An Examination of the Characteristics of Unidimensional IRT Parameter Estimates Derived From Two-Dimensional Data

Timothy N. Ansley and Robert A. Forsyth
The University of Iowa

The purpose of this investigation was to study the nature of the item and ability estimates obtained when the modified three-parameter logistic model is used with two-dimensional data. To examine the effects of two-dimensional data on unidimensional parameter estimates, the relative potency of the two dimensions was systematically varied by changing the correlations between the two ability dimensions. Data sets based on correlations of .0, .3, .6, .9, and .95 were generated for each of four combinations of sample size and test length. Also, for each of these four combinations, five unidimensional data sets were simulated for comparison purposes. Relative to the nature of the unidimensional estimates, it was found that the $\hat{a}$ value seemed best considered as the average of the true a values. The $\hat{b}$ value seemed best thought of as an overestimate of the true b_i values. The $\hat{\theta}$ value seemed best considered as the average of the true ability parameters. Although there was a consistent trend for these relationships to strengthen as the ability dimensions became more highly correlated, there was always a substantial disparity between the magnitudes of these values and of those derived from the unidimensional data. Sample size and test length had very little effect on these relationships.

Research related to item response theory (IRT) has dominated the psychometric literature in recent years. This is not surprising since this theory has the potential to resolve many problems frequently encountered in psychological and educational measurement (Lord, 1980). However, the mathematical models on which this theory is founded are based on some very strong assumptions. In particular, IRT models most commonly used assume that the response data are unidimensional in the reference population.

The importance of the unidimensionality assumption has been stressed by many authors (see e.g., Hambleton & Murray, 1983; Traub, 1983). Several researchers have hypothesized that failure to satisfy the assumption of unidimensionality was a major reason IRT models did not adequately fit their data. For example, Loyd and Hoover (1980) reported that multidimensional data might have contributed to the lack of fit of the Rasch model in a vertical equating setting. In comparing the fit of the one- and three-parameter logistic models to actual standardized test data, Hutten (1980) found that the potency of the major dimension (as assessed by the ratio of the first two eigenvalues of the matrix of inter-item tetrachoric correlations) was significantly related to the degree of fit.

Despite the discussions by these authors and others (e.g., Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978; Lord, 1980; Rentz & Rentz, 1979) concerning the importance of the unidimensionality assumption, there have been relatively few

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 1, March 1985, pp. 37–48

0146-6216/85/010037-12$1.85

studies directly assessing the effects of violating this assumption. Two such studies were undertaken by Drasgow and Parsons (1983) and by Reckase (1979). Reckase (1979) investigated the effects of using the one- and three-parameter logistic models with multidimensional data. He used real and simulated sets of test data. The simulated data were generated to fit a linear factor analysis model. For all data sets, the degree of multidimensionality was determined by the number and relative strength of the common factors.

Reckase's (1979) study did not offer any clear insights as to the nature of the item parameter estimates (derived using LOGIST, Wood, Wingersky, & Lord, 1976). However, with regard to ability parameters, he reported that for data sets with two or more equally potent factors, the one-parameter ability estimates were nearly equally correlated with the factor scores derived from each factor. Reckase concluded that for tests with several equally potent dimensions, the one-parameter ability estimates were best considered as the sum or average of the abilities required for each dimension. For data sets with a dominant first factor (accounting for 10% to 40% of the total variation), the one-parameter ability estimates were highly correlated with the scores for that factor. When the three-parameter model was applied to data sets with two or more equally potent factors, the ability estimates were highly correlated with the factor scores for just one of the common factors. For data sets with a dominant first factor, the three-parameter ability estimates were highly correlated with the scores for that factor.

In the Drasgow and Parsons (1983) study, data were generated to fit a hierarchical factor analysis model (Schmid & Leiman, 1957). For all data sets, there were five first-order common factors and a single second-order general factor. The potency of the general factor was controlled by manipulating the correlations between the first-order common factors. The first-order common factors varied in potency according to the number of items loading on each. Five data sets were simulated ranging from strictly unidimensional (first-order common factors perfectly correlated) to clearly multidimensional (first-order common factors correlated between .02 and .14). True discrimination and difficulty parameters were derived using the relationships between the two-parameter normal ogive model and the factor analysis model (Lord & Novick, 1968). True and estimated item parameters (estimates derived using LOGIST) were compared using root mean square differences. The ability estimates were correlated with the factor scores for both the general factor and the first-order factors.

For the strictly unidimensional data set and the two multidimensional data sets with first-order common factors most highly correlated (between .46 and .90), item parameter estimates were closely related to the true item parameters derived from the factor loadings for the general factor. Similarly, the ability estimates for these data sets were most highly correlated with the factor scores derived from the general factor. In the data set with first-order common factors correlated between .25 and .39, the discrimination and difficulty estimates seemed to represent a combination of the corresponding parameters derived from the factor loadings for the general factor and those derived from the factor loadings for the most potent first-order common factor.

The same was true for the ability estimates for this data set relative to the factor scores for the general factor and those for the most potent first-order common factor. For the data set with first-order common factors correlated between .02 and .14, the discrimination and difficulty estimates were most highly related to the parameters derived from the factor loadings associated with the most potent first-order common factor. Similarly, the ability estimates for this data set were most highly correlated with the factor scores for the most potent first-order common factor. Drasgow and Parsons concluded that as the potency of the second-order general factor decreased, the item and ability estimates produced by LOGIST were more closely related to the parameters associated with the most potent first-order common factor than to those associated with the general factor. They recommended that LOGIST should not be used with data sets having

oblique common factors that are correlated in the range from 0 to .4.

Reckase (1979) and Drasgow and Parsons (1983) used the factor analysis model to assess dimensionality as well as to generate simulated data. Since the relationship between the factor analysis model and the logistic model is not precisely defined, generating data to fit a factor analysis model might not yield a completely clear picture of the effects of using a unidimensional logistic model with multidimensional data. For studies of this type, it may be more reasonable to generate data to fit a multidimensional extension of a unidimensional IRT model.

The primary purpose of this study was to investigate the effects of using a unidimensional IRT model with multidimensional data. Specifically, this study attempted to provide some understanding of the nature of the item and ability parameter estimates derived from applying the unidimensional three-parameter logistic model to two-dimensional data. These data were simulated to fit a multidimensional extension of a unidimensional IRT model.

Method

Selection of Model

The simulated data had to represent realistic response data. Standardized achievement test data were used as a frame of reference for evaluating the representativeness of the simulated data sets.

The data generation procedure was similar to that used by Doody-Bogan and Yen (1983). Data were generated using a multidimensional IRT model. The first and perhaps most critical issue confronted was the choice of a multidimensional model. Several such models have been proposed (Reckase & McKinley, 1985). In selecting a model, a fundamental choice between extensions of the one- and three-parameter logistic models was required. It seemed clear that the three-parameter model would provide better model-to-data fit given that the data were to represent standardized achievement test data. Thus, a multidimensional version of the three-parameter logistic model was chosen for data simulation. (To avoid estimation problems, the guessing parameter was fixed at .2 throughout this investigation. Thus, the model actually used is referred to as a modified three-parameter logistic model.)

There have been three such models proposed. Hattie (1981) proposed the following multidimensional model:

$$P_{ij}(\theta_{ih}) = c_j + \frac{1 - c_j}{1 + \exp[-1.7 \sum_{h=1}^{n} a_{jh}(\theta_{ih} - b_j)]}, \quad (1)$$

where θ_{ih} is the ability parameter for person i for dimension h,
a_{jh} is the discrimination parameter for item j for dimension h,
b_j is the difficulty parameter for item j, and
c_j is the guessing parameter for item j.

Doody-Bogan and Yen (1983) modified this model as follows:

$$P_{ij}(\theta_{ih}) = c_j + \frac{1 - c_j}{1 + \exp[-1.7 \sum_{h=1}^{n} a_{jh}(\theta_{ih} - b_{jh})]}, \quad (2)$$

where b_{jh} is the difficulty parameter for item j for dimension h, and all other parameters are as defined above. These two models differ only in the number of difficulty parameters. The Hattie (1981) model specifies a single difficulty parameter for each item, whereas the Doody-Bogan and Yen model requires a difficulty parameter for each dimension.

Another model was proposed by Sympson (1978):

$$P_{ij}(\theta_{ih}) = c_j + \frac{1 - c_j}{\prod_{h=1}^{n} (1 + \exp[-1.7 a_{jh}\{\theta_{ih} - b_{jh}\}])}, \quad (3)$$

where all parameters are as defined above. Equation 3 can be distinguished from Equations 1 and 2 by comparing their respective denominators. In Equation 3, the denominator is simply the product of denominators (for each dimension), whereas in the Hattie model only the exponential terms in the denominator are multiplied. That is, there is no product of probabilities in the denominator as in Sympson's model.

This difference in denominators, and thus in the nature of the models, led Sympson to classify models like Hattie's as compensatory and models similar to his own as noncompensatory (Sympson actually labeled his model partially compensatory). Compensatory models, unlike noncompensatory models, permit high ability on one dimension to compensate for low ability on another dimension in terms of probability of correct response. In the noncompensatory models, the minimum factor (probability) in the denominator is the upper bound for the probability of a correct response. Thus, for a two-dimensional item, a person with very low ability on one dimension and very high ability on the other has a very low probability of correctly answering the item. However, if a compensatory model is employed, this same person would have at least some appreciable probability of correctly answering the item. For example, consider a reading comprehension item. Although reading comprehension is a complex process, two important traits involved in this process are vocabulary skills and the ability to interpret structural cues (Johnston, 1983). It is conceivable that persons with limited vocabularies could be successful on such an item if their skills in interpreting structural cues were great enough. That is, great facility with structural cues might compensate for weak vocabulary ability.

On the other hand, consider a mathematics problem-solving item involving two traits—developing a problem-solving strategy and computation. If a person is very poor in developing a problem-solving strategy, say $\theta_1 = -3$, it seems doubtful, even if $\theta_2 = 3$, that this person would have any chance of correctly answering the item, aside from guessing. That is, if a person cannot develop problem-solving strategies, no amount of computational skill can compensate for this deficiency and thus induce an appreciable probability of correctly responding.

Perhaps the real distinction between compensatory and noncompensatory models lies in the manner in which dimensions are defined. If dimensionality is considered in a factor analytic sense, a test consisting of two dimensions has a group of items tapping each dimension. In such a setting, a compensatory model seems reasonable, since here the test is being considered as a whole. From this point of view, it is not unreasonable to believe that compensation does take place. If a student has just one weak area in a subject being tested, strengths in other areas of that subject can compensate for the weak area and result in a relatively high test score.

However, if, for example, a two-dimensional test is considered to be one which requires the simultaneous application of two abilities to answer each item correctly, the noncompensatory model seems more appropriate. Even in the reading example given above, it could be argued that if a student had sufficiently weak vocabulary skills, no degree of facility with structural cues could overcome this deficiency. It would seem that the noncompensatory view of dimensionality is more reasonable, especially when considering IRT, since IRT does attempt to model the relationship between ability and item responses. Therefore, Sympson's model (1978) was chosen to generate the two-dimensional data used in this study.

Selection of Parameters

In order to simulate realistic data sets using Sympson's model (1978), a trial-and-error procedure was used. First, vectors of item discrimination and difficulty parameters were derived by creating sets of *a*s and *b*s that resembled those typically observed in working with the unidimensional three-parameter logistic model. The *a* values were distributed uniformly at 10 points in the interval from .5 to 2, while the *b* values were distributed uniformly at 30 points between -2 and 2. Next, a series of manipulations was carried out on these sets of *a*s and *b*s. These manipulations consisted of various linear transformations of the vectors of item parameters as well as several different correlations between the vectors of *a*s and between the vectors of *b*s. Based on the outcomes of these trials, the following strategy was chosen for simulating item parameters.

The original a_1 values were scaled to have a mean of 1.23 and a standard deviation of .34, while the a_2 values were centered at .49 with a standard

deviation of .11. The rationale for these values is as follows. Standardized achievement tests are typically written to tap one basic dimension. The presence of other dimensions results from either the nature of the subject matter or the nature of the test items. For example, reading ability is probably a secondary ability dimension on most tests. Since items on these tests are not written primarily to assess this secondary dimension, it seems reasonable that these items would not discriminate highly on that dimension. On the other hand, if, indeed, two ability dimensions are necessary to answer items on a given test, discrimination on the first or more potent dimension should clearly dictate the degree to which the item discriminates. Thus, the a_1 values were scaled to have a mean greater than 1.0. It was also observed that a vectors (scaled as described above) that were moderately negatively correlated ($r = -.29$), in conjunction with the other parameters, seemed to produce realistic data sets.

In the data generation procedure using Sympson's (1978) model, the b values played a major role in determining the realism of the data sets. Preliminary analyses showed that data sets simulated with b vectors centered at zero had average difficulties that were uncharacteristically low relative to those reported for standardized achievement tests. These low values were clearly a function of the noncompensatory nature of Sympson's model. If it is true that ability dimensions for a given test are noncompensatory, the scaling of item difficulty must be reconsidered. To avoid producing test data indicative of an unrealistically difficult test, the b values were scaled to have lower means. The b_1 values were scaled only slightly lower (mean = $-.33$, SD = .82), while the b_2 values were scaled sharply lower (mean = -1.03, SD = .82).

The rationale here was similar to that used for justifying the scaling of the a values. Standardized achievement tests are written to tap just one dimension, and therefore, difficulty levels on secondary ability dimensions are purposefully depressed. For example, on a mathematics problem-solving test there is clearly a reading dimension, but the items are typically written to require reading ability below the grade level of the individuals being tested. Vectors of difficulty parameters, scaled as described above, were generated to be moderately positively correlated ($r = .38$). This simulation strategy appeared to yield very realistic data sets.

Sympson's (1978) model requires one guessing parameter per item. It was found that if each item was assigned a guessing parameter of .2 (in conjunction with the other item parameters described above), realistic data sets (in terms of average difficulty) resulted.

Finally, the θ vectors were generated to fit a bivariate normal distribution, with both dimensions scaled to have a mean of 0.0 and a standard deviation of 1.0. In addition, the correlation between the two ability dimensions [$\rho(\theta_1, \theta_2)$] was varied. Data sets with $\rho(\theta_1, \theta_2)$ values of .0, .3, .6, .9, and .95 were simulated.

Simulation Procedure

Given parameters defined by the specifications detailed above, the following procedure was used to generate data sets:

1. Using the given parameters, the two-dimensional version of Sympson's model was used to create a person-by-item matrix of probabilities. This matrix was of order N (number of examinees) $\times$ k (number of items), with elements p_{ij}.
2. A random number matrix of order $N \times k$ (to be referred to as the comparison matrix) was generated with elements r_{ij} from a uniform distribution (in the range from 0.0 to 1.0).
3. A (0, 1) matrix of order $N \times k$ was generated with elements x_{ij} by application of the following rule:

$$x_{ij} = \begin{cases} 1 \text{ if } p_{ij} \geq r_{ij} \\ 0 \text{ if } p_{ij} < r_{ij} \end{cases} . \quad (4)$$

Data Sets

Four combinations of sample size and test length were used in this study. Specifically, two sample sizes (1,000, 2,000) were used in conjunction with

two test lengths (30, 60). (It has been suggested [Hulin, Drasgow, & Parsons, 1983] that the three-parameter model should not be applied to data sets with $N < 1{,}000$.) Within each of these combinations, 5 two-dimensional data sets were considered (one for each of the values of $\rho[\theta_1, \theta_2]$).

Table 1 presents information relevant to the characteristics of the data sets when $N = 2{,}000$ and $k = 60$. These data sets were representative of all the simulated data sets. The statistics listed in Table 1 indicate that these data sets were very similar to actual test data (cf. Brandenburg, 1972).

Analysis

This investigation was carried out by considering the parameter estimates derived from LOGIST (Wingersky, Barton, & Lord, 1982) from two different perspectives: (1) an item-level perspective, and (2) a total test-level perspective. The total test-level perspective involved a consideration of ability estimates for examinees. The unidimensional estimates of ability were correlated with the true multidimensional ability parameter values as well as with the averages of the true ability parameters. In addition, average absolute differences were used to compare these same quantities. These had the form:

$$\text{AAD} = \frac{\sum_{i=1}^{N} |\theta_{ih} - \hat{\theta}_i|}{N}, \qquad (5)$$

where θ_{ih} ($h = 1$ or 2) is the true ability parameter for person i for dimension 1 or 2,

$\hat{\theta}_i$ is the ability estimate derived by LOGIST for person i, and

N is the number of examinees.

At the item level, the unidimensional estimates of difficulty and discrimination were correlated with the true multidimensional item parameter values as well as with the averages of the true parameters. To further evaluate the relationship between the estimated values and the true parameters, average absolute differences were computed. These were of the form:

$$\text{AAD} = \frac{\sum_{j=1}^{k} |x_{jh} - \hat{x}_j|}{k}, \qquad (6)$$

where x_{jh} ($h = 1$ or 2) is the true discrimination or difficulty parameter of item j for dimension 1 or 2,

$\hat{x}_j$ is the corresponding estimate for item j derived by LOGIST, and

k is the number of items.

One-Dimensional Simulations

In an investigation of this type, it is important to note that deviations between parameters and estimates might be due, at least in part, to errors produced by the estimation procedure. Thus, it was also necessary to obtain an estimate of the accuracy of the estimation program. Therefore, it was decided to assess the relationship between parameters and estimates when LOGIST was applied to data that were truly unidimensional. Once this deter-

Table 1
Descriptive Statistics, Reliabilities (α), Ratios of Eigenvalues (λ_1/λ_2), Difficulty Indices (p), and Item-Total Biserial Correlations (bis) of Two-Dimensional Data Sets with N = 2000 and k = 60

				Range of p		Range of bis			Range of X			
$\rho(\theta_1,\theta_2)$	λ_1/λ_2	α	$\bar{p}$	Lo	Hi	Lo	Hi	$\bar{X}$	Lo	Hi	Skew.	Kurt.
0	8.06	.87	.51	.25	.86	.32	.55	30.70	8	59	.23	-.62
.3	10.45	.88	.52	.25	.85	.32	.58	30.95	8	59	.13	-.72
.6	11.79	.91	.53	.25	.86	.34	.62	31.94	7	59	.16	-.85
.9	12.97	.91	.53	.25	.85	.32	.63	31.61	8	60	.20	-.89
.95	13.76	.92	.54	.26	.86	.40	.66	32.18	7	60	.22	-.86

mination was made, it would provide a frame of reference for interpreting the results for the two-dimensional data sets.

For each two-dimensional data set described above, a corresponding unidimensional data set was generated for comparison purposes. Sets of ability parameters were generated to fit a standard normal distribution. The item parameters, a and b, were the a_1s and the b_1s from the two-dimensional data sets; the c parameter was set equal to .2 for each item. All of these parameters were then used in the unidimensional modified three-parameter logistic model to generate a matrix of probabilities. These probabilities were then compared with the elements of a randomly generated comparison matrix to obtain a 0,1 matrix as described above. As was the case with the two-dimensional data sets, traditional item analyses were carried out on these data. Table 2 summarizes the relevant statistics when $N = 2{,}000$ and $k = 60$.

The data in Table 2 clearly demonstrate that aside from skewness and average difficulty values, these data sets were similar to those typically associated with actual standardized achievement test data. The relatively high average difficulties and the resulting negatively skewed distributions of raw scores were caused by using the b_1 values for b. Recall that these b_1 values were scaled to have a mean of $-.33$ to be consistent with the noncompensatory nature of Sympson's (1978) model. Thus, the unidimensional data sets represented responses to relatively easy tests.

The 0,1 matrix for each unidimensional data set was entered into the LOGIST estimation program, and analyses similar to those described for the two-dimensional data sets were carried out.

Results

Item Parameter Estimates

Discrimination. Table 3 presents the means and standard deviations of the distributions of item discrimination and difficulty estimates for the data sets with $N = 2{,}000$ and $k = 60$. (Since the results of the analyses were highly similar for all combinations of sample size and test length, only data sets with $N = 2{,}000$ and $k = 60$ are discussed.) For all two-dimensional data sets, the means of the true a parameters were 1.23 and .49, respectively, while their standard deviations were .34 and .11. For the unidimensional data sets, the a_1 values were used as the true item discrimination parameters. It is clear from the data in Table 3 that for the unidimensional data sets, the true item discrimination parameters and the discrimination estimates were highly similar in terms of their means and standard deviations. However, for the two-dimensional data sets, the average estimated item discrimination values were between the means of the a_1 values and the a_2 values, with the value of the mean of the discrimination parameter estimates approaching the mean of the a_1 values as $\rho(\theta_1, \theta_2)$ increased. The standard deviations of the estimated a values were typically slightly less than the standard deviations

Table 2
Descriptive Statistics, Reliabilities, Ratios of Eigenvalues, Difficulty Indices, and Item-Total Biserial Correlations of Unidimensional Data Sets with N = 2000 and k = 60

				Range of p		Range of bis			Range of X			
Data Set	λ_1/λ_2	α	$\bar{p}$	Lo	Hi	Lo	Hi	$\bar{X}$	Lo	Hi	Skew.	Kurt.
1	13.52	.94	.66	.29	.93	.41	.82	39.60	8	60	-.24	-.92
2	15.27	.95	.66	.29	.92	.44	.83	39.60	6	60	-.26	-.94
3	15.40	.94	.67	.30	.93	.40	.83	39.96	7	60	-.28	-.92
4	14.67	.94	.66	.30	.92	.44	.84	39.87	9	60	-.28	-.96
5	14.35	.94	.65	.27	.92	.37	.83	39.16	7	60	-.25	-.91

Table 3
Means and Standard Deviations of the Distributions of True Item Discrimination and Difficulty Parameters and Estimates of those Parameters from Two-Dimensional and Unidimensional Data Sets

	True: Dimension		Two-Dimensional: $\rho(\theta_1,\theta_2)$					Unidimensional				
Parameter	1	2	0	.3	.6	.9	.95	1	2	3	4	5
Discrimination												
Mean	1.23	.49	.78	.81	.95	1.00	1.00	1.22	1.22	1.19	1.25	1.22
SD	.34	.11	.24	.24	.32	.34	.36	.33	.33	.31	.35	.34
Difficulty												
Mean	-.33	-1.03	.39	.36	.24	.28	.23	-.30	-.30	-.35	-.32	-.26
SD	.82	.82	.91	.86	.78	.75	.74	.87	.87	.89	.84	.86

of the true a_1 values when $\rho(\theta_1, \theta_2)$ was relatively low and slightly greater when $\rho(\theta_1, \theta_2)$ was relatively high.

Table 4 presents the correlations and average absolute differences between values of the estimates of item discrimination and the values of the true parameters. The last row of this table gives the median values of these statistics computed for the corresponding five unidimensional data sets.

In terms of correlations, clear trends are not present. It might have been expected, for example, to observe an increase in the values of the correlation between the $\hat{a}$ values and the a_1 values as $\rho(\theta_1, \theta_2)$ increased. There was at least a slight tendency for these correlations to increase as the θ vectors became more highly correlated, but the relationship was not monotonic. It is also clear that the true a_2 values were not strongly related to the $\hat{a}$ values; however, given the extremely small variability of the a_2 parameter (SD = .11), this is not surprising. In general, the $\hat{a}$ values were most highly correlated with the averages of the a_1 and a_2 values.

Of major interest is the clear disparity between the magnitudes of the correlations obtained from the two-dimensional data sets and the magnitudes of the correlations from the unidimensional data sets. It should be noted that when the θ vectors are perfectly correlated, Sympson's (1978) model is not equivalent to the unidimensional three-parameter logistic model. (It should also be noted that this is not unique to Sympson's model. The Doody-Bogan and Yen, 1983, model, and the Hattie, 1981, model, [both compensatory models] share this property.) This might explain, in part, the disparity between the two-dimensional values when $\rho(\theta_1, \theta_2) = .95$ and the unidimensional values. It did seem that given unidimensional data, LOGIST was able to estimate item discrimination with reasonable accuracy (at least in a rank order sense). Recall, however, that these results are for data sets

Table 4
Correlations and Average Absolute Differences Between Values of $\hat{a}$ and a_1, a_2, and Average a(aavg) for $\rho(\theta_1,\theta_2)$ Conditions

$\rho(\theta_1,\theta_2)$	$r_{\hat{a},a_1}$	$r_{\hat{a},a_2}$	$r_{\hat{a},aavg}$	$\frac{\Sigma\lvert a_1 - \hat{a}\rvert}{k}$	$\frac{\Sigma\lvert a_2 - \hat{a}\rvert}{k}$	$\frac{\Sigma\lvert \hat{a} - aavg\rvert}{k}$
0	.47	.02	.50	.46	.32	.16
.3	.58	-.01	.60	.43	.34	.14
.6	.58	-.01	.60	.32	.47	.22
.9	.67	-.04	.68	.26	.52	.23
.95	.64	-.05	.65	.27	.51	.23
Mdn 1-D Value	.94	-	-	.09	-	-

with $N = 2{,}000$. The results for data sets with $N = 1{,}000$ were very similar, but smaller data sets ($N < 1{,}000$) were not examined.

The differences between the $\hat{a}$ and the a_1 values, as shown by the AADs, generally decreased as $\rho(\theta_1, \theta_2)$ increased, while the differences between the $\hat{a}$ and the a_2 values generally increased. For the two data sets with vectors of θs correlated 0 and .3, the AAD values involving a_2 were less than those involving a_1, whereas for the other data sets, this situation was reversed. However, the strongest relationship, as defined by the AADs, was consistently that between the $\hat{a}$ values and the average of the true a values. As was the case with the correlations, the magnitudes of the differences for the two-dimensional data sets were clearly larger than the magnitudes of the differences for the unidimensional data sets.

Difficulty. Descriptive statistics for the item difficulty parameters and estimates were also computed. These values are reported in Table 3. The means of the true values of b_1 and b_2 were $-.33$ and -1.03, respectively, while their standard deviations were both .82. As was the case with the discrimination parameters, the means of the $\hat{b}$ values and the b values for the unidimensional data sets were very similar, but the standard deviations of the $\hat{b}$ values were consistently greater than the standard deviations of the b values. In contrast, the standard deviations of the $\hat{b}$ values for the two-dimensional data sets were consistently greater than those of either the b_1 or the b_2 values when the vectors of θs were weakly correlated, while the reverse was true as the correlation between θ vectors increased. Also, for the two-dimensional data sets, the means of the item difficulty estimates were consistently greater than the means of either the b_1 or the b_2 values. These greater average values were due to the noncompensatory nature of the IRT model used to generate the data. There was a clearly decreasing trend in both the means and standard deviations of the difficulty parameter estimates as the correlation between θ vectors increased. This appeals to intuition, in that a test with two clearly distinct dimensions (e.g., θ vectors correlated 0) would seem to be more difficult than a test with highly related dimensions (e.g., θ vectors correlated .95), assuming that other parameters were constant. These decreasing mean values were also consistent with the average p values for these data sets.

Table 5 presents the correlations and AADs used for comparing the true and estimated difficulty parameters. The correlations reported here are clearly much larger than those associated with the item discrimination parameters. The relationships between the b_1 and the $\hat{b}$ values were quite strong ($r > .87$). There were also rather strong relationships between the b_2 and the $\hat{b}$ values ($r > .69$). It is thus not surprising that the $\hat{b}$ values and the averages of b_1 and b_2 were very highly correlated ($r > .95$).

Once again, the last row in this table contains median values from the unidimensional data sets. As was the case with the item discrimination parameters, there was a clear disparity between the magnitudes of the correlations for the two-dimensional data sets with $\rho(\theta_1, \theta_2) = .95$ and the magnitudes of the correlations for the unidimensional data sets. The very high correlations for the unidimensional data sets indicated that given unidimensional data, LOGIST rank ordered the $\hat{b}$ values in a manner very similar to that of the corresponding true b values. Given two-dimensional data, it appeared that the rank order of the $\hat{b}$ values was most similar to that of the average of the b_1 and the b_2 values, with the potency of the major dimension having no role in the magnitude of this relationship.

AADs between values of true item difficulty parameters and difficulty estimates displayed clear trends across data sets. Differences between the $\hat{b}$ and the b_1 values and the $\hat{b}$ and the b_2 values decreased as the vectors of thetas became more highly correlated. As was indicated by the correlations, the $\hat{b}$ values were more highly related to the b_1 values than to the b_2 values in all data sets. The differences between the $\hat{b}$ values and the averages of b_1 and b_2 were always approximately midway between the differences between the $\hat{b}$ and the b_1 values and the differences between the $\hat{b}$ and the b_2 values. It is important to note that these AADs

Table 5
Correlations and Average Absolute Differences Between Values of $\hat{b}$ and b_1, b_2, and Average b(bavg) for $\rho(\theta_1,\theta_2)$ Conditions

$\rho(\theta_1,\theta_2)$	$r_{\hat{b},b_1}$	$r_{\hat{b},b_2}$	$r_{\hat{b},bavg}$	$\frac{\Sigma\lvert b_1 - \hat{b}\rvert}{k}$	$\frac{\Sigma\lvert b_2 - \hat{b}\rvert}{k}$	$\frac{\Sigma\lvert \hat{b} - bavg\rvert}{k}$
0	.88	.73	.97	.74	1.44	1.08
.3	.89	.72	.97	.70	1.40	1.04
.6	.90	.70	.96	.57	1.27	.92
.9	.90	.71	.97	.61	1.31	.96
.95	.90	.72	.97	.56	1.26	.91
Mdn 1-D Value	.998	-	-	.07	-	-

were for the most part simple algebraic differences, since the $\hat{b}$ values were consistently greater than the b_1 and the b_2 values. It is also useful to note the distinction between the information conveyed by correlations and AADs. The high correlations between the $\hat{b}$ values and the averages of b_1 and b_2 indicated that these two variables have similar rank orderings. On the other hand, the relatively small AADs between the $\hat{b}$ values and the b_1 values reflect the fact that $\hat{b}$ was most clearly an estimate of b_1, albeit consistently an overestimate. It should again be noted that the results reported here are for large data sets and may be unique to the model used for data generation.

As before, AADs involving item difficulty estimates were computed for the unidimensional data sets. These differences were clearly smaller than those from the two-dimensional data sets, even those with θ vectors correlated .95.

Ability Estimates

Correlations and average absolute differences between values of the true ability parameters (θ) and the ability estimates ($\hat{\theta}$) are presented in Table 6. The correlations involving θ clearly increased, and the AADs decreased as $\rho(\theta_1, \theta_2)$ increased. When the correlation between the θ vectors was weak, the $\hat{\theta}$ values were much more highly related to the θ_1 values than to the θ_2 values. As the correlation between the vectors of θs increased, the relationship between the $\hat{\theta}$ and the θ_1 values became somewhat stronger, while the relationship between the $\hat{\theta}$ and the θ_2 values became markedly stronger. For the highest values of $\rho(\theta_1, \theta_2)$, the $\hat{\theta}$ values were nearly equally related to the θ_1 and the θ_2 values. There was some evidence that the $\hat{\theta}$ values were most highly related to the averages of the true θs; however, it was clear that this rela-

Table 6
Correlations and Average Absolute Differences Between Values of $\hat{\theta}$ and θ_1, θ_2, and Average $\theta(\theta avg)$ for $\rho(\theta_1,\theta_2)$ Conditions

$\rho(\theta_1,\theta_2)$	$r_{\hat{\theta},\theta_1}$	$r_{\hat{\theta},\theta_2}$	$r_{\hat{\theta},\theta avg}$	$\frac{\Sigma\lvert \theta_1 - \hat{\theta}\rvert}{N}$	$\frac{\Sigma\lvert \theta_2 - \hat{\theta}\rvert}{N}$	$\frac{\Sigma\lvert \theta avg - \hat{\theta}\rvert}{N}$
0	.78	.32	.78	.50	.94	.48
.3	.78	.50	.81	.50	.83	.45
.6	.82	.70	.84	.43	.63	.37
.9	.85	.81	.85	.36	.44	.35
.95	.87	.85	.87	.33	.38	.32
Mdn 1-D Value	.93	-	-	.22	-	-

tionship was not much stronger than that between the $\hat{\theta}$ and the θ_1 values. This was not consistent with the results reported by Reckase (1979). He suggested that the θ estimates were most highly related to factor scores for the first dimension only when that dimension was clearly predominant. In data sets lacking a predominant dimension, it was reported that the θ estimates were most highly related to the factor scores for just one of the dimensions (not necessarily the first), instead of to an average of factor scores. Note, however, that the data generation procedures used by Reckase differed from those used in this study.

As before, the last row of this table contains the median values of correlations or differences calculated for the unidimensional data sets. For both indices, there was a clear disparity between the magnitudes of the statistics from the two-dimensional data sets with $\rho(\theta_1, \theta_2) = .95$ and the magnitudes of the statistics from the unidimensional data sets.

Summary

The purpose of this investigation was to examine the nature of the unidimensional parameter estimates derived from two-dimensional data. Before summarizing the results, it is important to note the major limitations of this study. Only two-dimensional data were considered. Also, this study was based entirely on data generated to fit a noncompensatory two-dimensional IRT model. Research has not substantiated that item response data are noncompensatory.

Within these limitations, it was found that the $\hat{a}$ values were best considered as averages of the true a_1 and a_2 values. On the other hand, the $\hat{b}$ values seemed best thought of as overestimates of the true b_1 values. And, finally, it was fairly clear that the estimated θ values were most highly related to the averages of the true θ values.

It thus seems clear that the results of this investigation demonstrate that violations of the assumption of unidimensionality do have an effect on parameter estimation for the modified three-parameter logistic model. It was almost always true that as the vectors of θs became more highly correlated, the values of statistics derived from the two-dimensional data sets approached the values of statistics derived from the unidimensional data sets. However, it was also consistently true that the magnitudes of those statistics derived from data sets with θ vectors correlated .95 and those derived from unidimensional data sets were clearly disparate.

Given that the application of a unidimensional IRT model to two-dimensional data yielded parameter estimates that were not clearly interpretable, and also given that many achievement tests are multidimensional, it appears that unidimensional parameter estimates derived from the application of the modified three-parameter logistic model to achievement test data must be interpreted with caution. Thus, until multidimensional IRT develops sufficiently so that it is clear which model should be applied in a given situation and how the parameters associated with that model can be reliably estimated, it seems reasonable to assume that if IRT is to fulfill its potential, measurement practitioners must develop tests specifically to fit unidimensional models.

References

Brandenburg, D. C. (1972). *The use of multiple matrix sampling in approximating an entire empirical norms distribution*. Unpublished doctoral dissertation, The University of Iowa.

Doody-Bogan, E., & Yen, W. M. (1983). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189–199.

Hambleton, R. K., & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 71–94). British Columbia: Educational Research Institute of British Columbia.

Hambleton, R. K., Swaminathan, H., Cook, L., Eignor, D., & Gifford, J. (1978). Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research, 48,* 467–510.

Hattie, J. (1981). *Decision criteria for determining unidimensionality*. Unpublished doctoral dissertation, University of Toronto, Canada.

Hulin, C., Drasgow, F., & Parsons, C. (1983). *Item response theory: Application to psychological measurement*. Homewood IL: Dow Jones-Irwin.

Hutten, L. R. (1980). *Some empirical evidence for latent trait model selection*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Johnston, P. (1983). *Reading comprehension assessment: A cognitive basis*. Newark DE: International Reading Association.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Lawrence Erlbaum.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 178–193.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Reckase, M. D., & McKinley, R. L. (1985). Some latent trait theory in a multidimensional latent space. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 248–280). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Rentz, R. R., & Rentz, C. C. (1979). Does the Rasch model really work? *Measurement in Education, 10,* 1–12.

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53–61.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82–98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57–70). British Columbia: Educational Research Institute of British Columbia.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton NJ: Educational Testing Service.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST: A program for estimating ability and item characteristic curve parameters*. Princeton NJ: Educational Testing Service.

Author's Address

Send requests for reprints or further information to Timothy N. Ansley, 318 Lindquist Center, University of Iowa, Iowa City IA 52242, U.S.A.