Think Different.
Comments on Alternative Regression Procedures

T. Mark Beasley, Guest Editor
St. John's University

I chose Apple Computers' slogan, not because I happen to be a Macintosh user, but because the issues raised in these three articles should lead us to "Think Different" as statisticians, data analysts, and researchers. One key issue underlying these articles is the ultimate question, "What are the data trying to tell us?" Several statistics texts have used the signal-to-noise analogy for analyzing data. Therefore, if we are simply trying to detect a signal amongst random, ambient noise then it does not seem as problematic to transform the data or to perform alternative procedures that potentially test different statistical hypotheses. If exact parameter estimation is of interest, however, data transformations may lead to interpretive difficulties.

Nevitt and Tam (*pp*. 54-69) approach this issue from the parameter estimation perspective of: What should be done in order to detect an *accurate* signal if the data are not "well behaved" or do not conform to the statistical assumptions of the regression model? These authors examine three general approaches for estimating parameters when data are not well behaved (i.e., nonnormal): (a) treat outliers differently (i.e., Trimming, Winsorizing), (b) transform the data (i.e., Monotonic Regression), or (c) compute parameter estimates in a different manner (i.e., LAD, Theil estimators).

The authors make an important distinction between robust and nonparametric estimators. Robust methods were developed for situations in which *symmetric* error distributions have heavy tails due to outliers in the observed data. Thus, the normality assumption is simply relaxed. Robust estimators are therefore resistant to violation of assumptions while testing the same null hypothesis as the normal theory methods (Draper & Smith, 1981). By contrast, nonparametric and distribution-free methods may involve (a) transforming data to ranks or other metrics or (b) computing the parameter estimate in an entirely different way. Therefore, the normality assumption may not apply whatsoever. In these cases, the statistical hypothesis tested, although conceptually similar, may be quite different than the hypothesis evaluated by a normal theory counterpart. Because of this difference, the performance of nonparametric methods relative to OLS methods is often hard to assess except under conditions where many parameters (i.e., skew and kurtosis) are held to normal theory assumptions which of course favors OLS procedures.

Recall the question posed in the foreword (*p*. 2), "How do these techniques integrate with what is already known about statistics?" There are extremely interesting relationships between OLS and nonparametric estimators of slope. By using the geometric definition of a regression slope and taking the $n(n - 1)$ pairwise slopes,

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i} \text{ where } x_i \neq x_j,$$

the Theil estimator of slope is the median of all the b_{ij} slopes. Interestingly, when all values of X are distinct, Sprent (1993) demonstrates that significance testing of the Theil median slope is based on Kendall's (1970) tau statistic which is related to Cliff's (1994) ordinal multiple regression (Long, *pp*. 45-53). This can be inferred from the fact that $n(n - 1)$ pairwise values are used in both procedures. Other relationships can be shown by making an aggregate of these slopes such that

$$\hat{\beta} = \frac{\sum_{i<j} w_{ij}\, b_{ij}}{\sum_{i<j} w_{ij}} \quad ,$$

OLS regression defines the weight as $w_{ij} = (x_j - x_i)^2$. Other nonparametric approaches defines $w_{ij} = |x_j - x_i|$ (Birkes & Dodge, 1993) which reduces to a sign function for the X variable over the sum of the absolute deviations of X (Huynh, 1978). For Kendall's tau, the weights would be defined in terms of the absolute value of both the Y and X deviations (i.e., $w_{ij} = |x_j - x_i|/|y_j - y_i|$) which would then reduce to a sign function for both Y and X. Kendall's tau is the simple (i.e., unweighted) average sign of the $n(n - 1)$ pairwise slopes.

In terms of Nevitt and Tam's methodology, it is questionable whether the sequential series of X is realistic. First of all, the sequential series of X is a uniform distribution. This implies that Nevitt and Tam are examining fixed-effects models because the underlying assumption of a random-effects model is that Y and X are sampled from a bivariate normal distribution (Hays, 1994). Although fixed-effects models are applied most commonly, even when random-effects are of interest (Clark, 1973), the use of a uniform distribution as the basis for the parameter model seems realistic only if the population relationship among ranks is of interest. Secondly, in the population, the uniform distribution of X yields a uniform distribution for Y through the linear transformation described in the methods section (*p*.

57). When a normal, random error component is added to Y then the conditional distribution of Y is normal which again is adequate for a fixed-effects model. In this case however, the overall distribution of Y is neither uniform nor normal which violates the bivariate normality assumption of a simple linear regression random-effects model. Thus in regression applications where random-effects models are of interest, the fact that the data for Y are nonnormal could stem from either (a) the structural component (i.e., population distributions of X and Y are nonnormal), (b) the error component being nonnormal, or (c) from both (a) and (b).

Thus, from this fixed-effects perspective, Nevitt and Tam's methodological approach assumes that "bad" (i.e., nonnormal) data originates from the error distribution of a regression model. Therefore, they rightfully suggest that "data analyses should always involve checking for outliers in the observed data and testing the underlying assumptions under OLS estimation" (*p*. 68). The idea that outliers and heteroscedasticity may stem from nonnormal error distributions is certainly interesting and leads to the question: How does one know if the error distribution is normal when the data are nonnormal? The possibilities are that: (a) the variable itself is nonnormal; (b) there are outliers present in a symmetric error distributions; or (c) the error distribution is skewed or nonnormal. Thus, there is an important distinction between: (1) a normal distribution with outliers that create skewness and (2) a skewed distribution such as reaction time. Nevitt and Tam's results show that as expected (Draper & Smith, 1981), robust estimators perform better under condition (1) but nonparametric methods perform better under condition (2). Nevitt and Tam report the surprisingly good overall performance of the Theil estimator. Furthermore, the Theil estimates were accurate especially with nonnormal error distributions. As expected with contaminated normal error distributions, the robust procedures (i.e., LAD, Trimming, Winsorizing) performed well. As a personal bias, however, I am not fond of Trimming because this form of discarding data creates a situation where the data are systematically missing which is know to lead to biased estimates.

One astounding and important result, mainly because of the common application of rank transformations, was the poor performance of Monotonic Regression. This finding should be viewed in a certain light, however. The authors note that their results substantiate the unacceptability of rank transformation in the form of Monotonic Regression with respect to Bias and root mean square error (RMSE). Namely, large Bias values reflect the inability of Monotonic Regression to "recover the true population values" (see *p*. 67). The fact that Nevitt and Tam used sequential values of X would seem to have benefited Monotonic (rank) Regression because the transformation was linear for X. With the addition of a random error distribution, however, the rank transformation for Y was not linear in most cases. Thus, Monotonic Regression did not perform well in general. Yet, procedures that transform the original data should not be expected to perform as well. How would rank values transform back to the original metric of Y if sequential X values were not used? Furthermore, one must consider that Monotonic Regression tests a different null hypothesis; it tests OLS hypotheses in the metric of ranks.

As with the Brockmeier et al. article (*pp*. 20-39), if the purpose of a study is simply to establish a relationship (i.e., just detecting the signal) then finding non-zero correlations (or standardized regression slopes) is the major issue rather than exact parameter estimation. Perhaps the rank transform procedure (Monotonic Regression) would not perform so poorly in these circumstances (e.g., a simulation study where Type I error and Power rates, instead of estimation bias, would be reported). Yet, if exact parameter estimation is of interest then the precision of both α and β parameter estimates is important. The RMSE and Bias reported by Nevitt and Tam are both valuable indicators because procedures such as Monotonic Regression can maintain stable Type I error rates and demonstrate superior power (e.g., Harwell & Serlin, 1989) yet provide consistently bad parameter estimates. Thus, it would appear that rank (as well as other non-linear) transformations are not appropriate when exact parameter estimates are to be "recovered." By contrast, if researchers are merely attempting to establish a relationship, then they could consider the signal-to-noise analogy where transformations (and other alternative approaches) do not seem so disabling.

To elaborate, in experimental designs and other group comparison research, ANOVA models that test for mean differences are employed. In contrast to single-sample statistics where a relevant population parameter must be known *a priori*, the fact that there is a comparison group makes the signal more detectable. One may think of this in terms of perceptual research which has demonstrated that judging the length or orientation of an object is much easier when there are perceptual cues that allow for comparisons (e.g., Witkin & Goodenough, 1981). Using the same analogy, violations of assumptions and other data problems can be viewed as the factors that create perceptual (statistical) distortions and illusions, and thus, the use of statistics in many behavioral research contexts may be seen as a field-dependent endeavor. Similar to the ANOVA model, a linear regression model is a comparison of means in the sense that as X increases the expected value of Y increases by the slope on average. Again, if one is simply trying to detect a signal, rather than estimating a parameter precisely, then the fact that Y generally increases with increases in X may be good enough. And it does not matter too much how the variables are expressed.

Popular sources such as Tabachnick and Fidell (1996) discuss transforming data when assumptions are violated. This is even more systematized as the "ladder of re-expression" when power and logarithmic transformations are used to transform data (Hoaglin, Mosteller, & Tukey, 1983). Yet most researchers have problems with such nonlinear transformation with the exception of the rank transform concept. That is, unlike taking the square root of a variable to quell an outlier or reduce asymmetry, ranking the data still retains the "meaning" of the data to many researchers (Zimmerman, 1996). Furthermore, there are conveniences because there are many rank-based tests already in existence and ranks have known means and variances. Despite these conveniences, Nevitt and Tam's results are consistent with other research (Zimmerman, 1996, 1998; Zimmerman & Zumbo, 1993) that has demonstrated serious problems in applying rank transformations. Therefore, the reliance on rank transformation may be somewhat superstitious because its historical prevalence and intuitive appeal are more convincing than empirical evidence showing its statistical viability for estimating parameters.

One issue that all researchers have with any re-expression is what do the data "mean" after transformation. For the sake of symmetry in a variable, a researchers may be left with the question: What does the square root of achievement scores mean? Cliff (1996) argues that researchers usually do not want their conclusions to be confined to the current, somewhat arbitrary version of the variables. Moreover the current measurements are often assumed to be manifest versions of latent variables that are not linearly related to them. Therefore, a poignant question for researchers to ask would be: What did my scores mean in the first place? From this perspective the central question of data analysis can be posed as: What question should I be asking? That is, the null hypotheses associated with OLS regression may not be what is really of interest. Cliff (1993, 1996) contends that most of the answers behavioral researchers want to get from their data are ordinal ones. Furthermore, most of the observed variables have only ordinal justification, at least as measures of the theoretical constructs they are used to represent. Therefore, because the questions asked are ordinal and the data are ordinal, ordinal methods are suggested.

Based on this perspective, **Long** (*pp*. 45-53) explicates another less common re-expression, the transformation of data into what Cliff (1993) has termed the "dominance" metric. Many of us have been familiarized with this concept through Kendall's (1970) measure of concordance. Not only does this notion lead to testing statistical hypotheses that are different from their OLS counterparts, the procedures require us to "Think Different" because the hypotheses are different conceptually. From the Pearsonian perspective, relationships are an issue of the average value of Y conditional on X. From the dominance perspective, however, relationships are expressed as the proportional alignment of Y with X. In terms of group comparisons where the OLS solution involves an ANOVA model, ordinal methods address what proportion of scores in group one are larger than the scores in group two. In terms of a linear regression, they assess what proportion of Y scores become larger as X increases.

Marascuilo and McSweeney (1977, *pp*. 439-440) discuss Kendall's tau as a measure of concordance and as a measure of correlation. However, Kendall's tau as a measure of correlation is not interpreted in the Pearsonian sense but as a measure of "array." That is, it is an index of the amount of agreement between two sets of ranks. When teaching the Pearson product-moment correlation, I prefer demonstrating the *z*-score formula and discussing the Pearson *r* as an averaged leverage (i.e., product-moment) value. Similarly, the notion of the dominance metric is appealing because it allows a perspective of what Kendall's tau (as well as ordinal multiple regression and Cliff's *d* statistic) actually measures. Thus, from the dominance matrix, it can be seen that Kendall's tau measures the proportional agreement between the dominance scores on two variables. Therefore, Kendall's tau coefficient, as a summary measure, is an average of proportional increase.

As is the case with OLS regression, a second predictor makes the interpretation more complicated but there are analogies in ordinal multiple regression (OMR). However, there are some unresolved issues in OMR which again force us to "Think Different." First of all, OMR does not yield truly partialled values. Similar to Marascuilo and McSweeney's discussion of the relationship of Kendall's tau to Pearson's *r*, one cannot interpret the coefficients that result from OMR as OLS regression weights. Furthermore, although Kendall (1970) developed a "partial tau," its properties are quite different from those in OLS. For example, suppose there are three variables that have positive intercorrelations and a trivariate normal distribution. Although the first two are statistically independent, conditional on the third ($r_{12.3} = 0$), Kendall's partial tau will not be zero (Cliff, 1996). Thus, Long rightfully warns that the "OMR function is much more ambiguous in its specification of the relationship between the weights and the criterion. In fact, an algebraic formula expressing the criterion in terms of the weighted predictors is not possible" (*p*. 47). This means that there is no final "regression equation" where a line or plane of best fit is described. Predicted values for each subject are not rendered. Long states that it would be possible "if $\hat{d}_{ihy}$ were used in the loss function" (*p*. 47). To elaborate, one can use either equation (4) or (6) and calculate, $\hat{d}_{ihy} = .40(d_{ih1}) + .33(d_{ih2})$, a "prediction equation" for the $n(n - 1)$ pairwise dominance scores. Then based on these $n(n - 1)$ predicted values an "average predicted dominance score" of the form $\hat{d}_{iy} = \sum_h(\hat{d}_{ihy})/(n - 1)$ can be computed for each of the n subjects. It can be shown that both $\hat{d}_{ihy}$ and $\hat{d}_{iy}$ sum to zero as would standardized predicted values from an OLS regression. However, this approach violates the logic of ordinal analysis. Therefore, only a verbal description of the functional relationship between the weights and the criterion is appropriate. Thus, the OMR weights are the constants that when applied to the predictor dominance scores best predict order on the criterion, "best" meaning that Q is optimal (*p*. 47). Thus using equation (5), $Q = .5945$; however, this is only the optimization of the weights. That is, Q can be viewed as analogous to Multiple R^2, but it is not the "variance accounted for" typically associated with OLS regression. Furthermore, to date there is not an omnibus test for Q analogous to the F-test for the full model R^2. Given the confidence interval approach taken by Long this may be less problematic. Still Q is only a statistic descriptive of the loss function. Thus, in its current state OMR has many statistical and interpretive limitations despite the compelling arguments of Cliff (1996). Possibly the OMR methodology forces us to "Think *too* Different." When applied to group comparison research, however, the dominance metric approach has many foreseeable advantages. In research in which two or more groups are compared, ANOVA models are applied to test differences in means which addresses the question: "Do the groups have different average values?" Cliff (1993) suggests that through using the dominance metric one can answer the question many behavioral researchers *really* want to ask, "Which group has higher scores?" Yet a similar and even more general question is: "Did the groups respond differently?"

Kromrey and Hogarty (*pp*. 70-82) address the differences among these three questions. They present an interesting situation in which two groups are compared on an ordered categorical response, as opposed to analyzing a dependent variable that is truly continuous in nature. This is a common practice in a variety of educational and psychological studies where Likert-type responses are elicited and groups are subsequently compared. Kromrey and Hogarty evaluate the statistical properties of four general procedures (*t*-test, Pearson chi-square test, Cliff's *d*, and Cumulative Logit model). They contend that despite the differences among the statistical null hypotheses tested, each of these procedures may be used to test the same, "conceptual" research hypothesis (*p*. 70). Although these methods may seem to address conceptually similar research questions, statistically they are *not* the same. Thus, a review of the procedures, their null hypotheses, and the questions addressed should be examined carefully.

Again, the most general question is, "Did the groups respond differently?" It is most likely to be addressed with the Pearson chi-square test for contingency tables which for two groups has the following null hypothesis:

$$H_{O(\pi)}: \pi_{1k} = \pi_{2k}, \text{ for all } k \text{ categories.}$$

The question of "Which group has higher scores?" is often thought of terms of the *t*-test. However, this issue is actually more in line with Cliff's *d* statistic. It tests the null hypothesis that the probability that a randomly selected member (i) of one population has a higher response than a randomly selected member (j) of the second population is equal to the reverse probability. That is, the probability that the scores from one group are higher minus the probability that a second group's scores are higher is equal to zero:

$$H_{O(\delta)}: \delta = \Pr(y_{i1} > y_{j2}) - \Pr(y_{i1} < y_{j2}) = 0.$$

These population probabilities are measured by the frequencies in the samples. It should be noted that the *d* statistic is equivalent to Kendall's tau performed with a dummy code representing the group distinctions, and thus, Cliff's *d* can extend into multiple group and factorial designs (Cliff, 1996).

The most specific of the three research questions is, "Do the groups have different average values?" It is addressed by the independent samples *t*-test with the following null hypothesis:

$$H_{O(\mu)}: \mu_1 - \mu_2 = 0.$$

A fourth approach investigated by Kromrey and Hogarty is a Cumulative Logit model suggested by Agresti (1989). In the current situation, this method treats the categorical response as an ordinal variable and the grouping variable as dichotomous. The impetus for the Cumulative Logit model is that the Pearson chi-square test was designed for variables that have unordered categories. Therefore, it detects any type of deviation from the null hypothesis $H_{O(\pi)}$. If the variable is ordinal, however, the categorical data may be represented with fewer degrees-of-freedom which for a fixed noncentrality structure increases the statistical power of a test. Thus, the Cumulative Logit model detects only monotonic deviations but these are the ones of most importance with ordinal variables (Agresti, 1989, *p*. 298).

To explicate this approach, Beasley and Schumacker (1995) demonstrated a method for orthogonally partitioning a contingency table using ANOVA contrast codes. One thing not pointed out by Beasley and Schumacker is that in the situation presented by Kromrey and Hogarty, the contingency table can be partitioned in order to test for mean differences (i.e., $H_{O(\mu)}$). Suppose a linear polynomial contrast (i.e., [-2 -1 0 1 2] for 5 categories, [-3 -2 -1 0 1 2 3] for 7 categories) is applied to the categorical variable. If this contrast variable is weighted by the frequencies and then correlated with a dummy code representing the two groups, the result is identical to the *t*-test. Therefore, with 5 ordered categories and two groups, the null hypothesis for the linear polynomial contrast (ψ) of population proportions in a contingency table,

$$H_{O(\psi)}\!: -2\pi_{11} -1\,\pi_{12} +0\,\pi_{13} +1\,\pi_{14} +2\,\pi_{15}$$
$$+2\pi_{21} +1\,\pi_{22} +0\,\pi_{23} -1\,\pi_{24} -2\,\pi_{25} = 0\ ,$$

is equivalent to evaluating differences in population means, $H_{O(\mu)}$: μ_1 - μ_2 = 0. The Cumulative Logit model uses a similar approach (Agresti, 1989, *p*. 294); however, a Logit model instead of a Pearsonian model is used. Thus, the null hypothesis associated with the cumulative Logit model ($H_{O(\beta)}$: $\beta = 0$, see *p*. 73 or Agresti, 1989 for details), although not identical to $H_{O(\mu)}$, is extremely similar in concept. Differences among these statistical hypotheses will be discussed later.

When evaluating the performance of these four procedures, one must consider that any test of statistical significance has assumptions. The assumption that each of the observations are *independent* of each other applies to all these procedures. Importantly, the independent *t*-test has the additional assumptions that the two groups are sampled from identical (i.e., *homogeneous variances*), *normal* (i.e., skew and kurtosis of zero) populations. The assumption of homogeneous variances translates into the notion that the group effect is "additive." As a point of distinction, the Cumulative Logit model can be interpreted as the multiplicative effect of the grouping variable on the cumulative odds. Because these odds for cumulative probabilities are expressed in logits, however, this multiplicative effect can be interpreted as "additive." The Cumulative Logit model investigated by Kromrey and Hogarty implies a uniform association of cumulative odds ratios and is referred to as the "Proportional Odds" model (Agresti & Finlay, 1997, *p*. 601). Therefore, this Cumulative Logit Model assumes that the group effect is the same for each cumulative probability, an assumption analogous to an additive model (Agresti, 1989, *p*. 293). Furthermore, under the conditions imposed by Kromrey and Hogarty, the Cumulative Logit model performed similarly to the *t*-test in terms of Type I error (e.g., Table 4, *p*. 74) and power rates (e.g., Table 6, *p*. 78). Unlike the *t*-test, however, the independence of the variables corresponds to the distribution of the response variable being identical, not necessarily *identical and normal*. Therefore, similar to the Pearson chi-square and Cliff's *d*, which are relatively "distribution-free," the Cumulative Logit model makes no assumption about the shape of the response variable. Moreover, like the Pearson chi-square and Cliff's *d*, it can be sensitive to differences in variance and shape even when population means are identical. Thus, the Cumulative Logit model tests a statistical hypothesis that is different from $H_{O(\mu)}$, a topic explicated later.

Kromrey and Hogarty's results confirmed that the Pearson chi-square test should not be used with small sample sizes which accentuates the need for an alternative procedure such as the Cumulative Logit model that reduces the hypothesis degrees-of-freedom in a contingency table analysis. That aside, it is interesting that most tests were generally acceptable for testing the null hypothesis of identical population distributions, but the *t*-test gave the most consistent Type I error rate (see Fig. 1, *p*. 76). The Type I error results (e.g., Table 4, p. 74) also showed that even when the conditional distribution of the dependent variable was highly skewed, the *t*-test was generally robust to violations of the normality assumption thus confirming the seminal work of Norton (1952, cited in Lindquist, 1956) and Boneau (1960). It should be noted, however, that under the conditions simulated the conditional distributions for *Y* were identical in that the population values for variance, skew, and kurtosis, as well as the population means, were the same for both groups. Therefore, the null hypotheses for all procedures were true. Thus, because of the robustness of the *t*-test to violations of the normality assumption, the three research questions are considered the same if the groups have identical distributions in terms of variance, skew, and kurtosis. However, one must consider that outside of violating the normality assumption, the Type I error simulation conditions favored the parametric *t*-test (i.e., identical conditional distributions). Although the Type I error results are valid, they are limited in the sense that there are many situations in which some of the null hypotheses are false while others are true. Moreover, it is difficult to reconcile one procedure being more "robust" when they have different assumptions. That is, a test cannot be robust to a condition for which it makes no assumption (Huber, 1991).

The Power results were even more difficult to interpret because in the conditions simulated, all three null hypotheses were false but to different extents (see Table 8, *p*. 82). Furthermore, because some tests are sensitive to different parameters, a researcher may confirm the "conceptual" research hypothesis for a variety of reasons. For example, the Pearson chi-square test was powerful because it can detect a variety of differences (i.e., mean, variance, skew, kurtosis). By contrast, the *t*-test detects very specific differences. It is designed to detect differences in means but can be sensitive to differences in variance. Thus, as compared to evaluating the robustness of these tests, it is even more problematic to discern which is most "powerful" when the null hypotheses tested are different. To elaborate, the chi-square null hypothesis is the most general. Consequently, if $H_{O(\pi)}$ is true, then $H_{O(\mu)}$, $H_{O(\delta)}$, and $H_{O(\beta)}$ are also true. However, a true $H_{O(\mu)}$ does not imply that $H_{O(\pi)}$ is true. Likewise, a true $H_{O(\delta)}$ does not imply that $H_{O(\delta)}$ is true. This is also the case for $H_{O(\beta)}$.

Imagine the following tables show the population probabilities (π_k) for each of the K = 5 ordered categories in each group. Situation One is identical to the moderately skewed distribution condition simulated by Kromrey and Hogarty for assessing Type I error rates.

Situation One

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	60	10	10	10	10
Group 2	60	10	10	10	10

In this case all four null hypotheses are true. By contrast, imagine the following scenario.

Situation Two

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	44	4	4	4	44
Group 2	10	20	40	20	10

In this case $\mu_1 = \mu_2 = 3$, and thus, $H_{O(\mu)}$ is true. $H_{O(\delta)}$ and $H_{O(\beta)}$ are also true. However, $H_{O(\pi)}$ is false again demonstrating that the Pearson chi-square test of $H_{O(\pi)}$ is sensitive to parameters other the mean differences. It should also be noted that the homoscedasticity assumption of the *t*-test is violated in that $\sigma^2_1 = 3.6$ and $\sigma^2_2 = 1.2$. Therefore, the *t*-test may not maintain a Type I error rate near the nominal alpha in this case (i.e., it can be sensitive to differences in variance).

In the following scenarios the difference between the *t*-test, Cliff's *d*, and the Cumulative Logit Model can be further demonstrated.

Situation Three

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	6	34	30	14	16
Group 2	16	14	30	34	6

In this case the two distributions have identical values for the population mean ($\mu_1 = \mu_2 = 3$), variance ($\sigma^2_1 = \sigma^2_2 = 1.36$), and kurtosis ($\gamma^4_1 = \gamma^4_2 = -0.80$). Therefore, the null hypothesis for the *t*-test is true. The population skews are different ($\gamma^3_1 = -0.39$, $\gamma^3_2 = 0.39$). Furthermore, $H_{O(\pi)}$, $H_{O(\delta)}$, and $H_{O(\beta)}$ are false. Importantly, Cliff's *d* and the Cumulative Logit Proportional Odds model can be sensitive to differences in skew even when population means and variances are identical. However, if the means are the same and both distributions are symmetric, not necessarily identical (e.g., Situation Two) both $H_{O(\delta)}$ and $H_{O(\beta)}$ are true (see Vargha & Delaney, 1998, for a discussion of what they call Stochastic Homogeneity).

To further accentuate how Cliff's ordinal method and Agresti's Cumulative Logit model forces us to "Think Different," all four population moments are different ($\mu_1 = 3.0$, $\mu_2 = 3.1$; $\sigma^2_1 = 0.96$, $\sigma^2_2 = 3.60$; $\gamma^3_1 = -0.34$, $\gamma^3_2 = 0$; $\gamma^4_1 = -0.55$, $\gamma^4_2 = -1.98$) in Situation Four, but $H_{O(\delta)}$ and $H_{O(\beta)}$ are true.

Situation Four

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	4	36	32	22	6
Group 2	44	4	4	4	44

Also, imagine a situation where Cliff's *d* would be equal to 1.0. That is, every subject in Group 1 (n_1 = 100) has a higher score than every subject in Group 2 (n_2 = 100). Furthermore, suppose that 50 people in Group 1 responded to category 5 and the other 50 endorsed category 4.

Situation Five

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	0	0	0	50	50
Group 2	0	50	50	0	0

In terms of maintaining a Cliff's *d* of 1.0, it does not matter what pattern of 3, 2, or 1 categories is endorsed by Group 2.

Situation Six

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	0	0	0	50	50
Group 2	50	0	0	0	0

That is, regardless of whether Group 2 responds to categories 3 and 2 only (Situation Five) or all of them endorse category 1 (Situation Six), Cliff's *d* would still be equal to 1.0. Thus, in this scenario, the *d* statistic can be contrasted with the *t*-test in the sense that Cliff's *d* considers rank position and dominance rather than average magnitude. Although the Cumulative Logit Proportional Odds model is somewhat sensitive to these differences in magnitude (Agresti, 1989), in general it seems more similar to Cliff's *d* than to the *t*-test, at least statistically.

It would be interesting to see how the Cumulative Logit model performs empirically under the various conditions elaborated, especially with between group differences in variance and skew (e.g., Situations Two through Four). Agresti (1989, *p.*

294) indicates that the independence of X and Y (i.e., $H_{O(\beta)}$ is true) corresponds to the distribution of the ordered categorical response (Y) being the same for each level of X (the grouping variable). In Situation Two, however, the β parameter is zero although the distributions of the ordered categorical responses are not identical for both groups which presents a violation of the Proportional Odds model. Therefore, it would also be interesting to determine whether the Cumulative Logit model performs more similar to Cliff's *d*, the Pearson chi-square, or to the *t*-test under such conditions.

Because of these statistical issues, the conceptual differences, and other previously elaborated arguments, Cliff (1996) contends that δ (and *Q* for OMR) are *NOT* just surrogates for OLS solutions; they are parameters worth estimating in their own right. From this perspective, Cliff's *d*, Kendall's tau, and OMR make parametric use of "nonparametric" statistics. For such statistical procedures, Bradley (1968) suggested the term "distribution-free," while Cliff prefers the term "ordinal methods."

Bradley (1968) and Zimmerman (1996) have pointed out that much of the confusion concerning the use of nonparametric methods lies in the treatment of nonparametric tests as "different" in most textbooks when actually many nonparametric tests are often algebraic reduction of OLS parametric tests performed on ranks (or signs or a dominance matrix). Under the basic assumptions of parametric tests, ranks have known means and variances. This allows the parametric formula to simplify which in turn makes it seem different. However, many of the problems associated with the original data can be inherited by the ranks (Zimmerman, 1996, 1998). Therefore, they may not be as "robust" as commonly believed. It is also true that the ordinal methods (i.e., Cliff's *d*, OMR) are OLS solutions for the dominance matrix (see Long, *p*. 46). Yet, the dominance matrix is a transformation that partially changes the *meaning* of the score. Therefore, the associated hypotheses are different both statistically and conceptually. Given its statistical similarity to Cliff's *d* , the same may also be said for the Cumulative Logit model.

The differences among the statistical hypotheses of parametric and alternative procedures has been seen as a drawback to employing "nonparametric" methods. Yet, Cliff (1996) argues that the hypotheses tested by alternative methods are often more in line with what behavioral researchers want to know from their data as compared to a null hypothesis of equal means. The point is that mean differences may not always be of interest (Olejnik, 1987). For example, in a randomized experiment if differences in variances occur then, an ANOVA model (*t*-test) may be inappropriate because a non-additive effect is suggested. That is, differences in variance indicate that the treatment did something to change the variability and thus a test of means may not be entirely appropriate. Furthermore, heterogeneous variances may also indicate some non-additive, interaction effect that has not been examined. This emphasizes the importance of data screening, data exploration, descriptive statistics, and graphical display in order to evaluate "What the data are trying to tell us." Moreover, instead of employing parametric statistical tests ritualistically, perhaps researchers should "Think Different" and perform alternative procedures. Again, the conclusions may be similar conceptually. Yet, there is the distinct possibility that the results from an alternative procedure may force investigators to "Think Different" about their research questions.

Address correspondence to:
T. Mark Beasley
School of Education
St. John's University
8000 Utopia Parkway
Jamaica, NY 11439
E-Mail: beasleyt@stjohns.edu

References

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

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

Beasley, T. M., & Schumacker, R. E. (1995). Multiple regression approach to analyzing contingency tables: Post-hoc and planned comparison procedures. *Journal of Experimental Education, 64*, 79-93.

Birkes, D., & Dodge, Y. (1993). *Alternative Methods of Regression*. New York: Wiley.

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

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

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

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

Cliff, N. (1994). Predicting ordinal relations. *British Journal of Mathematical and Statistical Psychology, 47*, 127-150.

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

Draper, N., & Smith, H. (1981). *Applied Regression Analysis* (2nd edition). New York, NY: Wiley.

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

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

Huber, P. (1991). *Robust Statistics*. New York: Wiley.

Huynh, H. (1978). A comparison of four approaches to robust regression. *Psychological Bulletin, 92*, 505-512.

Kendall, M. G. (1970). *Rank Correlation Methods* (4th ed.). London: Charles Griffin.

Lindquist, E. F. (1956). *Design and analysis of experiments in psychology and education* (2nd ed.). Boston: Houghton-Mifflin.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks-Cole.

Sprent, P. (1993). *Applied nonparametric statistical methods*. London: Chapman & Hall.

Norton, D. W. (1952). *An empirical investigation of some effects of non-normality and heterogeneity on the F-distribution*. Unpublished doctoral dissertation, State University of Iowa.

Olejnik, S. (April, 1987). Teacher education effects: Looking beyond the means. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational & Behavioral Statistics, 23*, 170-192.

Witkin, H. A., & Goodenough, D. R. (1981). *Cognitive styles: Essence and origins*. New York: International Universities Press.

Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *Journal of Experimental Education, 64*, 351-362.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67*, 55-68.

Zimmerman, D. W., & Zumbo, B. (1993). Relative power of the Wilcoxon test, the Friedman test, and the repeated-measures ANOVA on ranks. *Journal of Experimental Education, 62*, 75-86.