# An Investigation of Ordinal True Score Test Theory

John R. Donoghue, Educational Testing Service

Norman Cliff, University of Southern California

The validity of the assumptions underlying Cliff's (1989) ordinal true score theory (OTST) were investigated in a three-stage study. OTST makes only ordinal assumptions about the data, and provides a means of converting ordinal item information into summary ordinal information about examinees. Stage 1 was a simulation based on a classical (weak true score) test theory model. Stage 2 used a long empirical test to approximate the true order. Stage 3 was an extensive simulation based on the three-parameter logistic model. The results of all three studies were consistent; the assumption of local ordinal uncorrelatedness was violated in that partial item-item gamma ($\gamma$) correlations were positive instead of 0. The assumption of proportional distribution of ties was violated—pairs tied on one item were not distributed on the other as prescribed. The item-true order tau ($\tau$) correlation was consistently overestimated, although the estimated $\tau$ correlated highly with the true $\tau$. The $\tau$ correlation between total score and true order was also consistently overestimated. Stage 3 showed that these effects occurred under all conditions, although they were smaller under some conditions. *Index terms: classical test theory, item response models, local independence, monte carlo simulation, nonparametric test models, ordinal regression, ordinal test models, test theory.*

In many circumstances, test items provide only ordinal information about the trait being measured. For example, reading comprehension or vocabulary items often require the examinee to select the best response from available alternatives; the keyed response is "correct" only in that it is in some sense preferable to the alternatives. The uses of test data are often ordinal as well. The primary goal of testing is frequently to provide some type of ranking (e.g., percentile) of examinees. Other scaled scores, such as IQ or $T$ scores, are often useful only to the extent that they provide implicit information about an examinee's standing (order) relative to some meaningful reference group.

As a result, there have been repeated attempts to formulate a truly ordinal theory of testing. Early advocates of the position were Guttman (1950) and Loevinger (1947). More recent papers by Mokken (1971), Schulman and Haden (1975), Schulman (1976), Cliff (1977, 1979), Mokken and Lewis (1982), Stout (1988), and Cliff and Donoghue (in press) indicate that there is ongoing interest in this area. Recently, Cliff (1989) advanced ordinal true score theory (OTST). OTST makes only ordinal assumptions and provides a means of converting ordinal item information into summary ordinal information about examinees.

## Purpose

This study examined the validity of the assumptions under which OTST was derived. This was accomplished in a three-stage study. Stage 1 was a small simulation study in which data were generated according to a classical (weak) true score model and then analyzed using the procedures of OTST. Stage 2 used empirical data and investigated the effect of improving the estimate of the true order. In Stage 3, data were generated according to a three-parameter logistic item response theory (IRT) model and then analyzed using OTST procedures. The relationship between IRT parameters and OTST assumptions and estimates was examined. The reasons to expect the assumptions of OTST to hold when data are generated according to a different test model are

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 15, No. 4, December 1991, pp. 335-351*

*0146-6216/91/040335-17$2.10*

discussed below.

### Overview of OTST

Similar to many current test theories, OTST begins by assuming that a single, unidimensional latent ability underlies an examinee's responses to test items. It differs from most current formulations in that it is assumed that ability (i.e., the trait) is defined only up to a monotonic transformation. Hence, only ordinal statements can be made about ability. Ability is known as the "true order," emphasizing the state of knowledge about the trait. The formulation also assumes that the items provide ordinal information about the underlying trait. Thus, the goal of the model is to estimate the ordinal relations between examinees' orders on the items and the true order of examinees, and to describe how best to combine the ordinal information of the items (as in a total score) to make inferences about the true order of the examinees.

OTST considers only the probabilities of one examinee ranking ahead of another. The ranks as such are not used. Thus, the theory is consistent with the use of Kendall's tau ($\tau$) and the Goodman-Kruskal gamma ($\gamma$), rather than Spearman's rho.

$\tau$ describes the probability of agreement of the ordering of pairs of examinees. Consider the pair of examinees $i$ and $h$, each measured on two attributes $X_j$ and $X_k$. On variable $X_j$, examinee $i$ may have a larger value than $h$, $i$ may have a smaller value than $h$, or the two examinees may be tied. The same is true for $X_k$. $\tau$-A (Kendall, 1970) describes the probability that examinees $i$ and $h$ are ordered the same on the two variables. Ties are considered failures to order the two examinees. Thus, $\tau$ is the number of pairs ordered the same minus the number of pairs that are ordered oppositely, divided by the total number of pairs:

$$\tau_{jk} = p_{\text{same}} - p_{\text{opposite}} \quad . \tag{1}$$

$\gamma$ (Goodman & Kruskal, 1954) assesses the probability that a pair of examinees are ordered the same, conditional on the pair being ordered. No penalty is exacted for ties:

$$\gamma_{jk} = \frac{p_{\text{same}} - p_{\text{opposite}}}{p_{\text{same}} + p_{\text{opposite}}} \quad . \tag{2}$$

### Assumptions of OTST

Let $Y$ be the latent, true order of examinees on the trait being measured. It is assumed that $Y$ is a full order; there are no ties on the latent true order. Let $X_j$ ($j = 1, \ldots, p$) be scores on items on a test designed to measure $Y$. OTST has three underlying assumptions, all of which involve the local relations of items.

Local relations are ordinal relations of pairs of examinees, conditional on the true order of those examinees. Assume that the true order of a pair of examinees on $Y$ is known. Then the pair's scores on item $X_j$ may be in the correct order (+), incorrect order (–), or the pair may be tied (0) on $X_j$. The same is true for another item $X_k$. Thus, for a pair of items, each pair of examinees falls into one of nine possible categories of local relations. These pairwise relations may be collected into a 3 × 3 table of local relations, as shown in Table 1. This table is not a usual contingency table, because entries refer to *pairs* of examinees. For example, $S_{++}$ is the proportion of pairs correctly ordered by both $X_j$ and $X_k$, and $S_{0-}$ is the proportion of pairs tied on item $X_j$ and ordered incorrectly on item $X_k$.

**Table 1**
Local Relations Between Two Items

| | Item $X_k$ | | | |
|---|---|---|---|---|
| Item $X_j$ | + | 0 | – | Total |
| + | $S_{++}$ | $S_{+0}$ | $S_{+-}$ | $S_{+.}$ |
| 0 | $S_{0+}$ | $S_{00}$ | $S_{0-}$ | $S_{0.}$ |
| – | $S_{-+}$ | $S_{-0}$ | $S_{--}$ | $S_{-.}$ |
| Total | $S_{.+}$ | $S_{.0}$ | $S_{.-}$ | 1.00 |

The marginal (zero-order) item-item relations may be calculated from the cells of the table:

$$\gamma_{jk} = \frac{(S_{++} + S_{--}) - (S_{+-} + S_{-+})}{(S_{++} + S_{--}) + (S_{+-} + S_{-+})} \quad , \tag{3}$$

and item-true order relations may be calculated from the table's margins; for example, $\tau_{jY} = S_{+} - S_{-}$. Finally, as a direct extension of Kendall's (1970, pp. 117–122) development of partial $\tau$, Table 1 gives item-item ordinal relations with true order partialed out, for example,

$$\gamma_{jk|Y} = \frac{S_{++}S_{--} - S_{+-}S_{-+}}{S_{++}S_{--} + S_{+-}S_{-+}} \quad . \tag{4}$$

Table 1 is useful to illustrate the three assumptions of OTST. The first assumption of OTST is called *local ordinal uncorrelatedness*, and states that the conditional $\gamma$ correlation of two items is 0:

$$\gamma_{jk|Y} = 0 \quad . \tag{5}$$

The property of local ordinal uncorrelatedness is based on order, without regard to magnitude. Thus, local uncorrelatedness was proposed as a weaker version of local independence and, therefore, would hold whenever local independence held.

The last two assumptions are collectively termed *proportionality of ties*. They state that for all items $X_j$, the pairs that are tied on $X_j$—that is, give the same response—are assumed to occur on every other item $X_k$ in a manner proportional to the marginal distribution of pairs on $X_k$:

$$\frac{S_{0+}}{S_{0-}} = \frac{S_{.+}}{S_{.-}} \quad , \tag{6}$$

and conversely,

$$\frac{S_{+0}}{S_{-0}} = \frac{S_{+.}}{S_{-.}} \quad . \tag{7}$$

The three assumptions given in Equations 5 through 7 could be replaced by a single assumption that might be termed *local ordinal independence.* This assumption states that each cell in Table 1 is the product of the marginal quantities (i.e., $S_{+-} = S_{+.}S_{.-}$). This assumption is overly restrictive, however. It unnecessarily specifies the value of $S_{00}$ and restricts the values of the other cells that contain ties (e.g., $S_{+0}$) beyond the extent needed for the OTST derivations. The OTST formulation assumes only what is necessary for the derivations.

From these assumptions, Cliff (1989) has demonstrated that the interitem $\gamma$ correlation is the product of each item's $\gamma$ with the true order:

$$\gamma_{jk} = \gamma_{jY}\gamma_{kY} \quad . \tag{8}$$

Equation 8 implies that when OTST holds, a single factor will underlie the matrix of interitem $\gamma$ correlations. Further, the loadings on that factor give the $\gamma$ correlations between the examinees' order on the item and the true order of examinees, *Y.* These factor analyses are completely ordinal and are justified by the derivation of Equation 8 from the assumptions.

Using concepts borrowed from ordinal multiple regression (Cliff, 1986), Cliff has used the assumptions of Equations 5 through 7 to develop expressions for the ordinal reliability ($\tau$ correlation of the total score with the true order) and for the ordinal reliability of the optimally weighted total score. These relationships hold under two conditions of homogeneity: (1) where the conditional gamma ($\gamma_{jk|Y}$) = 0 for all $j,k$; and (2) a less restrictive case where the weighted average conditional $\gamma$ is equal to 0. If condition (1) is true, the $\gamma$ matrix should be unifactor. Condition (2) is intended for the common situation where a test is composed of different, but highly correlated, factors, such as a verbal test composed of vocabulary and reading comprehension items. In either case, Equation 8 indicates that the loading on the first centroid factor of the $\gamma$ matrix provides the estimated $\gamma$ correlation of the item with the true order. These estimated $\gamma$ correlations provide the basis for estimating the $\tau$ correlation of items with true order and, in turn, the $\tau$ correlation of observed total score with true order.

This study explored OTST by examining the degree to which the assumptions of the theory (local ordinal uncorrelatedness and proportionality of ties) were met in generated and empirical datasets. The degree to which the estimates of item-true order and total score-true order relations corresponded to the known characteristics of the artificial datasets was also examined.

## EMPIRICAL STUDIES

The assumptions of the OTST model were examined in three ways. In Stage 1, data were constructed according to a classical (weak true score) model for a homogeneous test and then analyzed using OTST. In Stage 2, empirical samples of 40 verbal items and 29 math items, each administered on two occasions, were analyzed. The empirical samples of verbal items and math items were used to investigate the effect of improving the estimate of the true order. Stage 3 was a large simulation study designed to investigate the conditions under which OTST corresponded to the IRT model (Birnbaum, 1968; Lord, 1952) and under which the models yielded different results.

### Stage 1: Generated Data Based on A Classical Test Model

#### Method

The generated data were tests of 40 items constructed according to a homogeneous test model. Individual items were formed according to a common factor model. For each dataset, 200 standard normal N(0,1) deviates, the true order, were randomly generated. Forty "errors" were constructed, each an independent N(0,1) variable $e_j$. Continuous item scores were then constructed by forming weighted combinations of these 40 error variables with the true order $Y$, that is,

$$z_{ij} = v_j y_i + e_{ij}(1 - v_j^2)^{1/2} \quad . \tag{9}$$

The weights $v_j$ were constant and either moderate (.5) or large (.8). The observed score matrix **X** was formed by dichotomizing item scores, with an average probability of passing each item fixed at .5,

$$x_{ij} = \begin{Bmatrix} 1, \ z_{ij} \geq 0 \\ 0, \ z_{ij} < 0 \end{Bmatrix} \quad . \tag{10}$$

Although it is not intuitively obvious, this model is mathematically identical to an IRT two-parameter normal ogive model:

$$P(X_{ij} = 1|Y_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(\frac{z^2}{2}\right) dz \tag{11}$$

where $x = a_j(Y_i - b_j)$

with a difficulty parameter $b_j = 0.0$ and discrimination parameter

$$a_j = \frac{v_j}{(1 - v_j^2)^{1/2}} \quad . \tag{12}$$

The discrimination parameter $a_j$ describes the steepness of the item response function (IRF), the function given in Equation 11, at the point $Y = b_j$. High values of $a_j$ indicate sharp discrimination between those examinees above and below $b_j$. The difficulty parameter $b_j$ describes the point of maximum slope (discrimination) for an item. The value of the difficulty parameter corresponds to the ability for which the probability of correct response is exactly .5. See Lord and Novick (1968, pp. 376–378) for a proof of the equivalence of these two models. An implication of this equivalence is that the IRT assumption of local independence holds for the classical test theory model as presented in Equation 9.

Classical test theory indicates that it is best for all items to be close to a probability correct of .5, and therefore of approximately the same difficulty. The data were generated as a best case according to classical test theory. If the assumptions of OTST are not compatible with such data, it is less likely that they will work with less optimal data.

The generated data were analyzed according to OTST using a program written by the authors. The accuracy of the program was insured by comparing results from subroutines to published analyses and output from commercial statistical packages. The "true score" used to generate the data, $Y$, defined the true order for OTST analyses. To test Assumption 1, the $\gamma$ between items conditional on the order on $Y$ was calculated for each pair of items. The assumption of proportionality of ties was assessed by calculating the average ratio ($Q_j$) of the conditional over the marginal occurrence of ties:

$$Q_j = \frac{\sum_{k \mp j}^{p} \left( \dfrac{S_{0+}/S_{0-}}{S_{.+}/S_{.-}} \right)}{p-1} \quad , \tag{13}$$

where $p$ is the number of test items. The numerator of the terms of $Q_j$ is the proportion of pairs tied on item $X_j$ and ordered correctly on item $X_k$, divided by the proportion of pairs tied on $X_j$ and ordered incorrectly on $X_k$. The denominator is the total (i.e., marginal) proportion of pairs ordered correctly on $X_k$ divided by the total proportion of pairs ordered incorrectly. If the assumption in Equation 6 is true, this ratio has a value of 1 for all items, which follows directly from Equations 6 and 7.

The implication of a unifactor γ matrix was tested by factor analyzing the interitem γ matrix. From the loadings, γ and τ correlations of items with true order were estimated by the model and compared with the actual correlations. Finally, the estimated τ of total observed score with true order was calculated for both uniform and optimal weighting of items, and compared to the OTST estimates. Only two datasets were generated for each consistency level because the results were so clear that additional replications were deemed unnecessary.

## Results

The first assumption, that the partialed γ is 0, was contradicted by the data. For the moderate loading case, the average partialed γ was .221 (compared to an average of .340 for the marginal γ), but it was .603 (compared to an average .764 for the marginal γ) for the high loading datasets. Thus, ordinal partialing based on the common factor order relations had little effect on reducing the γs. This is inconsistent with Equation 5, which states that the conditional γ is 0. This result suggests that there is a fundamental difference between local independence and local ordinal uncorrelatedness. This issue is discussed in more detail below.

For the assumption of proportionality of ties, the assumption was consistently violated. The average value of $Q_j$ was .928 for the moderate loading condition and .673 for the high loading condition. For every item of every dataset generated, the ratio $Q_j$ was less than 1.0. This indicates that the conditional proportions for ties were less than the assumption would predict. The conditional pluses and minuses were closer to equal than were the marginals.

Surprisingly, the interitem γ matrix was found to be substantially unifactor. For the moderate loading condition, the eigenvalue of the first factor was 12.8 and the second was 1.7. For the high loading case, the average first value was 30.7 and the second was .95. Despite the clear violation of all three assumptions, the factor analysis of the γ matrix appears to remain a good indicator of the underlying dimensionality of the data, at least in the equidifficulty case examined. This result indicates that although the OTST assumptions in Equations 5, 6, and 7 are sufficient for Equation 8 to hold, they may not be necessary. This issue is discussed in more detail below.

OTST yields an estimate of the γ of an item with the true order, the loading on the first centroid factor of the interitem γ matrix (Cliff, 1989).

**Table 2**
Average OTST Estimates and Observed Values for Test Quantities: Results of Stage 1

| | Moderate Loading Condition | | High Loading Condition | |
|---|---|---|---|---|
| Test Quantity | OTST Estimate | Observed Value | OTST Estimate | Observed Value |
| $\tau_{jY}$ | .281 | .232 | .437 | .393 |
| $\tau_{XY}$ | .985 | .780 | .953 | .892 |
| Optimally Weighted $\tau_{XY}$ | .986 | .781 | .958 | .895 |

This in turn yields an estimate of the $\tau$ of the item with the true order. Table 2 shows that the estimated $\tau$s of true order with items were uniformly higher than were the actual $\tau$s of true order with items. Yet, in spite of this bias, a close relation was found between the actual and estimated $\tau$s with true order; Pearson correlations ranged from .883 to .961.

The estimated $\tau$ correlation of the true order with total score was calculated, as was the estimated $\tau$ between true order and an optimally weighted combination of items. Table 2 gives the means for these variables. Both of these estimated $\tau$s were consistently larger than the actual $\tau$ between total score and the common factor. Taken as a group, the results of analyzing the data generated according to classical test theory indicate that the assumptions of classical test theory are inconsistent with those of OTST; when one theory is true, the other must be false (except in the trivial case where all item intercorrelations are 0).

## Stage 2: Analysis of Empirical Data

### Method

Two empirical tests were used: a 40-item vocabulary and a 29-item math test. Each test was administered twice, in the fall and the spring of the same school year. The sample was obtained on a random sampling basis from the sixth grade norms distributions of a nationally standardized achievement battery, and was obtained from a well-known test publisher. Data from 321 students with complete data for both administrations were used. The sample was equally divided between males and females.

A set of five items was selected from the fall administration to serve as the "test." The selected items were uniformly spaced to span the length of the full test.

For each scale, the interitem correlations across administrations of the test averaged approximately the same as the within-time interitem correlations. Based on preliminary factor analyses of both $\gamma$ and $\phi$ interitem correlations, and on reliability analyses, each of the scales was deemed to be highly consistent across administrations. Thus, the size of the item pool defining true order was increased by combining the two administrations (fall and spring) to form a single item pool.

Because the data were empirical, the true order of examinees was not known. However, given the internal reliability of the scales, the full set of items for each scale was deemed to be a reasonable approximation of the true order on that scale.

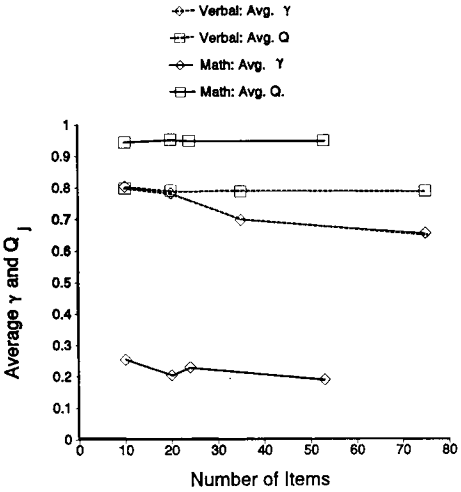
### Results

The assumption that the partialed $\gamma$ correlation is 0 was found to be violated by these data also. The average partialed $\gamma$s were .654 for the verbal items and .190 for the math items.

To determine if this result was an effect of having only an imperfect "true order," the use of even more imperfect true orders—defined by only a subset of the items—was tried. If it was found that the assumption held even more poorly for less well-defined "true orders," it might suggest that by extrapolation a true order defined by a 1,000-item test or item pool might be consistent with the assumption. Consequently, "true orders" were defined as follows: (1) 10 items from the fall administration, (2) 20 items from the fall administration, (3) all the remaining items at fall administration (35 items for the verbal data, 24 for math), and (4) all the remaining items for both administrations (75 items for the verbal data, 53 for math). These analyses sought to determine the effect on the validity of the assumptions of using a progressively better estimate of the true order.

As Figure 1 shows, the partialed $\gamma$ did decrease (albeit very slowly) toward meeting the assumption as the number of items used to define the pseudo true order increased. For example, the partialed $\gamma$ for the 10-item verbal true order was .805, but for the 75-item true order it was .654. This pattern appeared to hold true for the math items, although it was not as clear. There was an

**Figure 1**
The Relationship of Partialed γ and the Ratio $Q_j$ to Number of Items Defining True Order for Math and Verbal Items



increase in the average partialed γ between the 20-item and 24-item definitions of true order, but this increase was relatively small, and partialed γ for the 53-item definition was less than that of both the 20-item and 24-item definitions of true order. However, the decline appears to be too slow to suggest that the bias would be 0 for any reasonable definition of true order.

Figure 1 also shows that the assumption of proportionality of ties was violated in the data. The ratio $Q_j$ was consistently approximately .95 for the math items and approximately .80 for the verbal items, indicating—as in the simulation—that the conditional distribution of pluses and minuses was more nearly equal than were the marginal distributions. The ratio of proportionality of ties held constant across true order estimates; the maximum difference in $Q_j$ was .02 across the various definitions of true order within any given dataset. Thus, it appears that the ratio is a characteristic of the data.

As was the case with the generated data, the factor analysis of the interitem γ matrix indicated that the matrix was substantially unifactor, despite violation of the assumptions. There was only one interitem γ matrix because there was only one "test" involved. The estimated τ of item with true order was consistently larger than the actual τ, although the relationship again appeared to be linear. The estimated τ of total score with true order was again found to be much higher than the actual correlation. As with the partialed γ, the actual correlation did increase toward the theoretical estimate with increasing length of true order, although the increase was relatively slow. As was the case for the data generated according to classical test theory, these results indicate that although the OTST assumptions in Equations 5, 6, and 7 are sufficient for Equation 8 to hold, they may not be necessary. This issue is discussed in more detail below.

Finally, the test-retest γ correlations between the full verbal (40 items) and math (29 items) scales were calculated. Analogous to classical test theory, OTST predicts that the test-retest γ reliability should equal the product of the γ of item total and true order for each of the scales. These results are summarized in Table 3. As can be seen, the estimated τ correlations are too large to generate the test-retest γ correlation. Indeed, six of the estimated γs of total with true order were greater than 1.0.

## Stage 3: IRT Simulation Study

### Method

As was noted above, the method of data generation for the classical true score model corresponds to a normal ogive IRT model. In order to directly investigate the relationship between IRT models and OTST, this stage of the research was designed to determine the conditions under which the IRT and OTST models agreed and the conditions under which they disagreed. Data were generated according to the three-parameter logistic model (Birnbaum, 1968). This model gives the probability of a correct response as:

$$P_j(\theta) = c_j + \frac{(1 - c_j)}{1 + \exp[-1.7a_j(\theta - b_j)]} \tag{14}$$

where:

Table 3
Test-Retest τ Correlations and Correction for Ties
($\gamma_{XY} = \tau_{XY}/\tau_{YY}$) for Verbal and Math Tests

| Test and | Fall | | Spring | | | |
|---|---|---|---|---|---|---|
| Weights | $\tau_{FY}$ | $\gamma_{FY}$ | $\tau_{FY}$ | $\gamma_{SY}$ | $\gamma_{FY}\gamma_{SY}$ | $\gamma_{FS}$ |
| Verbal | .965* | | .964* | | | |
| Weights | | | | | | |
| Unit | .974 | 1.01 | .961 | .997 | 1.01 | .744 |
| Optimal | .980 | 1.02 | .970 | 1.01 | 1.03 | .744 |
| Math | .957* | | .958* | | | |
| Weights | | | | | | |
| Unit | .962 | 1.01 | .959 | 1.00 | 1.01 | .694 |
| Optimal | .969 | 1.01 | .963 | 1.01 | 1.02 | .694 |

*Proportion of paired comparisons not tied on that test.

θ is the ability that underlies the responses to the test items,
$c_j$ is the guessing parameter, and corresponds to the probability that an examinee of very low ability will answer the item correctly,
$b_j$ is the difficulty parameter, and
$a_j$ is the discrimination parameter.

The simulation design used a wide range of values for the item parameters in an effort to investigate the OTST model in a wide range of conditions. The factors manipulated were:

1. Mean of the discrimination parameters, AMEAN. Two levels were used: Low ($\mu_a = .8$) and High ($\mu_a = 1.3$).
2. Variability of the discrimination parameters, ASTD. The two levels were None ($\sigma_a = 0.0$) and Some ($\sigma_a = .3$).
3. Mean of the difficulty parameters, BMEAN. Three levels were used: Easy test ($\mu_b = -.5$), Medium Difficulty ($\mu_b = 0.0$), or High Difficulty ($\mu_b = .5$).
4. Variability of the difficulty parameters, BSTD. The two levels were Low ($\sigma_b = .7$) and High ($\sigma_b = 1.3$).
5. Value of the common guessing parameter, COMC. Two levels were used: No Guessing ($c_j = 0.0$) and Chance Level Guessing ($c_j = .2$).

These five factors were fully crossed to yield 48 cells. Ten replications were run for each cell, yielding a total of 480 datasets. The number of simulated examinees (simulees) was held constant (200), as was the number of items (40). As preliminary work had indicated, increasing either of these values yielded no appreciable change in the performance of the OTST model.

Item responses were generated as follows: for each dataset, 200 abilities $\theta_i$ were generated from a N(0,1) distribution. Item parameters were then generated for each of the 40 items. The difficulty parameters were generated from a normal distribution with mean $\mu_b$ and standard deviation $\sigma_b$. The discrimination parameters were generated from a normal distribution with mean $\mu_a$ and standard deviation $\sigma_a$. The guessing parameter was constant for all items, with the value $c_j$. For each simulee, the probability of correctly answering each item $P_j(\theta_i)$ was then calculated using Equation 14. A uniform [0,1] number $U$ was generated for each item. If $U$ was less than or equal to $P_j(\theta_i)$, the simulee was deemed to have correctly answered the item; otherwise, the response was scored as incorrect.

In analyzing the datasets, θ was used to define the true order. For each dataset, four outcome measures were obtained: (1) the average partialed γ, AVGAM; (2) the average $Q_j$ ratio, AVRAT; (3) the average discrepancy between the estimated item-true order τ and actual item-true order τ, DITEMT; and (4) the discrepancy between the estimated total score-true order τ and the actual item total-true order τ, DTOTAL.

**Table 4**
Descriptive Statistics for Dependent Variables

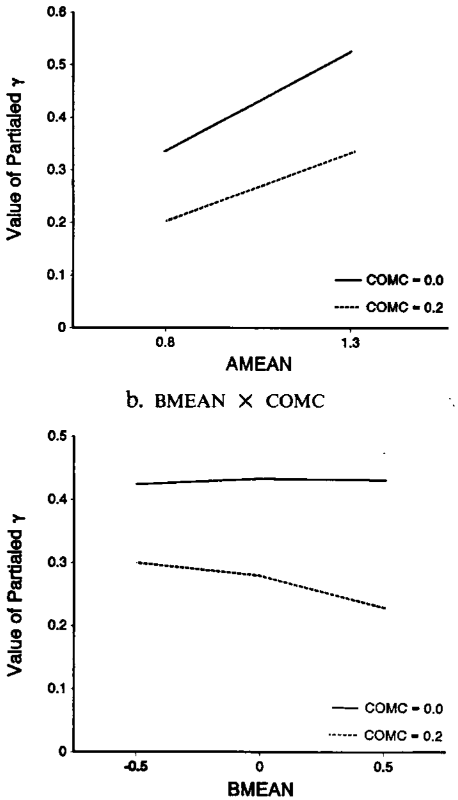
| Statistic | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|
| AVGAM | .34955 | .12507 | .10041 | .63820 |
| AVRAT | .87745 | .07211 | .65828 | .97666 |
| DITEMT | .03974 | .00618 | .02046 | .05872 |
| DTOTAL | .15585 | .04854 | .04720 | .26128 |

## Results

Table 4 gives the dependent variable means, standard deviations, minima, and maxima over all 480 datasets. Considering the OTST assumptions, Table 4 shows that the average partial γ was positive in all cases—the minimum was .10 and the average was substantial, .35. The average $Q_j$ ratio was always below 1.0, averaging .88, although the maximum was as high as .98. The average estimated item-true order τ was always an overestimate, with an overall bias of .04 and an observed minimum of .02. The total score-true order τ was also always an overestimate, with an average discrepancy of .16 and an observed low of .05. Thus, when item responses were simulated with the three-parameter logistic model, the OTST assumptions did not hold. Furthermore, they led to appreciable overestimates of item-true order and total score-true order correlations.

Table 5 shows that, in addition to the very large effect for "constant" (i.e., bias in the overall mean), all main effects had highly significant influences on the average partial γ. The influences of AMEAN and COMC were the largest. Several two- and three-way interactions had small to moderate effects. Means for the largest of these, AMEAN × COMC and BMEAN × COMC, are plotted in Figure 2. Both of these interactions indicate that for conditions for which guessing had more impact on the data (highly discriminating items or a difficult test), the presence of guessing (COMC = .2) had a larger effect of decreasing the average partialed γ. The results generally indicate that discrepancies will be smallest with the least consistent data—that is, with lower discriminations and/or in the presence of guessing, and with variable difficulties. It was under conditions with the most consistency of response that the largest discrepancies occurred.

**Figure 2**
Means for Significant Interactions for AVGAM



a. AMEAN × COMC

b. BMEAN × COMC

Table 5 also reports the effects of AVRAT, the proportional distribution of tied pairs. Here AMEAN and COMC had the largest effects, although BSTD was also appreciable. Figure 3 gives the means for the largest two-way interaction, AMEAN × COMC. In tests composed of

**Table 5**
Results of the Analysis of Variance for AVGAM, AVRAT, DITEMT, and DTOTAL

| | | Sum of | | |
|---|---|---|---|---|
| Source | *df* | Squares | *F* | *P* |
| AVGAM (R² = .915) | | | | |
| Constant | 1 | 58.649 | | |
| Main Effects | | | | |
| AMEAN | 1 | 3.068 | 2090.51 | .0001 |
| ASTD | 1 | .056 | 38.18 | .0001 |
| BMEAN | 2 | .093 | 31.52 | .0001 |
| BSTD | 1 | .228 | 155.02 | .0001 |
| COMC | 1 | 3.083 | 2100.69 | .0001 |
| Two-Way Interactions | | | | |
| AMEAN × BSTD | 1 | .036 | 24.24 | .0001 |
| AMEAN × COMC | 1 | .100 | 68.41 | .0001 |
| BMEAN × COMC | 2 | .125 | 42.68 | .0001 |
| All Other 2-Way | 10 | .013 | .82 | n.s. |
| Three-Way Interactions | | | | |
| AMEAN × BMEAN × COMC | 2 | .022 | 7.45 | .0007 |
| AMEAN × BSTD × COMC | 1 | .012 | 8.45 | .0038 |
| All Other 3-Way | 13 | .011 | .52 | n.s. |
| Other Interactions | | | | |
| All 4-Way and 5-Way | 11 | .012 | .74 | n.s. |
| Error | 432 | .634 | | |
| Total | 479 | 7.492 | | |
| AVRAT (R² = .915) | | | | |
| Bias in constant | 1 | 7.2089 | | |
| Main Effects | | | | |
| AMEAN | 1 | .8981 | 1826.42 | .0001 |
| ASTD | 1 | .0039 | 7.95 | .0050 |
| BMEAN | 2 | .0023 | 2.34 | .0976 |
| BSTD | 1 | .0646 | 131.38 | .0001 |
| COMC | 1 | 1.1297 | 2297.27 | .0001 |
| Two-Way Interactions | | | | |
| AMEAN × BSTD | 1 | .0046 | 9.50 | .0022 |
| AMEAN × COMC | 1 | .1413 | 287.35 | .0001 |
| BMEAN × COMC | 2 | .0037 | 3.80 | .0231 |
| BSTD × COMC | 1 | .0078 | 15.87 | .0001 |
| All Other 2-Way | 9 | .0032 | .72 | n.s. |
| Three-Way Interactions | | | | |
| AMEAN × BSTD × COMC | 1 | .0019 | 4.07 | .0444 |
| BMEAN × BSTD × COMC | 2 | .0036 | 3.71 | .0253 |
| All Other 3-Way | 13 | .0060 | .94 | n.s. |
| Other Interactions | | | | |
| All 4-Way and 5-Way | 11 | .0074 | 1.37 | n.s. |
| Error | 432 | .2124 | | |
| Total | 479 | 2.4909 | | |
| DITEMT (R² = .697) | | | | |
| Bias in Constant | 1 | .75805 | | |
| AMEAN | 1 | .00019 | 14.85 | .0001 |
| ASTD | 1 | .00019 | 14.99 | .0001 |
| BMEAN | 2 | .00186 | 72.55 | .0001 |

*continued on the next page*

**Table 5, continued**
Results of the Analysis of Variance for AVGAM, AVRAT, DITEMT, and DTOTAL

| Source | *df* | Sum of Squares | *F* | *P* |
|---|---|---|---|---|
| BSTD | 1 | .00338 | 263.02 | .0001 |
| COMC | 1 | .00235 | 182.64 | .0001 |
| Two-Way Interactions | | | | |
| AMEAN × ASTD | 1 | .00027 | 21.38 | .0001 |
| AMEAN × BMEAN | 2 | .00030 | 11.72 | .0001 |
| AMEAN × COMC | 1 | .00096 | 74.35 | .0001 |
| BMEAN × BSTD | 2 | .00034 | 13.39 | .0001 |
| BMEAN × COMC | 2 | .00190 | 73.94 | .0001 |
| All Other 2-Way | 6 | .00008 | 1.04 | n.s. |
| Three-Way Interactions | | | | |
| AMEAN × BMEAN × COMC | 2 | .00033 | 12.71 | .0001 |
| BMEAN × BSTD × COMC | 2 | .00023 | 9.01 | .0001 |
| All Other 3-Way | 12 | .00018 | 1.17 | n.s. |
| Other Interactions | | | | |
| All 4-Way and 5-Way | 11 | .00019 | 1.34 | n.s. |
| Error | 432 | .00555 | | |
| Total | 479 | .01832 | | |
| DTOTAL ($R^2$ = .830) | | | | |
| Bias in Constant | 1 | 11.6588 | | |
| AMEAN | 1 | .4238 | 953.35 | .0001 |
| ASTD | 1 | .0030 | 6.78 | .0095 |
| BMEAN | 2 | .0734 | 82.57 | .0001 |
| BSTD | 1 | .0246 | 55.44 | .0001 |
| COMC | 1 | .2535 | 570.40 | .0001 |
| Two-Way Interactions | | | | |
| AMEAN × BMEAN | 2 | .0103 | 11.58 | .0001 |
| AMEAN × COMC | 1 | .0233 | 52.49 | .0001 |
| ASTD × BMEAN | 2 | .0036 | 4.03 | .0185 |
| BMEAN × BSTD | 2 | .0050 | 5.59 | .0040 |
| BMEAN × COMC | 2 | .0895 | 100.60 | .0001 |
| All Other 2-Way | 5 | .0005 | .23 | n.s. |
| Three-Way Interactions | | | | |
| AMEAN × ASTD × COMC | 1 | .0017 | 3.90 | .0491 |
| AMEAN × BMEAN × COMC | 2 | .0165 | 18.58 | .0001 |
| All Other 3-Way | 13 | .0058 | 1.00 | n.s. |
| Other Interactions | | | | |
| All 4-Way and 5-Way | 11 | .0020 | .41 | n.s. |
| Error | 432 | .1920 | | |
| Total | 479 | 1.1288 | | |

more highly discriminating items, for which guessing had more impact on the data, the absence of guessing (COMC = 0.0) led to larger deviations from the assumption of proportionality of ties. Again, all the effects were overshadowed by the fact that the overall mean discrepancy was large. The direction of the effects was again that the discrepancies were smaller for less consistent data; under the combination of guessing, low discrimination, and variable difficulty, the mean ratio was .96; but it was as low as .77 with no guessing, high discrimination, and small variability in difficulty.

Table 5 shows that there were a variety of influences on the accuracy of estimating $\tau_{jY}$, analyzing the discrepancy $(\hat{\tau}_{jY} - \tau_{jY})$, DITEMT.

**Figure 3**
Cell Means for the AMEAN × COMC Interaction for AVRAT

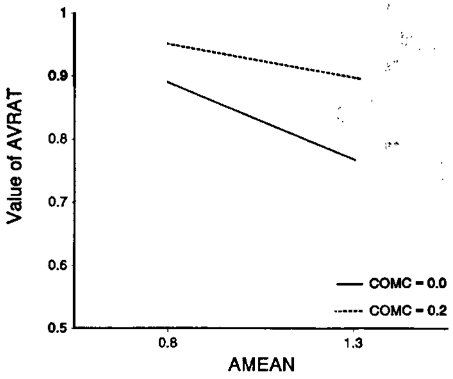
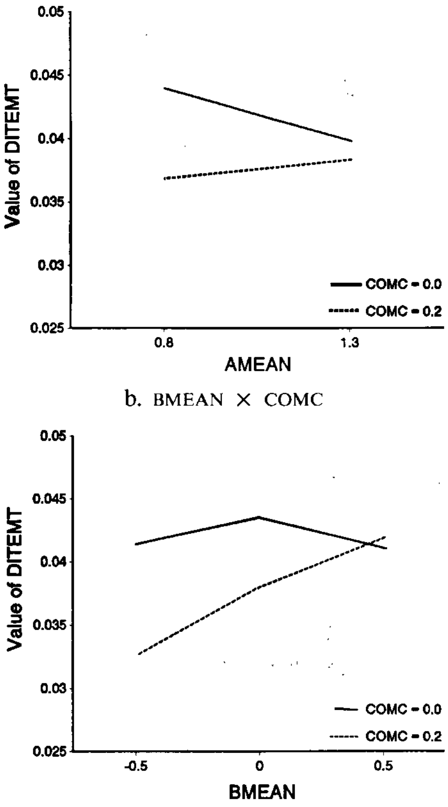


**Figure 4**
Cell Means for Significant Interactions for DITEMT

a. AMEAN × COMC



b. BMEAN × COMC

Here, the consistency of the data had a relatively small, although significant, effect on the bias. The larger main effects were found for BMEAN, BSTD, and COMC. Several two- and three-way interactions had moderately large effects. Means for the largest two-way interactions, AMEAN × COMC and BMEAN × COMC, are plotted in Figures 4a and 4b, respectively. Both of these interactions indicate that conditions for which guessing had more impact on the data (highly discriminating items or a difficult test), the value of COMC paradoxically had a *smaller* effect on the bias in estimating $\tau_{jY}$. The smallest bias occurred when the presence of guessing was combined either with high discrimination and low difficulty, or with low discrimination and variable difficulty. Under these circumstances, the average item-true order $\tau$ was overestimated by approximately .03 relative to the true $\tau_{jY}$ of approximately .20.

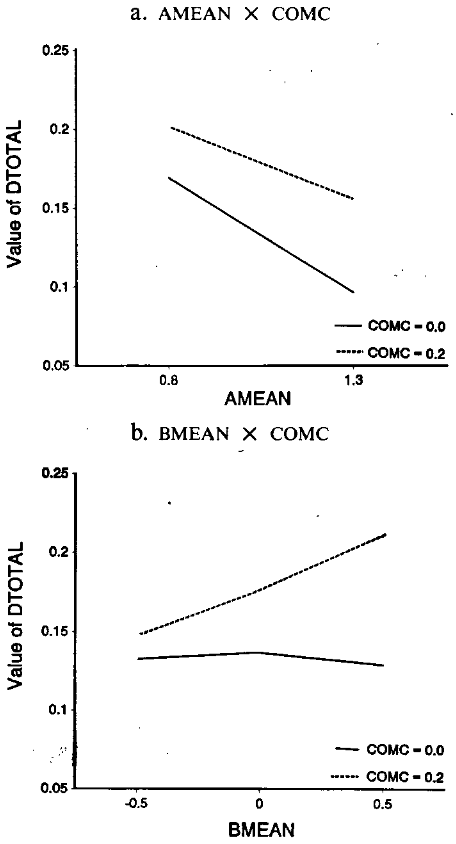
In addition to revealing a substantial tendency to overestimate the total score-true order $\tau$, Table 5 indicates that all the design factors influenced the size of this discrepancy. AMEAN and COMC had the largest effects, and an appreciable interaction, but there were several other significant main effects and interactions as well. Figure 5a shows the means for the AMEAN × COMC interaction, and Figure 5b shows the BMEAN × COMC means. Here, the discrepancies tended to be largest with the least consistent data: with lower discriminations and guessing. The latter interacted appreciably with BMEAN—the largest cell mean occurred with low discrimination, guessing, and high difficulty. These are contrary to the effects on assumption violation, where the largest effects occurred with high discrimination and no guessing.

In general, however, the results demonstrate the lack of agreement of Cliff's OTST model with the simulated data. The assumptions were violated in all cells, and all cells showed appreciable overestimation of correlations with the true order.

## DISCUSSION AND CONCLUSIONS

Several conclusions may be drawn from these analyses. First, the assumptions of classical test theory and OTST are incompatible. It was

**Figure 5**
Cell Means for Significant Interactions for DTOTAL


a. AMEAN × COMC

b. BMEAN × COMC

originally thought that the local uncorrelatedness assumption of OTST was a weaker version of local independence. This belief was based in part on Kendall's (1970, pp. 117-122) demonstration that the computational formula for partialed $\tau$ is identical to that of the partialed Pearson correlation.

Based on the results obtained, this belief was reexamined and analytically demonstrated to be false. The common factor model implies that the partialed interitem tetrachoric correlation should be 0.0. Kendall (1970, p. 126) has shown that for bivariate normal variables

$$E(\tau) = (2/\pi)\sin^{-1}(\rho) \quad , \tag{15}$$

where $\rho$ is the product-moment correlation. By substituting into the formula for the partialed Pearson correlation, it was found that the partialed $\tau$ is greater than 0 for the continuous variables. By definition, $\gamma$ is greater than or equal to $\tau$; hence, the expected partialed $\gamma$ is greater than 0. By the same argument, when OTST is true, classical test theory must be false. Thus, the two theories make different assumptions which cannot be simultaneously true. This is because local ordinal uncorrelatedness, defined in Equation 5, is not a weaker form of local independence.

For OTST to be consistent with classical test theory, the local ordinal pair relations shown in Table 1 would have to be redefined in terms of pairs of examinees who are *tied* on the true score *Y*, which is counter to the assumption that there are no ties on the true order. In this case, the subscripts + and – have no meaning, and the table collapses to a three-way classification: pairs ranked in the same direction, pairs ranked in the opposite direction, and pairs tied on the variables. Also, the assumptions of proportionality of ties (Equations 6 and 7) have no meaning in this case. Note that when *Y* is a complete ranking (i.e., there are no ties on the true order), all such tables of local ordinal relations are empty. In summary, a redefinition of the assumption in Equation 5 would be necessary to be compatible with local independence and would require the complete reformulation of OTST, which is beyond the scope of this paper.

The relationship between OTST and IRT models is less clear. The OTST assumption that the partialed $\gamma$ correlation was 0 was uniformly violated. The reason for this is the same as that of the classical test theory case; there is a fundamental difference between local independence and the pairwise local ordinal uncorrelatedness assumed by OTST.

Table 6 shows that the pattern of main effects on the ratio $Q_j$ mirrored those of the bias in the item sum-true order $\tau$, DTOTAL. The pattern of effects on the item-true order $\tau$ were different, as were those of the average partialed $\gamma$. Noteworthy is the fact that for AVGAM, $Q_j$, and DTOTAL, decreasing item discrimination and guessing were both associated with better estimates. Yet, both

Table 6
Means for Main Effects of Design Factors

| Statistic and Level | AVGAM | AVRAT | DITEMT | DTOTAL |
|---|---|---|---|---|
| AMEAN | | | | |
| .8 | .26960 | .92071 | .040372 | .18556 |
| 1.3 | .42951 | .83419 | .039111 | .12613 |
| ASTD | | | | |
| .0 | .36036 | .87460 | .040375 | .15334 |
| .3 | .33875 | .88030 | .039108 | .15835 |
| BMEAN | | | | |
| -.5 | .36205 | .87507 | .036984 | .14017 |
| .0 | .35642 | .87692 | .040768 | .15697 |
| .5 | .33019 | .88036 | .041472 | .17040 |
| BSTD | | | | |
| .7 | .37133 | .86585 | .042395 | .14868 |
| 1.3 | .32778 | .88905 | .037088 | .16301 |
| COMC | | | | |
| .0 | .42970 | .82894 | .041953 | .13286 |
| .2 | .26941 | .92596 | .037531 | .17883 |

of these factors indicate a loss of information. It appears that a test constructed to satisfy the OTST assumptions would be a poor test. Because the item parameters were selected to be more extreme than would be likely to occur empirically, it appears unlikely that such a test exists. Nor is it likely that such a test would lead to reliable measurement.

The bias in partial $\tau$ mentioned above for the normal case is part of a more general phenomenon. In partial $\tau$ it is not the *value* of the control variable that is held constant, as in the ordinary partial correlation, but rather it is the *relation of pairs* on the control variable that is held constant. A conditional $\tau$ could be defined as the average $\tau$ between items for constant values of the true score. This conditional $\tau$ would be 0 whenever the partial covariance between the items is 0, because of the relation between $\tau$ and covariance in the dichotomous case. Another interpretation would be as the average for pairs that are tied on the true score. Partial $\tau$, on the other hand, "holds constant" the *relation of pairs* on the control variable. Somers (1976) provided some discussion of this issue, although failure to appreciate the distinction remains fairly widespread (e.g., Hettmansperger, 1984).

The item-true order estimation functioned fairly well. Figure 6 gives a plot of the estimated and actual item-true order $\tau$ for the means of each of the 48 cells of the design. All the points fall above the line $Y = X$, confirming that the estimate is uniformly too large. The linear regression is also plotted. As can be seen, the plot is highly linear, indicating that much of the error of estimation is systematic and may be correctable. The correlation was .994. This is not an artifact of aggregating the data; correlations within the 480 individual tests ranged from .851 to .998, with a mean of .975.

The results in Figure 6 were obtained in spite of the persistent violation of both the assumptions of local uncorrelatedness and proportionality of ties. Thus, although the assumptions underlying the derivation of $\hat{\tau}_{jY}$ were not valid in the cases examined, the estimate appears to have some promise nonetheless. Further, the relationship between the estimated and actual item-true order $\gamma$ and $\tau$ were both highly linear for every case examined. Therefore, this method of estimation should be examined further in the future.

These results indicate that the assumptions in Equations 5 through 7 are sufficient but not necessary for the result in Equation 8 to hold.

**Figure 6**
Cell Means for Estimated and Actual Item-True Order τ Correlations

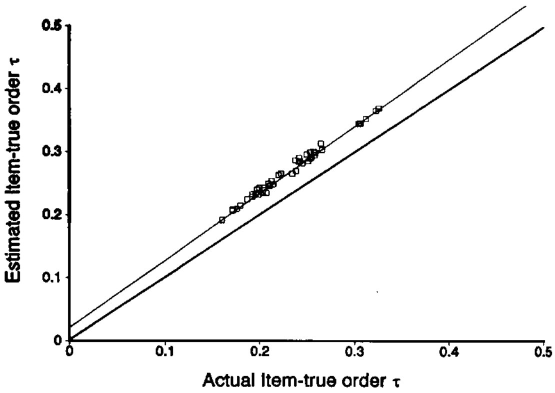
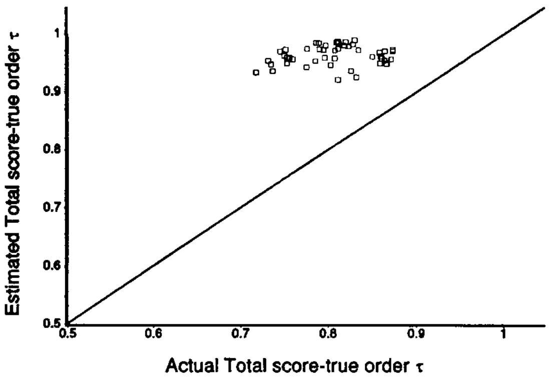


**Figure 7**
Cell Means for Estimated and Actual Total Score-True Order τ Correlations



It would be useful to examine other sets of assumptions that could lead to Equation 8. However, the nature of γ makes this task difficult because no restriction is placed on the number or allocation of tied cases. Thus far, attempts to determine alternative sets of assumptions have not yielded any revealing results.

The estimate of the τ correlation of the true order with the total score appeared to be a consistent overestimate of the actual value, but there was little relation between the estimated and true values. Figure 7 shows a plot of the estimated and actual total score-true order τ for the means of each of the 48 cells of the design. All the points fall above the line $Y = X$, indicating that the estimate was consistently too large. As can be seen, there appears to be little systematic relationship between the estimated and actual values ($\tau = .091$). Further, the estimated τ shows much less variability than does the actual τ. The Pearson correlation was .152. In contrast, the Pearson correlation of the actual total score-true order τ with estimated item-true order τ was .883. Thus, the procedure did not accurately estimate the total score-true order τ.

There are two possible sources for the observed bias. The first issue is the estimation of the item-true order relationship, $\tau_{jY}$. These estimates were derived from a weighted version of the local ordinal uncorrelatedness assumption, by centroid factoring of the matrix of γ correlations. The method of determining these weights is basically arbitrary and was selected for its intuitive appeal. It may be that this procedure introduced bias into the estimation of $\tau_{XY}$.

A second explanation is that the value of $\tau_{XY}$ is derived under an assumption of normality, based on the central limit theorem. Cliff (1986) and Charlin (1987) have found the estimate to be relatively accurate, despite radical departures from normality. However, in the context of their work, the indicators had very few ties; the dichotomous items used here had a large number of ties. Regardless of the reason, however, the total score-true order estimate functioned poorly.

### Future Directions for Ordinal Test Theory

As was noted above, the goal of deriving an ordinal theory of test scores is an old one in the field of psychometrics. The results of the present study indicate that OTST is no longer a viable candidate, because the assumptions under which the theory was derived did not hold for empirical or generated data.

What alternatives are available for using an ordinal test theory? Guttman's (1950) scalogram analysis suffers from being deterministic and not acknowledging the stochastic nature of examinees' responses to test items. Lovinger's (1947) and Mokken's (1971) ordinal test theories do not provide much specific information at the item level. Schulman and Haden's (1975) ordinal test theory treats the ranks as meaningful numbers. This has the undesirable effect that their results are entirely dependent on the specific sample under study, because the addition of a single case can alter every other examinee's true rank. Thus, each of the ordinal formulations above suffers from some shortcoming.

Two bodies of research seem most promising for ordinal test theory. A recent ordinal item sampling formulation by Cliff and Donoghue (in press) is an ordinal test theory in much the same spirit as OTST. They conceptualize items as being sampled from a larger universe of possible items and define the "universe order" $Y_{ih}$ of examinees $i$ and $h$ as the total of their orders (dominances) on the $P$ items in the universe:

$$Y_{ih} = \sum_{J=1}^{P} d_{ihJ} \tag{16}$$

where the item dominances are defined as

$$d_{ihJ} = \left\{ \begin{array}{l} 1,\ X_{iJ} > X_{hJ} \\ 0,\ X_{iJ} = X_{hJ} \\ -1,\ X_{iJ} < X_{hJ} \end{array} \right\} . \tag{17}$$

Using sample quantities, they estimate the ordinal correlations of the items with the universe order, the ordinal correlation of the total score with the universe order, and the probability that a difference in observed total score reflects a correct ordering in the universe order. A small simulation in Cliff and Donoghue (in press) and a more extensive evaluation in Donoghue (1990) indicate that this approach functions well for both empirical and simulated data. Thus, the ordinal item sampling formulation appears to offer promise as an ordinal test theory.

The second body of ordinal work is nonparametric IRT—models in which the shape of the IRF is not restricted to a particular form. The work of Mokken (1971; Mokken & Lewis, 1982) may be considered to fall into this category.

Holland (1981) has examined the minimal conditions under which IRT makes meaningful predictions about data, and his "Dutch identity" (Holland, 1990) yields some interesting conjectures as to the number of parameters that may be estimated from data. The "ordinal IRT" work of Stout and Junker (Stout, 1988, 1990; Junker, 1989, 1990) also uses nonparametric IRT, but weakens the assumption of local independence to what they term "essential independence." Finally, there have been applications of kernel smoothing (e.g., Ramsay, 1990) to obtain empirical estimates of the IRFs. The nonparametric IRT work also appears to hold promise for the future of ordinal test theories.

## Conclusions

Based on the results obtained here, OTST is not recommended for use in practical testing situations. The assumptions under which the theory was derived were violated in both empirical and simulated tests. Further, the estimate of the total score-true order $\tau$ correlation functioned poorly.

On the other hand, the estimates of item-true order $\tau$ correlations functioned well, and the deviations observed were systematic. This method of estimating item-true order relations should therefore be studied further. The factor analysis of the $\gamma$ matrix also appeared to function as an indication of the underlying dimensionality, but the results obtained are only suggestive. Further study, especially of the method's ability to detect multidimensionality, is needed.

## References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores* (pp. 395–479). Reading MA: Addison-Wesley.

Charlin, V. (1987). *Implementation and inferential issues of a multiple regression system with ordinal variables.* Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Cliff, N. (1977). A theory of consistency of ordering generalizable to tailored testing. *Psychometrika, 42,* 375–399.

Cliff, N. (1979). Test theory without true scores? *Psychometrika, 44,* 373–393.

Cliff, N. (1986). Ordinal analogues to multiple regression. *Proceedings of the Social Science Section of the American Statistical Association*, 320-324.

Cliff, N. (1989). Ordinal consistency and ordinal true scores. *Psychometrika, 54,* 75-91.

Cliff, N. & Donoghue, J. R. (in press). Ordinal test theory from item sampling assumptions. *Psychometrika.*

Donoghue, J. R., Jr. (1990). *Properties of ability estimates derived from logistic item response test theory and ordinal item sampling test theory.* Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Goodman, L., & Kruskal, W. B. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732-764.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer (Ed.), *Measurement and prediction* (pp. 60-90). Princeton NJ: Princeton University Press.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks.* New York: Wiley.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79-92.

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika, 55,* 5-18.

Junker, B. W. (1989, June). *Conditional association, essential independence and local independence.* Paper presented at the annual meeting of the Psychometric Society, Los Angeles CA.

Junker, B. W. (1990, June). *Essential independence and structural robustness in item response theory.* Paper presented at the annual meeting of the Psychometric Society, Princeton NJ.

Kendall, M. G. (1970). *Rank correlation methods* (4th ed.). London: Charles Griffin and Co.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs, 61* (4, Whole No. 285).

Lord, F. M. (1952). A theory of test scores. *Psychological Monographs,* No. 7.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* Hawthorne NY: Mouton and Co.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417-430.

Ramsay, J. O. (1990, June). *Kernel smoothing approaches to nonparametric item characteristic curve estimation.* Paper presented at the annual meeting of the Psychometric Society, Princeton NJ.

Schulman, R. S. (1976). Correlation and prediction in ordinal test theory. *Psychometrika, 41,* 329-340.

Schulman, R. S., & Haden, R. L. (1975). A test theory for ordinal measurements. *Psychometrika, 40,* 455-472.

Somers, R. H. (1976). A caution regarding partial rank correlation based on the product-moment model. *Social Forces, 54,* 694-700.

Stout, W. (1988). *A nonparametric multidimensional IRT approach with applications to ability estimation and test bias* (Report No. ONR-88-1). Champaign: University of Illinois at Urbana-Champaign, Department of Statistics.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293-325.

## Acknowledgments

*The first author was supported by an All-University Fellowship from the University of Southern California. An earlier version of this paper was presented at the meeting of the Psychometric Society in Los Angeles, June 1988. This paper has benefited from the comments of two anonymous reviewers. The authors thank them for their detailed suggestions and clarifications on an earlier version of this paper.*

## Author's Address

Send requests for reprints or further information to John R. Donoghue, Mail Stop 02-T, Educational Testing Service, Rosedale Road, Princeton NJ 08541, U.S.A.