods, A5*(14), 1349-1368.

Conover, W. J. (1980). Practical nonparametric statistics. 2nd Ed., NY: John Wiley.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician, 35* 124-129.

Eaton, M. L. & Efron, B. (1970). Hotelling's T^2 test under symmetry conditions. *Journal of the American Statistical Association, 65*, 702-711.

Everitt, B. S. (1979). A monte carlo investigation of the robustness of Hotelling's one- and two-sample T^2 test. *Journal of the American Statistical Association. 74*(365), 48-51.

Gaito, J. (1986). Some issues in the measurement-statistics controversy. *Canadian Psychology, 27*, 63-68.

Granger, C. V., Hamilton, B.B., Keith, R.A., Zielezny, M., & Sherwin, F.S. (1986). Advances in functional assessment for medical rehabilitation. *Topics in Geriatric Rehabilitation, 1*, 59-74.

Granger, C. V., Cotter, A. C., Hamilton, B. B., Fiedler, R. C., & Hens, M. M. (1990). Functional assessment scales: A study of persons with multiple sclerosis. *Archives of Physical Medicine & Rehabilitation, 71*, 870-875.

Hakstian, A. R., Roed, J. C., & Lind, J. C. (1979). Two-sample T^2 procedure and the assumption of homogeneous covariance matrices. *Psychological Bulletin, 86,* 1255-1263.

Hamilton, B. B., Laughlin, J. A., Granger, C.V., & Kayton, R.M. (1991). Interrater agreement of the seven level Functional Independence Measure (FIM). *Archives of Physical Medicine & Rehabilitation, 72*, 790.

Harwell, M. R. & Serlin, R. (1995). An empirical study of the Type I error rates of five multivariate tests for the single-factor repeated measures model. Paper presented at the annual meeting of the American Educational Research Association. April, San Francisco.

Heeren, T. & D'Agostino, R. (1987). Robustness of the two independent samples t-test when applied to ordinal scaled data. *Statistics in Medicine, 6*, 79-90.

Holloway, L. N. & Dunn, O. J. (1967). The robustness of Hotelling's T^2. *Journal of the American Statistical Association, 62,* 124-136.

Hopkins, J. W. & Clay, P. P. F. (1963). Some empirical distributions of bivariate T^2 and homoscedasticity criterion M under unequal variance and leptokurtosis. *Journal of the American Statistical Association,* December. 1049-1053.

Hora, S. C., & Conover, W. J. (1984). The F statistic in the two-way layout with rank-score transformed data. *Journal of the American Statistical Association, 79*, 668-673.

Hora, S. C. & Iman, R. L. (1988). Asymptotic relative efficiencies of the rank-transformation procedure in randomized complete block design. *Journal of the American Statistical Association, 83*, 462-470.

Hsu, T. C. & Feldt, L. S. (1969). The effect of limitations on the number of criterion score values on the significance level of the F-test. *American Educational Research Journal, 6*(4), 515-527.

Hunter, M. A., & May, R. B. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology, 34*(4), 384-389.

Iman, R. L. (1974). A power study of a rank transform for the two-way classification model when interaction may be present. *The Canadian Journal of Statistics, 2*(2), 227-239.

Iman, R. L., Hora, S. C. & Conover, W. J. (1984). Comparison of asymptotically distribution-free procedures for the analysis of complete blocks. *Journal of the American Statistical Association, 79*(387), 674-685.

Ito, K. & Schull, W. J. (1964). On the robustness of the T^2 test in multivariate analysis of variance when variance-covariance matrices are not equal. *Biometrika, 51*(1), 71-82.

Jensen, D. R. (1982). Efficiency and robustness in the use of repeated measurements. *Biometrics, 38*, 813-825.

Johnson, R. A. & Wichern, D. W. (1982). Applied multivariate statistics. Englewood Cliffs, NJ: Prentice Hall, Inc.

Kariya, T. (1981). A robustness property of Hotelling's T^2 test. *The Annals of Statistics, 9*(1), 211-214.

Katz, S., Ford, A. B., Moskowitz, R. W., Jackson, B. A., & Jaffe, M. W. (1963). Studies of illness in the aged: The index of ADL. A standardized measure of the biological and psychosocial function. *Journal of the American Medical Association, 185*, 914-919.

Keith, R. A., Granger, C. V., Hamilton, B. B., & Sherwin, F. S. (1987). The Functional Independence Measure: A new tool for rehabilitation. In M. G. Eisenberg & R. C. Grzesiak (Eds.), *Advances in clinical rehabilitation, 1*, 6-18. New York: Springer.

Kelley, L. D., & Sawilowsky, S. S. (1997). Nonparametric alternatives to the F statistic in analysis of variance. *Journal of Statistical Computing and Simulation*, 58-343-359.

Keppel, G. (1975). *Design and analysis: A researchers handbook.* Englewood Cliffs, NJ: Prentice Hall.

Lord, F.M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8*, 750-751.

Mardia, K.V. (1975). Assessment of multinormality and the robustness of Hotelling's T^2 test. *Applied Statistics, 24*(2), 163-171.

Matyas, T.A., & Ottenbacher, K.J. (1993). Confounds of insensitivity and blind luck: Statistical conclusion validity in stroke rehabilitation clinical trials. *Archives of Physical Medicine & Rehabilitation, 74*, 559-565.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

Nanna, M. J., & Sawilowsky, S.S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods, 3*, 55-67.

Nath, R. (1982). An investigation of the rank transform in the multivariate one-sample location problem. *Journal of Statistical Computing and Simulation, 16*, 139-155.

Nath, R., & Duran, B.S. (1983). A robust test in the multivariate two-sample location problem. *American Journal of Mathematical and Management Sciences, 3*(3), 225-249.

Ottenbacher, K. J., & Barrett, K. A. (1989). Measures of effect size in the reporting of rehabilitation research. *American Journal of Physical Medicine & Rehabilitation, 68*(2), 52-58.

Ottenbacher, K. J. & Barrett, K. A. (1990). Statistical conclusion validity of rehabilitation research: A quantitative analysis. *American Journal of Physical Medicine & Rehabilitation, 69*(2), 102-107.

Ottenbacher, K. J. (1991). Statistical conclusion validity: Multiple inferences in rehabilitation research. *American Journal of Physical Medicine & Rehabilitation, 70*(6), 317-322.

Ottenbacher, K. J. (1992). Statistical conclusion validity and Type IV errors in rehabilitation research. *Archives of Physical Medicine & Rehabilitation, 73*, 121-125.

Ottenbacher, K. J. (1995). Why rehabilitation research does not work (as well as we think it should). *Archives of Physical Medicine and Rehabilitation, 76*, 123-129.

Ottenbacher, K.J., Hsu, Y., Granger, C.V., & Fiedler, R.C. (1996). The reliability of the Functional Independence Measure: A quantitative review. *Archives of Physical Medicine & Rehabilitation, 77*, 1226-1232.

Sawilowsky, S.S., Blair, R.C., & Higgins, J.J. (1989). An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics, 1*(3), 255-267.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60*(1), 91-126.

Sawilowsky, S. S. (1993). Comments on Using Alternatives to Normal Theory Statistics in Social and Behavioural Science. *Canadian Psychology, 34*, 432-439.

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t-test to departures from population normality. *Psychological Bulletin, 111*(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 & Clinical Psychology, 60*(2), 240-243.

Sedlmeir, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309-316.

Serlin, R. C., & Harwell, M. R. (1989). A comparison of Hotelling's T^2 and Puri and Sen's rank test for the single-factor, repeated measures design. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Siegal, S. (1956). Nonparametric statistics. New York: McGraw-Hill.

Stevens, S. S. (1946). On the theory of scales and measurement. *Science, 103*, 677-680.

Stineman, M. G., Shea, J. A., Jette, A., Tassoni, C. J., Ottenbacher, K. J., Fiedler, R., & Granger, C. V. (1996). The Functional Independence Measure: Tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. *Archives of Physical Medicine & Rehabilitation, 77*, 1101-1108.

Tiku, M. L. & Singh, M. Robust statistics for testing mean vectors of multivariate distributions. *Communications in Statistics - Theoretical Methods, 11*(9), 985-1001.

Utts, J. M., & Hettmansperger, T. P. (1980). A robust class of tests and estimates for multivariate location. *Journal of the American Statistical Association, 75*, 939-946.

Zimmerman, D. W. (1991). Failure of the Mann-Whitney test: A note on the simulation study of Gibbons and Chakraborti. *Journal of Experimental Education, 60*, 359-364.

Zumbo, B. D., & Zimmerman, D. W. (1991). Levels of measurement and the relation between parametric and nonparametric statistical tests: a review of recent findings. *A handbook for data analysis in the behavioural sciences, 1*, 481-517.

Zumbo, B. D. & Zimmerman, D. W. (1993a). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology, 34*, 390-400.

Zwick, R. (1986). Rank and normal scores alternatives to Hotelling's T^2. *Multivariate Behavioral Research, 21*, 169-186.