Analysis Options for Testing Group Differences on Ordered Categorical Variables: An Empirical Investigation of Type I Error Control and Statistical Power

Jeffrey D. Kromrey **Kristine Y. Hogarty**
University of South Florida

Type I error control and statistical power of four methods of testing group differences on an ordered categorical response variable were evaluated in a Monte Carlo study. Data were analyzed using the independent means *t*-test, the chi-square test of homogeneity, the delta statistic, and a cumulative logit model. The number of categories of the response variable, sample size, population distribution shape, and effect size were examined. These experimental conditions were crossed with each other providing a total of 192 conditions. The independent means *t*-test provided the best control of Type I error, but was rarely the most powerful. For the 5-point response scale, the chi-square was most often the most powerful. Results varied for the 7-point response scale. Small power differences (in many instances) among these procedures suggest that researchers' choices should be driven by the interpretations that are appropriate for the research questions being addressed.

Response variables that are measured as ordered categories, such as Likert scale and other rating scale items, present a variety of analysis options for researchers. For example, in testing for the equality of two groups on such a response variable, the data are usually analyzed using either a Pearsonian chi-square test of homogeneity or a test for the equality of population means such as the independent means *t*-test. Implicit in the former analysis is the treatment of the response variable as nominal-level measurement, while the latter analysis implies an assumption of interval-level data. In between these two extremes are analysis options that are infrequently seen in applied educational research, specifically, logistic regression models (Agresti, 1996; Agresti & Finlay, 1997) and ordinal indices of association (Cliff, 1996a). Arguments about the relationship between levels of measurement and appropriate statistical analyses have been ongoing since Stevens' (1951) classic work, and, no doubt, will continue in the future.

Although the present paper is not intended to directly address the logical arguments related to Stevens' levels of measurement issues, the influence of his work is unavoidable. For example, recent arguments on the level-of-measurement/appropriate-statistics issue have been advanced by Davidson and Sharma (1988) and by Velleman and Wilkinson (1993). Rather than examining such analysis issues in terms of "appropriate statistics," the issues surrounding the analysis of ordered categorical data may be productively addressed in terms of Type I error control and statistical power. For example, Cliff (1996a) has argued that ordinal measures of association such as Tau and delta are useful both descriptively and inferentially because of their robustness properties when compared to traditional parametric tests such as the independent means *t*-test. Similarly, Agresti (1989) suggested that researchers may realize power advantages in the use of cumulative logit models rather than Pearsonian chi-square tests when testing hypotheses about ordered categorical data. Unfortunately, neither Cliff nor Agresti presented empirical evidence of the magnitude of power differences or the extent of improvement in robustness when these ordinal-level statistics are used.

It is important to recognize that different statistical null hypotheses are tested with each of these procedures. For example, the independent means *t*-test provides a test of the null hypothesis of equivalence of population means and the chi-square test of homogeneity tests the equivalence of the population proportions at each level of the response variable. In contrast, the G^2 statistic used in testing the cumulative logit model provides a test of the null hypothesis of equal cumulative log odds, while the delta statistic is used to test equivalence of probabilities of scores in each group being larger than scores in the other (the property that Cliff (1993) referred to as "dominance"). However, as Cliff (1993, 1996a) has pointed out, despite the differences in statistical null hypotheses tested, each of these procedures may be used to test the same, conceptual research hypothesis (e.g., "the two groups respond differently on the dependent variable").

The purpose of the present study was to empirically compare the Type I error control and statistical power of four tests of group differences on ordered categorical response data: a parametric test of mean differences (independent means *t*-test), the

Pearsonian chi-square test of homogeneity, the cumulative logit model recommended by Agresti (1989, 1996), and the delta statistic recommended by Cliff (1993, 1996a). Such a comparison was made for a variety of sample sizes and distribution shapes likely to be encountered in educational research. Although previous research has investigated the Type I error control and statistical power of parametric and nonparametric statistics (primarily comparisons of the *t*-test and the Wilcoxon-Mann-Whitney *U* test), such comparisons have typically been conducted using continuous outcome variables (see, for example, Blair & Higgins, 1980, 1985). A notable exception is the recent work of Nanna and Sawilowsky (1998), comparing the *t*-test with the Wilcoxon rank-sum test based on resampling from actual data obtained on ordered categorical variables.

Test Statistics Examined

Four test statistics were examined in this study. These test statistics will be presented in reference to the set of data presented in Table 1. These data, consisting of responses to a 5-point Likert item, were obtained from six members of an experimental group and ten members of a control group. The research question to be addressed is whether the two populations from which the samples were obtained differ in their response to this item.

Table 1. Sample of Two Groups' Responses to a 5-Point Likert Item

Control Group	Experimental Group
1	1
1	2
2	3
2	4
2	4
3	5
3	
3	
4	
5	

Independent Means t-test. The independent means *t*-test is used to test the null hypothesis of equivalent population means (H$_O$: $\mu_1 = \mu_2$). The test statistic is given by

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{[(n_1 - 1) + (n_2 - 1)]S_{pl}}$$

where $(\bar{X}_1 - \bar{X}_2)$is the difference in sample means, n_1 and n_2 are the sample sizes, S_{pl} is a pooled estimate of the population standard deviation given by

$$S_{pl} = \sqrt{\frac{(SS_1 + SS_2)}{(n_1 + n_2 - 2)}}$$

and SS$_1$ and SS$_2$ are the sums of squares computed in each of the samples. The obtained value of this test statistic is compared to the sampling distribution of *t* with degrees of freedom equal to $n_1 + n_2 - 2$.

For the sample of data presented in Table 1, the means for the experimental and control groups are 3.167 and 2.600, respectively, and the pooled variance estimate is 1.802. The obtained value of *t* for these data is -0.817, and the probability associated with this value under the null hypothesis is 0.427. The *t*-test, thus, fails to reject the null hypothesis of equal population means.

Chi-Square Test of Homogeneity. In contrast to the *t*-test which compares sample means, the Pearsonian chi-square test of homogeneity tests the null hypothesis of equivalent population proportions in each response category (H$_0$: $\pi_{1j} = \pi_{2j}$, for all *j*). For computation of the chi-square statistic, the data may be arranged in a contingency table as illustrated in Table 2. The sample value of this test statistic is given by

$$\chi^2 = \frac{\sum_i \sum_j (O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed frequency in cell *ij* of the contingency table, E_{ij} is the expected frequency in the cell under the null hypothesis of homogeneity, and the summation is over all of the cells in the table. The obtained value of χ^2 is compared to the sampling distribution of χ^2 with degrees of freedom equal to $(n_{\text{rows}} - 1)(n_{\text{cols}} - 1)$.

Table 2. Contingency Table for the Sample Data

	Response Category				
Group	1	2	3	4	5
1	2	3	3	1	1
2	1	1	1	2	1

For the sample of data presented in Table 1, the obtained value of chi-square is 1.778. In comparison to a four degree of freedom chi-square sampling distribution, this value has a probability of 0.777 under the null hypothesis. Thus, like the *t*-test, the chi-square test fails to reject the null hypothesis of equal population proportions for each level of the response variable.

Delta Statistic. Cliff (1993, 1996a) has proposed the use of the delta statistic for testing null hypotheses about group differences on ordinal level measurements. The population parameter for which such tests are intended is the probability that a randomly selected member of one population has a higher response than a randomly selected member of the second population, minus the reverse probability. That is,

$$\text{delta} = \Pr(x_{i1} > x_{j2}) - \Pr(x_{i1} < x_{j2}) ,$$

where x_{i1} is a member of population one and x_{j2} is a member of population two.

A sample estimate of this parameter can be obtained by enumerating the number of occurrences of a sample one member having a higher response value than a sample two member, and the number of occurrences of the reverse. This gives the sample statistic

$$d = \frac{\#(x_{i1} > x_{j2}) - \#(x_{i1} < x_{j2})}{n_1 \; n_2}$$

This statistic, and inferential methods associated with it, are readily addressed by considering the data in an arrangement called a dominance matrix. This n_1 by n_2 matrix has elements taking the value of 1 if the row response is larger than the column response, -1 if the row response is less than the column response, and 0 if the two responses are identical. The sample value of *d* is simply the average value of the elements in the dominance matrix. The dominance matrix for the Table 1 data is presented in Table 3. The row and column marginals of this table provide mean values of the elements in the respective rows and columns of the matrix. These marginals are used in the inferential statistics associated with *d*. The null hypothesis tested in such inferential statistics (representing no relationship between the grouping variable and the response variable) is that delta is equal to zero.

Table 3. Dominance Matrix for the Sample Data

	1	2	3	4	4	5	*di.*
1	0	-1	-1	-1	-1	-1	-0.833
1	0	-1	-1	-1	-1	-1	-0.833
2	+1	0	-1	-1	-1	-1	-0.500
2	+1	0	-1	-1	-1	-1	-0.500
2	+1	0	-1	-1	-1	-1	-0.500
3	+1	+1	0	-1	-1	-1	-0.167
3	+1	+1	0	-1	-1	-1	-0.167
3	+1	+1	0	-1	-1	-1	-0.167
4	+1	+1	+1	0	0	-1	0.333
5	+1	+1	+1	+1	+1	0	0.833
d.j	0.8	0.3	-0.3	-0.7	-0.7	-0.9	-0.250

Cliff (1996b) presented three methods of inference for *d*. The first method uses an "unbiased" estimate of the variance of *d*. This estimate is given by

$$S_d^2 = \frac{n_2^2 \sum_i (d_{i.} - d)^2 + n_1^2 \sum_j (d_{.j} - d)^2 + \sum_i \sum_j (d_{ij} - d)^2}{n_1 n_2 (n_1 - 1)(n_2 - 1)}$$

where $d_{i.}$ is the marginal value of row i, $d_{.j}$ is the column marginal of column j, and d_{ij} is the value of element *ij* in the matrix.

For the sample data in Table 1, the value of *d* is -0.25 and the value of S_d^2 is 0.098. The square root of this variance is used as the denominator of a *z* statistic: $z_{\text{unbiased}} = d \,/\, S_d$

For the sample data, the value of *z* is -0.798, yielding a probability under the null hypothesis of 0.425. The unbiased test fails to reject the null hypothesis of delta = 0.

The second method of inference for d uses a “consistent” estimate of the variance:

$$S_{dc}^2 = \frac{(n_2 - 1)S_{di.}^2 + (n_1 - 1)S_{d.j}^2 + S_{dij}^2}{n_1 \; n_2}$$

where $S_{di.}{}^2 = \Sigma(d_{i.} - d)^2/(n_1\text{-}1)$, $S_{d.j}{}^2 = \Sigma(d_{j.} - d)^2/(n_2\text{-}1)$, and $S_{dij}{}^2 = \Sigma\Sigma(d_{ij} - d)^2 \,/\, [(n_1 - 1)(n_2 - 1)]$.

As with the "unbiased" estimate of variance, the square root of this "consistent" estimate of the variance of d can be used as the denominator of a z statistic: $z_{\text{consistent}} = d \,/\, S_{dc}$.

For the Table 1 data, the value of $S_{dc}{}^2$ is 0.106, yielding a value for $z_{\text{consistent}}$ of -0.768, with a probability under the null hypothesis of 0.443. The conclusion with this sample is the same as that reached with the unbiased test, that is, a failure to reject the null hypothesis of delta = 0.

The final method of inference regarding *d* uses S_{dc} to construct an asymmetric confidence interval around the sample value of *d*. When such an interval does not include the value of zero, the null hypothesis of delta = 0 can be rejected. The limits of this asymmetric confidence interval are given by

$$\frac{d - d^3 \pm Z_{\alpha/2} S_{dc}[(1 - d^2)^2 + Z_{\alpha/2}^2 S_{dc}^2]}{1 - d^2 + Z_{\alpha/2}^2 S_{dc}^2}$$

where $Z_{\alpha/2}$ is the normal deviate corresponding to the $(1 - \alpha/2)^{\text{th}}$ percentile of the normal distribution.

For the Table 1 data, the lower limit of the 95% confidence interval is -0.713, and the upper limit is 0.364. Because this interval contains the value of zero, the null hypothesis is not rejected at the .05 level.

Cliff (1996a) has pointed out that the well-known Mann-Whitney-Wilcoxon statistic can also provide a test of delta = 0 (because *d* and *U* are related by $d = 2U/[n_1 n_2 - 1]$). However, the rank test is not recommended by Cliff because it is actually testing for the equivalence of the two groups’ distributions rather than focusing on the parameter delta.

Cumulative Logit Models. Logistic regression is a technique used to construct models of the probabilities of values of categorical variables. In its simple, binary form, a model relating the probability of response 1 as a function of an explanatory or predictor variable X, can be thought of as:

$$\pi = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$$

where π is the probability of response 1, exp is the exponential function or the antilog function of the natural logarithms, and α and β are regression parameter estimates. This equation describes an S-shaped curve called the logistic regression model. However, the relationship between π and X is often expressed as logits, yielding the linear logit model: $\text{logit}(\pi) = \log[\pi/(1 - \pi)] = \alpha + \beta X$.

A relatively minor modification of this linear logit model can be used with ordinal response variables having more than two levels (Agresti, 1990; McCullough & Nelder, 1989). With a response variable having J ordered categories, the probability associated with any category j can be denoted π_j, where $\Sigma\pi_j = 1$. The cumulative logit model is formed from logits of cumulative probabilities. For example, the probability of a response less than or equal to an arbitrary category j is given by

$$\begin{aligned}\text{logit}[\Pr(Y \leq j)] &= \log[(\pi_1 + ... + \pi_j)/(\pi_{j+1} + ... + \pi_J)] \\ &= \alpha_j + \beta X\end{aligned}$$

This model treats the response as binary by forming the cumulative probability over the first j categories, and the remaining (J - j) categories. This model has J - 1 values of α_j, one for each of the adjacent category differences. The parameter of primary interest in this model is β, which describes the relationship of the X variable to the cumulative probabilities of response. When β is equal to zero, the variable X is not related to the response variable.

Two methods for testing the null hypothesis that $\beta = 0$ are available. The first method uses the standard error of the sample estimate of β to form a z test (or an equivalent chi-square test), called the Wald test. The standard error of β is obtained from the inverse of the information matrix, the matrix of second partial derivatives of the log likelihood function. For the Table 1 data, the sample estimate of β is -0.834, with a standard error of 0.936. The value of the Wald z test is -0.891, with a probability of 0.373 under the null hypothesis. As with the other tests examined thus far, the Wald test fails to reject the null hypothesis of $\beta = 0$.

The second method of testing the null hypothesis that $\beta = 0$ is with a likelihood ratio test. This test is based on the likelihood ratio statistic:

$$G^2 = 2\Sigma_j O_j \log(O_j/E_j),$$

where O_j and E_j are the observed and expected counts, respectively, and log is the natural logarithm.

The likelihood ratio test of $\beta = 0$ is obtained as the difference in the values of G^2 for the model that includes X, and the model that does not (i.e., a model with intercepts only). This difference in G^2 values is distributed as a chi-square with a single degree of freedom. For the sample data, the value of G^2 for the model that includes X is 49.808, while that for the intercept only model is 50.586. The difference in these G^2 values is 0.737, which has a probability of 0.391 under the null hypothesis. Thus, consistent with the other tests conducted on these data, the likelihood ratio tests does not reject the null hypothesis that $\beta = 0$.

Method

This research was a Monte Carlo study designed to provide an empirical comparison of the Type I error control and statistical power of the four methods of testing group differences on an ordered categorical response variable. Two of these tests are frequently used with ordered categorical data: the independent means t-test and Pearsonian chi-square test of homogeneity. The other two methods, the cumulative logit model and the delta statistic, have been recommended for the analysis of ordinal level data because of increased power (relative to the chi-square test) or increased robustness (relative to tests of mean differences). Although four methods for testing group differences were examined in this study, a total of seven statistical tests were compared (i.e., three tests associated with the d statistic and two tests associated with the logistic regression method).

All of the conditions simulated provided tests of differences between two groups on an ordered categorical dependent variable. Four factors were investigated in the Monte Carlo study: number of categories of the response variable, sample size, population distribution shape, and effect size. The number of categories of the response variable was examined at two levels (5-category and 7-category responses). Six sample sizes were examined (equal sizes of 10:10, 30:30, and 100:100; and unequal sizes of 10:30, 10:100, and 30:100). Four population distribution shapes were investigated (a uniform response distribution, a moderately skewed distribution, a highly skewed distribution, and a unimodal symmetric distribution). Finally, small, medium and large population effect sizes (Cohen, 1988) were examined as well as a null condition. These experimental conditions were crossed with each other providing a total of 192 conditions examined.

Table 4. Type I Error Rate Estimates for 5 Point Response Scale at nominal α = .05

Marginal	Sample	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Square	t-test	Unbiased	Consistent	CI	Wald	LR
1:1:1:1:1	10, 10	0.033	0.055	0.081	0.073	0.043	0.057	0.068
	10, 30	0.042	0.048	0.071	0.067	0.049	0.052	0.057
	10,100	0.045	0.054	0.084	0.082	0.067	0.055	0.059
	30, 30	0.046	0.050	0.057	0.056	0.046	0.053	0.054
	30,100	0.051	0.050	0.056	0.055	0.049	0.051	0.052
	100,100	0.055	0.051	0.054	0.054	0.050	0.052	0.053
6:1:1:1:1	10, 10	0.015	0.050	0.073	0.066	0.046	0.039	0.067
	10, 30	0.046	0.049	0.078	0.076	0.060	0.038	0.061
	10,100	0.053	0.048	0.086	0.085	0.075	0.031	0.059
	30, 30	0.043	0.054	0.060	0.059	0.051	0.053	0.057
	30,100	0.048	0.050	0.056	0.056	0.050	0.047	0.051
	100,100	0.050	0.046	0.048	0.048	0.046	0.046	0.047
16:1:1:1:1	10, 10	0.004	0.033	0.084	0.038	0.035	0.004	0.083
	10, 30	0.038	0.036	0.115	0.114	0.108	0.026	0.070
	10,100	0.083	0.046	0.132	0.132	0.126	0.036	0.101
	30, 30	0.018	0.048	0.054	0.053	0.050	0.034	0.055
	30,100	0.050	0.046	0.062	0.062	0.059	0.035	0.053
	100,100	0.049	0.048	0.051	0.051	0.049	0.047	0.052
1:2:4:2:1	10, 10	0.030	0.049	0.077	0.071	0.042	0.047	0.063
	10, 30	0.048	0.049	0.074	0.071	0.055	0.052	0.057
	10,100	0.049	0.051	0.079	0.078	0.063	0.053	0.054
	30, 30	0.046	0.050	0.058	0.056	0.048	0.050	0.053
	30,100	0.047	0.050	0.057	0.056	0.051	0.051	0.052
	100,100	0.053	0.050	0.053	0.052	0.050	0.051	0.052

Programming for the Monte Carlo Study. The program for the Monte Carlo study was written in SAS/IML version 6.12. The data were generated using uniform random numbers on the zero to one interval (the SAS RANUNI function). A separate seed value was used for each execution of the simulation and the accuracy of the program code was verified using benchmark data sets. To simulate samples, a separate series of random numbers was generated for each of the two groups. The observations were then assigned to values of the ordered categorical response variable based upon the value of the random number.

For example, with a 5-point response scale with equal marginals and an effect size of zero, two series of random numbers were generated. Observations with random numbers between zero and .20 were assigned to the first category of the response variable, those with random numbers between .20 and .40 were assigned to the second category, etc. This procedure yields tables in which the expected proportion in each cell is equal, providing a uniform response across the five categories and the two groups.

The marginal skewness of the response variable was controlled by assigning larger or smaller ranges of the uniform random numbers to each of the ordered categories. For example, to simulate a 60:10:10:10:10 marginal distribution, 60% of the observations were assigned to the first value of the response variable, and 10% to each of the other values. Four marginal distributions were examined in this study. The equal marginal condition provided equal proportions at each level of the response variable. A slightly skewed marginal distribution was produced by generating data in which 60% of the observations were in the first category of the response variable, and the remaining 40% were evenly dispersed over the other values. Similarly, a more highly skewed marginal was produced by generating data in which 80% of the observations were at the first value and the remaining 20% were evenly distributed over the remaining values. Finally, a unimodal symmetric distribution was generated with the mode at the middle of scale and descending proportions of observation for scale values towards the scale endpoints.

Non-null effects were generated by assigning observations to response categories in proportions that differed from the products of the row and column marginal proportions. By varying the extent of discrepancy between the products of the marginals and the actual proportions of observations, effect sizes corresponding to *w* values of 0.10, 0.30, and 0.50 (Cohen, 1988) were produced.

Table 5. Type I Error Rate Estimates for 7 Point Response Scale at nominal $\alpha = .05$

Marginal	Sample	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Square	t-test	Unbiased	Consistent	CI	Wald	LR
1:1:1:1:1:1:1	10, 10	0.020	0.053	0.077	0.068	0.041	0.058	0.066
	10, 30	0.038	0.049	0.070	0.066	0.048	0.055	0.057
	10,100	0.044	0.050	0.084	0.082	0.067	0.054	0.055
	30, 30	0.047	0.052	0.061	0.059	0.048	0.055	0.057
	30,100	0.050	0.051	0.060	0.058	0.052	0.051	0.053
	100,100	0.050	0.048	0.050	0.049	0.047	0.049	0.049
9:1:1:1:1:1:1	10, 10	0.006	0.049	0.072	0.064	0.044	0.037	0.065
	10, 30	0.043	0.048	0.077	0.074	0.061	0.037	0.060
	10,100	0.061	0.044	0.088	0.086	0.076	0.031	0.058
	30, 30	0.032	0.052	0.058	0.057	0.048	0.051	0.055
	30,100	0.048	0.050	0.058	0.058	0.054	0.049	0.054
	100,100	0.045	0.049	0.048	0.048	0.046	0.047	0.048
24:1:1:1:1:1:1	10, 10	0.001	0.028	0.083	0.038	0.034	0.004	0.082
	10, 30	0.037	0.040	0.118	0.117	0.111	0.030	0.069
	10,100	0.096	0.041	0.125	0.125	0.121	0.031	0.093
	30, 30	0.008	0.046	0.056	0.056	0.050	0.035	0.057
	30,100	0.048	0.042	0.063	0.062	0.060	0.035	0.051
	100,100	0.037	0.051	0.052	0.052	0.051	0.048	0.052
1:2:3:8:3:2:1	10, 10	0.018	0.049	0.075	0.070	0.044	0.047	0.063
	10, 30	0.040	0.050	0.073	0.070	0.054	0.053	0.058
	10,100	0.052	0.050	0.078	0.077	0.063	0.054	0.054
	30, 30	0.034	0.054	0.064	0.061	0.051	0.056	0.060
	30,100	0.046	0.051	0.061	0.060	0.053	0.053	0.054
	100,100	0.049	0.053	0.055	0.054	0.051	0.053	0.054

For each of the 192 conditions, 10,000 samples were generated using SAS IML, version 6.12 (SAS, 1992). The use of 10,000 samples provides an adequate level of precision for this study, yielding maximum 95% confidence intervals of ±.0098 around the observed proportion of null hypotheses rejected (Robey & Barcikowski, 1992). For each condition, seven test statistics were computed: (a) the independent means *t*-test, (b) Pearson's chi-square test of homogeneity, (b) Cliff's *Unbiased* test of *d*, (c) Cliff's *Consistent* test of *d*, (d) Cliff's asymmetric confidence interval (CI) for *d*, (e) the Wald test associated with the cumulative logit model, and (f) the likelihood ratio (LR) test associated with the cumulative logit model. Estimates of the Type I error control and the statistical power of each test were conducted at nominal alpha levels of .10, .05, and .01.

Results and Discussion

Before turning to an examination of statistical power, attention must first focus on a comparison of the relative ability of the seven tests to control Type I error. Estimates of Type I error rate were calculated for each of the seven procedures based on 10,000 randomly generated samples for each null condition under examination. Bradley's (1978) liberal criterion of robustness (actual α within $\alpha_{nominal} \pm 0.5\alpha_{nominal}$) was used to evaluate the capacity of each of the seven procedures to control Type I error under the various conditions. To save space, results are provided only for nominal alpha equal to .05. Type I error rates and power estimates for alpha level equal to .10 and .01 are available from the first author.

Estimates of Type I Error Control

Five Point Response Scale. The estimates of Type I error rates for the 5-point scales are provided in Table 4. A broad overview of the robustness of all of the seven tests across all conditions at alpha = .05 is presented in a series of box and whisker plots in Figure 1. The two horizontal lines in this figure are Bradley's limits of robustness. Examination of these plots revealed the *t*-test best able to control Type I error, followed closely by the LR, the Wald test, Cliff's confidence interval, and the chi-square test. Considerably less control was exhibited by Cliff's consistent and unbiased tests. The *t*-test stood alone in its ability to maintain the appropriate level across all conditions. The CI, Wald, LR, and Chi-Square were able to maintain alpha within acceptable limits for all but the most skewed conditions coupled with small and unequal sample sizes. Both the unbiased and consistent tests failed to maintain acceptable control in several instances when small and unequal samples were involved.

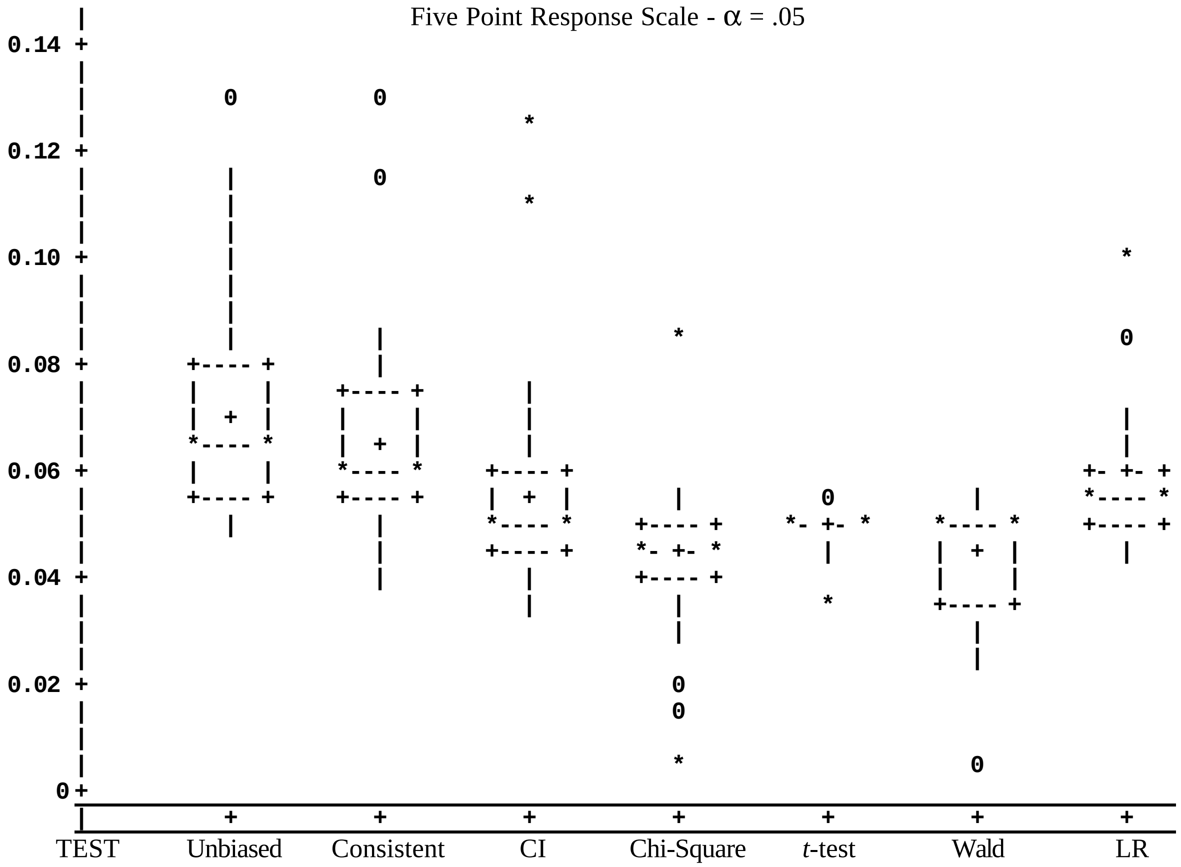

Figure 1. Distribution of Type I Error Rate Estimates for Seven Tests across Experimental Conditions.

Seven Point Response Scale. The estimates of Type I error rates for the 7-point scales are provided in Table 5. The box and whisker plots presented in Figure 2 provide a general overview of the robustness of all seven procedures across all conditions when alpha was set equal to .05. Again, the t-test maintained Type I control across all conditions. Generally, the seven procedures maintained alpha within acceptable limits when large sample sizes were examined. The most skewed condition presented problems for several of the tests, as liberal estimates were observed, on several occasions, for the LR, and Cliff's Confidence Interval, Unbiased, and biased tests. However, there were a few instances in which the Chi-Square and Wald test became conservative. For the unimodal, symmetric distribution, Cliff's Unbiased and consistent tests were liberal only for the unequal sample sizes of 10 and 100, while the Chi-Square test was conservative with the smallest samples.

Estimates of Statistical Power

Five Point Response Scale. Table 6 contains power estimates for the seven procedures. Statistical power estimates are provided only for conditions in which Type I error was controlled. In addition, the Wald test used with the cumulative logit model was not calculable for most samples when the distribution was highly skewed and a non-null condition was simulated (conditions which typically yielded a singular covariance matrix). Estimates of power for only those samples in which it was calculable would be misleading, so these power estimates have also been omitted.

An examination of statistical power at nominal α = .05 revealed the chi-square to be superior to all other tests under the equal marginal and slightly skewed marginal conditions. Under the highly skewed marginal conditions, the Chi-Square was the most powerful only under the largest samples examined. For smaller samples, or unequal samples, other tests were more powerful. For example, Cliff's Consistent test and CI produced the highest power under a highly skewed, small sample condition with a large effect size (power = .625 for both). However, it should be noted that in this instance only one other test, the *t*-test, was able to control Type I error.

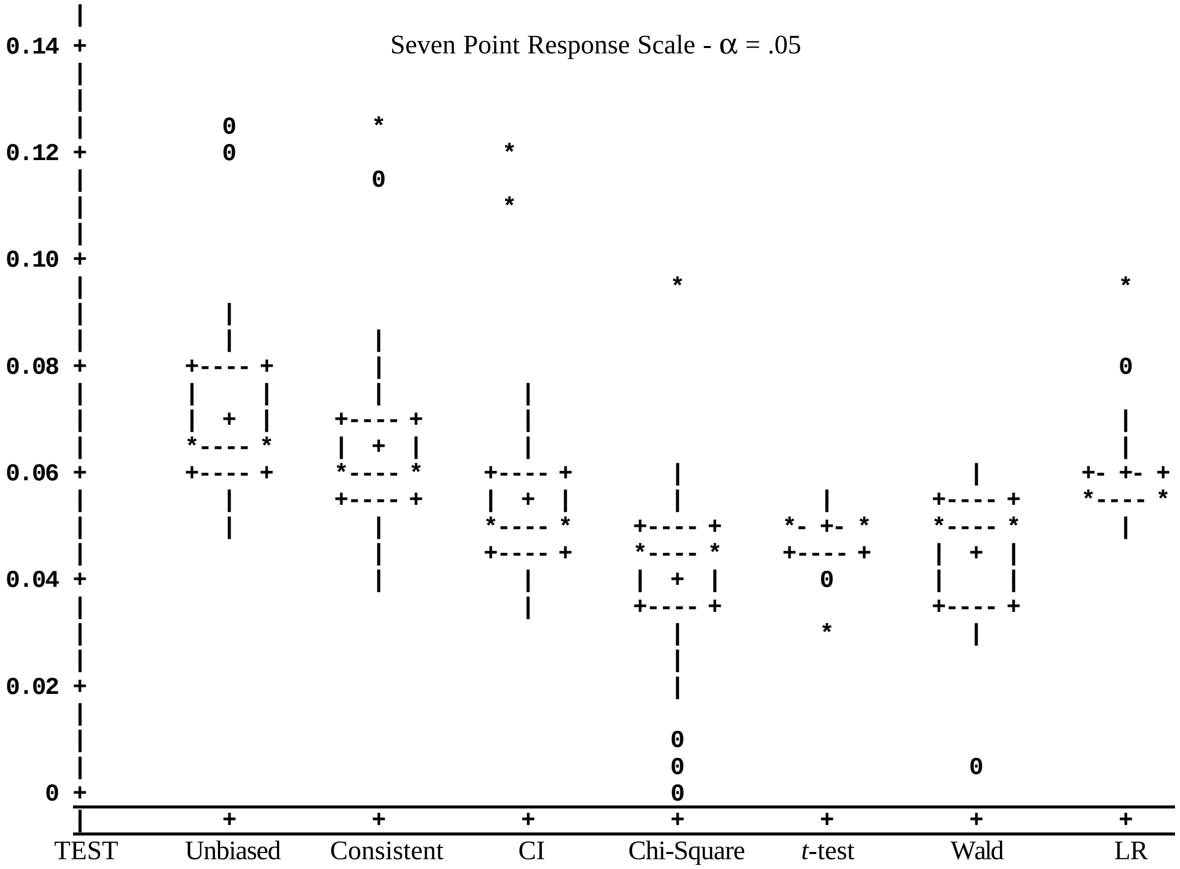

Figure 2. Distribution of Type I Error Rate Estimates for Seven Tests across Experimental Conditions

The *t*-test produced the highest power under highly skewed and unbalanced design conditions, but again, it was one of only two tests that were able to control Type I error under these conditions. For the unimodal, symmetric distribution, the chi-square test was never the most powerful. Rather, for samples drawn from this distribution shape, either the LR test or Cliff's Unbiased or Consistent tests were the most powerful.

Seven Point Response Scale. Table 7 contains power estimates for the seven procedures for nominal alpha level equal to .05. Examination of these results, revealed the Chi-Square to be the most powerful test only under the equal marginal condition, except with small sample sizes. When small sample sizes were examined, Cliff's Consistent test and the LR produced more power than the other tests. Under the slightly skewed and highly skewed marginal distributions, the power produced by several tests was very similar. For example, under the slightly skewed condition with small samples, Cliff's delta tests and the LR produced similar estimates. With larger samples under this condition, it was difficult to choose a superior test from among Cliff's delta tests, the Wald test, or the LR. Similar circumstances surrounded the highly skewed distribution with large sample sizes. For small sample sizes under this condition, the consistent test and CI produced the most power, but many of the tests were unable to control Type I error. For the unimodal, symmetric distribution, the most powerful tests were typically Cliff's Unbiased or Consistent tests. As with the results obtained with the 5-point scales, neither the Chi-Square nor the *t*-test were the most powerful in any sample size condition with this distribution shape.

The differences in the results obtained between the 5-point and 7-point data prompted a further examination of the populations from which samples were generated. Recall that these populations were constructed based on differences between proportions at each scale point to produce desired values of Cohen's *w* (the effect size for differences in population proportions). These populations were examined in terms of the effect size for standardized mean difference (Cohen's *d*) and Cliff's delta. Although the latter is not an effect size, per se, it represents the proportional non-overlap of the two populations from which samples were drawn. These results are presented in Table 8.

Table 6. Statistical Power Estimates for 5 Point Response Scale at nominal α = .05

Marginal	Sample	Effect	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Size	Square	*t*-test	Unbiased	Consistent	CI	Wald	LR
1:1:1:1:1	10, 10	.10	0.040	0.061	-----	0.078	0.048	0.067	0.075
	10, 10	.30	0.115	0.139	-----	0.161	0.104	0.147	0.162
	10, 10	.50	0.345	0.296	-----	0.310	0.226	0.321	0.339
	10, 30	.10	0.061	0.068	0.086	0.082	0.062	0.076	0.079
	10, 30	.30	0.220	0.211	0.206	0.200	0.158	0.242	0.234
	10, 30	.50	0.638	0.482	0.391	0.385	0.307	0.535	0.505
	10,100	.10	0.060	0.074	-----	-----	0.080	0.082	0.083
	10,100	.30	0.310	0.265	-----	-----	0.185	0.298	0.280
	10,100	.50	0.807	0.597	-----	-----	0.337	0.646	0.596
	30, 30	.10	0.080	0.079	0.087	0.084	0.071	0.080	0.083
	30, 30	.30	0.427	0.320	0.330	0.325	0.293	0.336	0.336
	30, 30	.50	0.924	0.710	0.689	0.685	0.651	0.729	0.723
	30,100	.10	0.103	0.108	0.111	0.109	0.099	0.115	0.114
	30,100	.30	0.649	0.482	0.410	0.407	0.383	0.513	0.489
	30,100	.50	0.994	0.888	0.769	0.768	0.743	0.900	0.875
	100,100	.10	0.170	0.142	0.146	0.145	0.140	0.145	0.146
	100,100	.30	0.950	0.777	0.768	0.766	0.759	0.779	0.778
	100,100	.50	1.000	0.996	0.994	0.994	0.993	0.996	0.995
6:1:1:1:1	10, 10	.10	-----	0.058	0.078	0.071	0.048	0.039	0.071
	10, 10	.30	-----	0.111	0.115	0.109	0.080	0.071	0.112
	10, 10	.50	-----	0.225	0.192	0.186	0.143	0.138	0.191
	10, 30	.10	0.045	0.050	-----	-----	0.080	0.031	0.065
	10, 30	.30	0.147	0.113	-----	-----	0.150	0.056	0.119
	10, 30	.50	0.551	0.261	-----	-----	0.284	0.127	0.237
	10,100	.10	0.063	0.047	-----	-----	0.100	0.022	0.063
	10,100	.30	0.243	0.117	-----	-----	0.201	0.036	0.126
	10,100	.50	0.770	0.297	-----	-----	0.370	0.111	0.268
	30, 30	.10	0.058	0.068	0.070	0.069	0.058	0.062	0.068
	30, 30	.30	0.394	0.218	0.192	0.189	0.171	0.185	0.194
	30, 30	.50	0.956	0.508	0.430	0.426	0.400	0.435	0.443
	30,100	.10	0.082	0.072	0.094	0.093	0.087	0.063	0.074
	30,100	.30	0.602	0.290	0.298	0.297	0.286	0.223	0.256
	30,100	.50	0.998	0.662	0.622	0.620	0.605	0.546	0.589
	100,100	.10	0.165	0.109	0.097	0.096	0.093	0.095	0.097
	100,100	.30	0.958	0.572	0.491	0.489	0.481	0.495	0.496
	100,100	.50	1.000	0.952	0.898	0.897	0.893	0.907	0.906

Note that, for the null condition, the populations are identical regardless of how population “differences” are represented. Further, when differences are represented in terms of Cohen's *w*, the 5-point and 7-point populations have identical effect sizes. However, when differences are represented by Cohen's *d*, the effect sizes differ across the two sets, and the difference is not consistent across the distribution shapes. For example, with the "small effect" populations under the slight skew condition, Cohen's *d* was 0.10 for the 5-point data and 0.17 for the 7-point data. A similar difference was evident for the high skew. However, for the unimodal, symmetric distributions, the Cohen's *d* values were nearly identical (0.17 vs. 0.19). Similar differences were noted across the remaining non-null conditions examined. Such discrepancies were also evident when the population differences were measured as Cliff's delta. These observed deviations across effect sizes reflect variations in the magnitude of population differences that result from the design variables of distribution shape and number of scale points. These design variables produced differential effects when inequalities were measured as discrepancies in population standardized mean difference or proportion of non-overlap.

Table 6 (continued). Statistical Power Estimates for 5 Point Response Scale at nominal α = .05

Marginal	Sample	Effect	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Size	Square	*t*-test	Unbiased	Consistent	CI	Wald	LR
16:1:1:1:1	10, 10	.10	-----	0.032	-----	0.038	0.033	-----	-----
	10, 10	.30	-----	0.068	-----	0.058	0.056	-----	-----
	10, 10	.50	-----	0.542	-----	0.625	0.625	-----	-----
	10, 30	.10	0.036	0.025	-----	----	-----	0.015	0.076
	10, 30	.30	0.089	0.035	-----	----	-----	-----	0.116
	10, 30	.50	0.054	0.670	-----	----	-----	-----	0.981
	10,100	.10	-----	0.021	-----	----	-----	0.018	-----
	10,100	.30	-----	0.013	-----	----	-----	-----	-----
	10,100	.50	-----	0.874	-----	----	-----	-----	-----
	30, 30	.10	-----	0.060	0.062	0.062	0.057	0.044	0.063
	30, 30	.30	-----	0.197	0.146	0.145	0.138	0.105	0.147
	30, 30	.50	-----	0.998	1.000	1.000	1.000	-----	1.000
	30,100	.10	0.073	0.056	0.093	0.093	0.090	0.035	0.065
	30,100	.30	0.584	0.212	0.257	0.256	0.251	0.098	0.180
	30,100	.50	0.998	1.000	1.000	1.000	1.000	-----	1.000
	100,10	.10	0.154	0.099	0.083	0.083	0.081	0.078	0.083
	100,10	.30	0.972	0.495	0.340	0.340	0.335	0.331	0.343
	100,10	.50	1.000	1.000	1.000	1.000	1.000	-----	1.000
1:2:4:2:1	10, 10	.10	0.035	0.070	-----	0.095	0.062	0.069	0.089
	10, 10	.30	0.105	0.222	-----	0.283	0.210	0.227	0.274
	10, 10	.50	0.319	0.536	-----	0.625	0.533	0.481	0.628
	10, 30	.10	0.063	0.078	0.115	0.111	0.088	0.086	0.094
	10, 30	.30	0.216	0.299	0.379	0.372	0.322	0.334	0.352
	10, 30	.50	0.654	0.725	0.787	0.783	0.733	0.778	0.799
	10,100	.10	0.064	0.079	-----	-----	0.100	0.086	0.089
	10,100	.30	0.305	0.378	-----	-----	0.411	0.411	0.424
	10,100	.50	0.821	0.821	-----	-----	0.797	0.871	0.872
	30, 30	.10	0.075	0.103	0.123	0.120	0.104	0.111	0.116
	30, 30	.30	0.417	0.540	0.616	0.611	0.581	0.603	0.612
	30, 30	.50	0.934	0.948	0.974	0.974	0.968	0.974	0.975
	30,100	.10	0.103	0.139	0.162	0.161	0.149	0.150	0.153
	30,100	.30	0.650	0.729	0.782	0.780	0.764	0.785	0.787
	30,100	.50	0.995	0.994	0.996	0.996	0.995	0.998	0.998
	100,100	.10	0.166	0.235	0.266	0.265	0.257	0.262	0.264
	100,100	.30	0.954	0.967	0.982	0.982	0.981	0.981	0.981
	100,100	.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Estimates are based on 10,000 samples of each condition. Power estimates are provided only for conditions in which Type I error was controlled.

Conclusions

The results of this research need to be interpreted in the light of the limitations of the study. First, only analyses based on two independent groups were conducted. Although all of the statistical procedures investigated can be extended to multiple group applications, the resulting Type I error rates and power estimates will not necessarily be comparable to those obtained here. Secondly, a limited number of distribution shapes were examined in this study. Extensions to other shapes, such as bimodal distributions, are important areas to explore because distribution shape was seen to influence both Type I error control and the relative power of these tests. Finally, in the consideration of statistical power, the nature of the differences between groups can assume several forms. Although ordered categorical data preclude the consideration of simple shifts in location (because of the boundedness of the response scale), types of non-null effects other than those modeled here need to be investigated.

In light of these limitations, the superiority of the *t*-test and the cumulative logit model in their control of Type I error is evident in these data. Problems with the control of Type I error rates were frequently encountered in conditions with skewed marginal distributions and with unbalanced or small samples. Specific limitations in Type I error control were

Table 7. Statistical Power Estimates for 7 Point Response Scale at nominal $\alpha = .05$

Marginal	Sample	Effect	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Size	Square	t-test	Unbiased	Consistent	CI	Wald	LR
1:1:1:1:1	10, 10	.10	-----	0.053	-----	0.071	0.041	0.059	0.067
	10, 10	.30	-----	0.060	-----	0.079	0.046	0.067	0.077
	10, 10	.50	-----	0.081	-----	0.106	0.064	0.093	0.104
	10, 30	.10	0.054	0.055	0.076	0.072	0.052	0.059	0.062
	10, 30	.30	0.181	0.073	0.093	0.087	0.066	0.083	0.084
	10, 30	.50	0.548	0.103	0.126	0.120	0.090	0.119	0.121
	10,100	.10	0.064	0.053	-----	-----	0.063	0.061	0.059
	10,100	.30	0.265	0.078	-----	-----	0.081	0.097	0.090
	10,100	.50	0.757	0.127	-----	-----	0.111	0.156	0.145
	30, 30	.10	0.067	0.052	0.059	0.056	0.047	0.053	0.055
	30, 30	.30	0.347	0.083	0.095	0.092	0.077	0.088	0.091
	30, 30	.50	0.868	0.143	0.160	0.155	0.136	0.151	0.155
	30,100	.10	0.084	0.056	0.064	0.063	0.055	0.058	0.058
	30,100	.30	0.578	0.111	0.114	0.112	0.102	0.122	0.119
	30,100	.50	0.988	0.212	0.205	0.203	0.187	0.230	0.223
	100,100	.10	0.137	0.065	0.068	0.067	0.064	0.066	0.067
	100,100	.30	0.923	0.165	0.171	0.170	0.163	0.168	0.169
	100,100	.50	1.000	0.378	0.383	0.381	0.372	0.383	0.384
9:1:1:1:1	10, 10	.10	-----	0.064	0.092	0.084	0.059	0.048	0.084
	10, 10	.30	-----	0.207	0.293	0.278	0.218	0.172	0.278
	10, 10	.50	-----	0.521	0.665	0.646	0.568	0.394	0.654
	10, 30	.10	0.017	0.065	-----	0.132	0.116	0.048	0.100
	10, 30	.30	0.028	0.260	-----	0.424	0.385	0.241	0.382
	10,100	.10	0.023	0.060	-----	-----	----	0.030	0.106
	10,100	.30	0.022	0.296	-----	-----	----	0.295	0.441
	10,100	.50	0.350	0.761	-----	-----	----	0.670	0.879
	30, 30	.10	0.048	0.094	0.117	0.114	0.102	0.107	0.114
	30, 30	.30	0.317	0.521	0.631	0.626	0.603	0.615	0.628
	30, 30	.50	0.862	0.939	0.981	0.981	0.978	0.972	0.982
	30,100	.10	0.036	0.120	0.181	0.180	0.170	0.141	0.158
	30,100	.30	0.406	0.680	0.806	0.805	0.791	0.797	0.807
	30,100	.50	0.971	0.990	0.998	0.998	0.997	0.992	0.999
	100,100	.10	0.139	0.229	0.283	0.283	0.276	0.279	0.282
	100,100	.30	0.916	0.958	0.986	0.986	0.985	0.986	0.986
	100,100	.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000

observed for the tests of delta suggested by Cliff (1993, 1996a). Of special interest is that in many conditions, Cliff's Confidence Interval approach to inferences regarding delta were superior to the two *z* test approaches examined. The asymmetric approach to the confidence interval estimation appeared to improve the control of Type I errors in several of the conditions examined in this study. However, for researchers working with small samples or unequal sample sizes, the *t*-test or cumulative logit model appear to be the tests of choice to maintain Type I error control.

Finally, in terms of statistical power, although the independent means *t*-test provided the best control of Type I error rates across the conditions examined, this test was rarely the most powerful, and, consequently, should not be the first choice in most applications. For the 5-point response scales, the chi-square test of homogeneity was clearly the most powerful test for those conditions in which it maintained Type I error control. In contrast, for the 7-point scales, the Chi-Square test was only the most powerful when the marginal distribution was symmetric. For the skewed marginal distributions, the cumulative logit models or the tests of delta tended to be the most powerful. However, the variation in power across these scales should not be interpreted as a simple function of the number of scale points. Rather, such variations represents changes in the magnitude of the population differences in terms of standardized mean difference or proportion of non-overlap of the populations.

Table 7 (continued). Statistical Power Estimates for 7 Point Response Scale at nominal α = .05.

Marginal	Sample	Effect	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Size	Square	*t*-test	Unbiased	Consistent	CI	Wald	LR
24:1:1:1:1	10, 10	.10	-----	0.046	-----	0.063	0.057	-----	-----
	10, 10	.30	-----	0.173	-----	0.232	0.217	-----	-----
	10, 10	.50	-----	0.506	-----	0.622	0.622	-----	-----
	10, 30	.10	0.011	0.024	-----	-----	-----	0.007	0.126
	10, 30	.30	0.001	0.145	-----	-----	-----	-----	0.441
	10, 30	.50	0.004	0.620	-----	-----	-----	-----	0.982
	10,100	.10	-----	0.011	-----	-----	-----	0.005	-----
	10,100	.30	-----	0.107	-----	-----	-----	-----	-----
	10,100	.50	-----	0.818	-----	-----	-----	-----	-----
	30, 30	.10	-----	0.105	0.129	0.128	0.119	0.090	0.131
	30, 30	.30	-----	0.575	0.694	0.692	0.677	0.535	0.699
	30, 30	.50	-----	0.999	1.000	1.000	1.000	-----	1.000
	30,100	.10	0.017	0.100	0.224	0.223	0.220	0.094	0.167
	30,100	.30	0.129	0.707	0.894	0.894	0.890	0.708	0.859
	30,100	.50	0.960	1.000	1.000	1.000	1.000	-----	1.000
	100,10	.10	0.120	0.233	0.288	0.288	0.284	0.278	0.288
	100,10	.30	0.920	0.972	0.992	0.992	0.992	0.991	0.992
	100,10	.50	1.000	1.000	1.000	1.000	1.000	-----	1.000
1:2:3:8:3:2:1	10, 10	.10	-----	0.072	-----	0.098	0.066	0.072	0.089
	10, 10	.30	-----	0.245	-----	0.307	0.232	0.244	0.293
	10, 10	.50	-----	0.611	-----	0.680	0.596	0.539	0.667
	10, 30	.10	0.050	0.082	0.116	0.111	0.089	0.091	0.096
	10, 30	.30	0.208	0.369	0.410	0.402	0.345	0.401	0.408
	10, 30	.50	0.632	0.810	0.799	0.792	0.741	0.821	0.837
	10,100	.10	0.082	0.087	-----	-----	0.098	0.095	0.094
	10,100	.30	0.352	0.432	-----	-----	0.405	0.472	0.464
	10,100	.50	0.882	0.891	-----	-----	0.802	0.910	0.902
	30, 30	.10	0.054	0.114	0.131	0.126	0.111	0.118	0.123
	30, 30	.30	0.310	0.602	0.653	0.645	0.612	0.634	0.643
	30, 30	.50	0.876	0.978	0.984	0.983	0.979	0.981	0.983
	30,100	.10	0.088	0.139	0.156	0.154	0.144	0.150	0.152
	30,100	.30	0.624	0.803	0.806	0.804	0.790	0.831	0.829
	30,100	.50	0.994	0.998	0.997	0.997	0.997	0.999	0.998
	100,100	.10	0.136	0.259	0.282	0.280	0.272	0.276	0.279
	100,100	.30	0.919	0.982	0.987	0.987	0.987	0.987	0.987
	100,100	.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Estimates are based on 10,000 samples of each condition. Power estimates are provided only for conditions in which Type I error was controlled.

Further, the power differences among these procedures were small suggesting that researchers' choices may be based on the types of interpretations that are appropriate for the research questions being addressed. For interpretations based on simple dominance, the d statistics and their inferential tests would be the most appropriate. In contrast, a more rigorous modeling of response probabilities is provided by the cumulative logit models.

In summary, ordered categorical data, such as those investigated in this study, are frequently encountered in educational research. Unfortunately, the analysis strategies most frequently employed with these types of data are not necessarily the best strategies to use. This research has provided information about the operating characteristics (Type I error control and statistical power) of the commonly used tests employed with ordered categorical data, and has provided evidence of the advantages (in some data conditions) associated with two recently recommended options for testing hypotheses. Although additional research is certainly needed to further explore the performance of these tests and their limitations, this initial examination suggests that for many data conditions, the choice of an appropriate test statistic is vitally important to the validity of research inferences.

Table 8. Indices of Differences in the Simulated Populations

Population		Index of Group Difference					
Group	Marginal	Effect size *W*		Effect Size *d*		Cliff's *d*	
Differences	Distribution	5-point	7-point	5-point	7-point	5-point	7-point
Null Model	Uniform	0.00	0.00	0.00	0.00	0.00	0.00
	Slight Skew	0.00	0.00	0.00	0.00	0.00	0.00
	High Skew	0.00	0.00	0.00	0.00	0.00	0.00
	Unimodal Sym	0.00	0.00	0.00	0.00	0.00	0.00
Small Effect	Uniform	0.10	0.10	0.13	0.05	0.07	0.03
	Slight Skew	0.10	0.10	0.10	0.17	0.05	0.10
	High Skew	0.10	0.10	0.09	0.18	0.03	0.08
	Unimodal Sym	0.10	0.10	0.17	0.19	0.10	0.11
Medium Effect	Uniform	0.30	0.30	0.39	0.14	0.21	0.08
	Slight Skew	0.30	0.30	0.31	0.53	0.14	0.29
	High Skew	0.30	0.30	0.28	0.57	0.09	0.24
	Unimodal Sym	0.30	0.30	0.54	0.58	0.31	0.32
Large Effect	Uniform	0.50	0.50	0.67	0.23	0.36	0.13
	Slight Skew	0.50	0.50	0.52	0.95	0.23	0.49
	High Skew	0.50	0.50	1.40	1.36	0.40	0.40
	Unimodal Sym	0.50	0.50	1.00	1.05	0.51	0.54

Correspondence should be directed to
Jeffrey D. Kromrey
Educational Measurement & Research,
University of South Florida
4202 East Fowler Avenue, Tampa, FL 33620

References

Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychological Bulletin*, *105*, 290-301.

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.

Agresti, A. & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Blair, R. C. & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of the student's t statistic under various non-normal distributions. *Journal of Educational Statistics*, *5*, 309-335.

Blair, R. C. & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, *97*, 119-128.

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

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*, 494-509.

Cliff, N. (1996a). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, *31*, 331-350.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Davidson, M. L. & Sharma, A. R. (1988). Parametric statistics and levels of measurement. *Psychological Bulletin*, *104*, 137-144.

McCullough, P. & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

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

Robey, R. R. & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283-288.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: John Wiley.

Velleman, P. F. & Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *American Statistician*, *47*, 65-72.