Application of Unidimensional Item Response Theory Models to Multidimensional Data

Fritz Drasgow
University of Illinois

Charles K. Parsons
Georgia Institute of Technology

A simulation model was developed for generating item responses from a multidimensional latent trait space. The model permits the prepotency of a general latent trait underlying responses to all simulated items to be varied systematically. Five levels of prepotency were used to generate data sets. The levels of prepotency ranged from a truly unidimensional latent trait space to a very weak general latent trait. Simulated item pools with guessing and without guessing were analyzed by the LOGIST computer program. The general latent trait was recovered in data sets where the prepotency of the general latent trait was only moderate. Consequently, it appears that item response theory models can be applied to moderately heterogenous item pools under the conditions simulated here.

Lord and Novick (1968, p. 383) stated, "it can be taken for granted that every [item response theory] model is false and that we can prove it so, if only we collect a sufficiently large sample of data." One way in which most currently available item response theories (IRTs) are surely incorrect is in their assumption of a unidimensional latent trait space. This assumption, which implies local independence of item scores, is an essential part of the theory underlying most currently available parameter estimation procedures.

The implications of violations of the unidimensionality assumption for testing practitioners and substantive researchers require careful examination. The present research examines the effects of a multidimensional latent trait space on estimation of item and person parameters by the widely used computer program LOGIST (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976). In this program the method of maximum likelihood and a number of ad hoc techniques are used to estimate item and person parameters of the one-, two-, or three-parameter logistic models. The assumption of unidimensionality is used in the derivation of the likelihood function maximized by LOGIST (see Lord, 1980, p. 19). Consequently, estimates computed by LOGIST have a theoretical justification only in the case of a unidimensional latent trait space. From a practical perspective it is important to know the extent to which LOGIST's parameter estimates are robust to violations of unidimensionality. In particular, when is an item pool "sufficiently unidimensional" for parameter estimates to be useful to testing practitioners and substantive researchers?

The intuitions used in the present research to formalize the notion of "sufficiently unidimensional" are based on three examples:

1. A test of verbal ability that has subsections composed of antonyms, analogies, and paragraph comprehension questions;
2. A test of algebra achievement that asks questions based on each of several parts of a high school algebra course; and
3. An instrument measuring overall job satisfac-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 2, Spring 1983, pp. 189–199

0146-6216/83/020189-11$1.55

tion that asks questions about workers' affective responses to a number of job characteristics including their supervision, pay, and coworkers.

In each of these examples, interest is centered upon a single latent trait that underlies responses to all items in the item pool. However, it is clear that none of the three instruments is truly unidimensional. In particular, clusters of items are likely to be more highly related than expected on the basis of a single latent trait. An item pool is defined to be "sufficiently unidimensional" to allow application of an IRT model and estimation procedure if the estimation procedure recovers the general latent trait that underlies responses to all items in the item pool.

Previous research studying the accuracy of LOGIST has generally used simulated item responses that meet the assumptions (including unidimensionality) of the model fitted to the data. Under these conditions, Lord (1975), Swaminathan and Gifford (1979), Hulin, Lissak, and Drasgow (1982), and others have found that LOGIST provides effective parameter estimation when sample size (N) and test length (n) are sufficiently large. Provided that the assumptions are satisfied, $N \geq 1000$ and $n \geq 50$ appear to be adequate for the three-parameter logistic model. Minimum requirements for the two-parameter logistic model are less restrictive.

Reckase (1979) conducted one of the few studies that examined the effectiveness of LOGIST (or any estimation technique) when the unidimensionality assumption is violated. He generated a data set with an underlying dominant latent trait that was related to all items as well as weaker latent traits that affected clusters of items. LOGIST was found to be robust to these minor violations of the unidimensionality assumption in the sense that the dominant latent trait was well recovered. Reckase also simulated a test composed of items that were factorially pure measures of two statistically independent latent traits. Here LOGIST was drawn to one of the two latent traits: ability estimates were highly correlated ($r = .93$) with estimated factor scores on one factor and nearly uncorrelated ($r = .29$) with estimated factor scores on the other factor. In addition, estimates of the item discrimination parameter were generally greater than 1.70 for items related to the former trait and less than .15 for items related to the latter trait.

It is useful to think of the prepotency of a general latent trait as varying along a continuum. At one extreme the latent space is truly unidimensional. Reckase's case of a dominant general trait corresponds to a small move along the prepotency continuum away from unidimensionality. Note that this simulated item pool was sufficiently unidimensional for LOGIST to recover the general trait. Reckase's hypothetical item pool composed of factorially pure items measuring two independent traits lies at the other extreme of the continuum. Here there was no general factor, which is equivalent to a general factor with zero prepotency. It is obviously impossible for any estimation technique to recover a general trait at this end of the prepotency continuum.

In the present research, several item pools were simulated that ranged from truly unidimensional to an inconsequential general latent trait. Item pools with intermediate levels of prepotency of the general latent trait were also constructed. These item pools were used to determine the degree of prepotency that is required by LOGIST in order to recover the general latent trait and *not* be drawn to a latent trait underlying a cluster of items.

Method

Simulation Model and Parameters

The simulation model used in the present research consisted of the following components. First, correlated common factors were simulated using the hierarchical factor analysis model proposed by Schmid and Leiman (1957). A single second-order *general* factor controlled the correlations of the first-order common factors. Latent *item propensity scores* were generated for n hypothetical items using the underlying factors. Finally, dichotomous item scores were created by determining whether item propensity scores were above or below their respective threshold values. In the remainder of

this section, details of the simulation model are described more fully.

The common factor model can be written

$$\mathbf{x} = \mathbf{Ay} + \mathbf{Be}, \quad [1]$$

where

- **x** is a vector containing the n observed variables x_i (here, however, "observed" variables are not observed but instead are considered as item propensity variables that underlie observed item responses);
- **A** is the $n \times k$ matrix of loadings on the k common factors;
- **y** is a vector containing the k common factors y_i;
- **B** is an $n \times n$ diagonal matrix with loadings on unique factors along its diagonal; and
- **e** is a vector containing the n unique variables e_i.

The unique variables are assumed to be mutually uncorrelated and uncorrelated with the common factors. Let α_{ij} denote the loading of the ith item propensity variable on the jth factor and let β_i denote the single loading of the ith item propensity variable on the ith unique variable. The item propensity variables, common factors, and unique factors are all scaled to have mean zero and unit variance.

The factor loading matrix **A** used in the present research is shown in Table 1. Note that each of the 50 item propensity variables loads on a single common factor and that there are five common factors. The first common factor is related to 15 item propensity variables, the second is related to 5 item propensity variables, and the rest are related to 10 item propensity variables. Since the magnitudes of the factor loadings are comparable across all five common factors, it is apparent that the first common factor is the most influential in this particular item pool.

The simple pattern of factor loadings in Table 1 was selected for several interrelated reasons. First, simple structure (Thurstone, 1947, chap. 14) is a convenient means for specifying a frame of reference in the factor space that eliminates the rotational indeterminacy of factors. Moreover, rotations to approximate simple structure seem possible for each of the three examples described previously. With actual data from ability tests, achievement tests and attitude assessments, it is usually necessary for the common factors to be correlated in order to rotate to simple structure. Let $\mathbf{\Phi}$ denote the matrix of correlations of first-order common factors after rotation to oblique simple structure.

A fundamental assumption of the simulation model is that a single *second-order general factor* accounts for the first-order common factor correlations in $\mathbf{\Phi}$. The substantive motivation for this assumption can be seen by again considering the examples previously mentioned. For each of the three instruments, all items should be affected by a general latent trait: general verbal ability, general algebra achievement, and overall job satisfaction, respectively. The way in which the IRT unidimensionality assumption is violated is by clusters of items with larger within-cluster correlations than would be expected on the basis of a single first-order common factor. To account for these within-cluster correlations, it is usually necessary to extract several first-order common factors. Rotation to approximate simple structure might yield a factor-loading matrix analogous to the idealized **A** matrix in Table 1 and a factor correlation matrix $\mathbf{\Phi}$ with large factor intercorrelations. If $\mathbf{\Phi}$ were then factor analyzed, a single general factor (i.e., a second-order common factor) would emerge. This would result because all items are affected by a single general latent trait.

Schmid and Leiman (1957) proposed a convenient model for the factor analysis of correlated common factors. A special case of this general model is used in the present research. Here common factors **y** are represented by

$$\mathbf{y} = \mathbf{fz} + \mathbf{Dv}, \quad [2]$$

where

- **f** is the k element vector containing loadings of the first-order common factors on the single general factor z,
- **D** is a $k \times k$ diagonal matrix containing loadings of the first-order common factors on the k *second-order group factors* (i.e.,

Table 1
Factor Loading Matrix **A** and Item Threshold Values

	Common Factor							Common Factor					
Item	1	2	3	4	5	γ	Item	1	2	3	4	5	γ
1	.4					.85	26			.4			- .53
2	.5					- .53	27			.5			1.30
3	.6					- .85	28			.6			- .85
4	.7					- .26	29			.7			- .26
5	.8					.26	30			.8			.26
6	.4					-1.30	31				.4		-1.30
7	.5					.53	32				.5		.00
8	.6					- .85	33				.6		.85
9	.7					.00	34				.7		.53
10	.8					1.30	35				.8		-1.30
11	.4					1.30	36				.4		- .85
12	.5					-1.30	37				.5		.26
13	.6					.53	38				.6		- .26
14	.7					- .53	39				.7		- .53
15	.8					.00	40				.8		1.30
16		.4				- .85	41					.4	.53
17		.5				-1.30	42					.5	1.30
18		.6				.85	43					.6	-1.30
19		.7				.00	44					.7	-1.30
20		.8				- .26	45					.8	- .26
21			.4			-1.30	46					.4	- .53
22			.5			-1.30	47					.5	.85
23			.6			.53	48					.6	.26
24			.7			.85	49					.7	- .85
25			.8			.00	50					.8	.00

Note: Only nonzero factor loadings are shown; zero loadings have been left as blanks.

second-order specific factors), and **v** is a vector containing the k group factors. The group factors are assumed to be mutually uncorrelated and uncorrelated with z. Again, all factors are scaled to have unit variance.

The second-order general factor z can be interpreted as the general latent trait underlying responses to all items in the item pool. The elements v_j of **v** correspond to additional latent variables, uncorrelated with z, that cause clusters of items to be more highly related than would be expected on the basis of a unidimensional latent trait. For example, if z were general verbal ability, then v_j might represent an individual's ability to solve, say, analogies, after controlling for general verbal ability.

The elements f_j of **f** and diagonal elements δ_j of **D** control the relative importance of the general and group factors. The combinations of **f** and **D** used in the present research are shown in Table 2. The first combination shown in Table 2, labeled Latent Structure 1, has five common factors that are perfectly correlated; therefore, the latent trait

Table 2
Loadings of Common Factors on General and Group Factors

	Latent Structure									
Common	1		2		3		4		5	
Factor	f_j	δ_j	f_j	δ_j	f_j	δ_j	f_j	δ_j	f_j	δ_j
1	1.00	.00	.95	.31	.70	.71	.65	.76	.40	.92
2	1.00	.00	.90	.44	.65	.76	.55	.84	.35	.94
3	1.00	.00	.85	.53	.70	.71	.60	.80	.10	.99
4	1.00	.00	.80	.60	.80	.60	.45	.89	.20	.98
5	1.00	.00	.95	.61	.75	.66	.55	.84	.25	.97

Note: Correlations between common factors ranged from .68 to .90 for latent structure 2, from .46 to .60 for latent structure 3, from .25 to .39 for latent structure 4, and from .02 to .14 for latent structure 5.

space is truly unidimensional. Results for this latent structure are used as baseline values; it is unlikely that actual data sets would ever be truly unidimensional. The prepotency of the general factor gradually decreases across the remaining four combinations. Latent Structure 2 was designed to simulate the test of verbal ability described previously. The oblique common factors, which might correspond to factors associated with the various item types, have intercorrelations that range from .68 to .90. Analytic rotation methods designed to rotate to oblique simple structure would be likely to encounter difficulties in recovering such highly correlated factors. Moderately heterogeneous achievement tests and attitude assessment instruments are simulated by Latent Structure 3. Here intercorrelations of the oblique common factors range from .46 to .60. Latent Structure 4 was designed to simulate broad ranged achievement tests and attitude assessment instruments. The oblique common factors from this latent structure have intercorrelations that range from .25 to .39. Finally, the general factor from Latent Structure 5 is very weak, perhaps corresponding to "method variance," rather than a psychologically meaningful trait.

The item propensity variables x_i in Equation 1 can be obtained directly from the second-order general and group factors and first-order unique factors by

$$\mathbf{x} = \mathbf{A}[\mathbf{f} \vdots \mathbf{D}] \begin{bmatrix} z \\ \cdots \\ \mathbf{v} \end{bmatrix} + \mathbf{Be}$$

$$= [\mathbf{g} \vdots \mathbf{S}] \begin{bmatrix} z \\ \cdots \\ \mathbf{v} \end{bmatrix} + \mathbf{Be}, \qquad [3]$$

where $\mathbf{g} = \mathbf{Af}$ is the n element vector containing loadings of item propensity variables on the general factor and $\mathbf{S} = \mathbf{AD}$ is the $n \times k$ matrix containing loadings of item propensity variables on the group factors. When $\mathbf{A}$ has a simple pattern of loadings as in Table 1, the ith diagonal element of $\mathbf{B}$ can be computed by

$$\beta_i = \sqrt{1 - \sum_j \alpha_{ij}^2}. \qquad [4]$$

In the present research Equation 3 was used to generate the item propensity variables; and $\mathbf{A}$, $\mathbf{f}$, and $\mathbf{D}$ from Tables 1 and 2 were used to compute $\mathbf{g}$ and $\mathbf{S}$. Note that the first group factor v_1 is related

to 15 item propensity variables; v_2 is related to 5 item propensity variables; and v_3, v_4, and v_5 are each related to 10 item propensity variables. The group factor z, the v_j, and the e_i were all generated as independent, normal (0,1) variables by the IMSL (1975) subroutine GGNPM.

Item responses u_i were simulated by dichotomizing the item propensity variables:

$$u_i = \begin{cases} 1 \text{ if } x_i \geq \gamma_i \\ 0 \text{ if } x_i < \gamma_i. \end{cases} \quad [5]$$

The threshold values γ_i are presented in Table 1. These values were sampled from a uniform distribution on approximately the nine decile points of the normal distribution.

Data Sets

Samples of $N = 1{,}000$ simulated examinees were created using Equations 3 and 5 for each of the five latent structures in Table 2. These samples are labeled Data Sets 1 through 5, respectively. Note that item responses were generated by a process that is exactly equivalent to the two-parameter normal ogive model (Lord & Novick, 1968, chap. 16) when the truly unidimensional Latent Structure 1 is used in Equation 3, i.e., in Data Set 1.

Guessing can be simulated by first generating item responses by Equations 3 and 5 and then, if $u_i = 0$, rescoring u_i to be correct with probability c_i. Item responses with guessing for samples of $N = 1{,}500$ were generated using each of the five latent structures in Table 2, with $c_i = .15$ for even-numbered items and $c_i = .20$ for odd-numbered items. These samples are labeled Data Sets 6 through 10.

Criteria for Evaluation of Parameter Estimates

Item parameters. Lord and Novick (1968, p. 375) have derived two important relations between the two-parameter normal ogive model and factor analysis model. For Data Set 1 the methods used in the present simulation correspond exactly to the assumptions made by Lord and Novick. Using these assumptions, Lord and Novick showed that

$$a_i = \frac{g_i}{\sqrt{1 - g_i^2}}, \quad [6]$$

where a_i is the item discrimination parameter for item i, and g_i is the loading of the ith item propensity variable x_i on the general factor. They also proved that the item difficulty parameter b_i is

$$b_i = \frac{\gamma_i}{g_i}. \quad [7]$$

Note that Equation 7 implies that b_i is undefined when g_i is zero.

For the multidimensional Data Sets 2 through 5, Equation 6 can be applied to the loading g_i on the general factor. An equation analogous to Equation 6 can also be defined:

$$\breve{a}_{ij} = \frac{s_{ij}}{\sqrt{1 - s_{ij}^2}}, \quad [8]$$

where s_{ij} is the loading of the ith item propensity variable x_i on the jth group factor v_j in Equation 3.

It appears reasonable to conclude that LOGIST is robust to violations of unidimensionality to the extent that elements of **g**, after transformation by Equations 6 and 7, are related to estimates of a and b, and elements of **S** after transformation by Equation 8, are *not* related to estimates of a. The measures of association used in the present research are root mean squared differences:

$$\text{RMSD for a} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{a}_i - a_i)^2}, \quad [9]$$

$$\text{RMSD for } \breve{a}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{a}_i - \breve{a}_{ij})^2}, \quad [10]$$

and

$$\text{RMSD for b} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{b}_i - b_i)^2}, \quad [11]$$

where $\hat{a}_i$ and $\hat{b}_i$ are estimates of a_i and b_i, respectively, obtained from LOGIST. Equations 6 and 7

can also be used when there is guessing. Thus the RMSDs in Equations 9 through 11 can be computed, and the RMSDs for Data Sets 1 and 6 can be used as baseline values for interpreting RMSDs for Data Sets 2 through 5 and 7 through 10.

Person parameters. Lord and Novick (1968) showed that the person parameter θ of the two-parameter normal ogive model is the general factor z under conditions equivalent to those implied by the first latent structure of Table 2 when there is no guessing. Their results can also be used to derive the same relation when guessing occurs. Again, it is reasonable to conclude that LOGIST is robust to violations of unidimensionality to the extent that the estimates $\hat{\theta}$ of person parameters are strongly related to z and are not related to the group factors v_i. Convenient measures of association are the product-moment correlations $r(\hat{\theta}, z)$ between $\hat{\theta}$ and z and $r(\hat{\theta}, v_i)$ between $\hat{\theta}$ and each of the v_i.

Results

Item Parameters

Estimates of a and b parameters were obtained for Data Sets 1 through 5 (i.e., the data sets without guessing) by LOGIST using default convergence criteria. Values of c_i were fixed at 0 so that LOGIST was fitting the two-parameter logistic model to item responses.[1]

Based on earlier studies (e.g., Hulin et al., 1982), it was expected that the numbers of items (n = 50) and examinees (N = 1,000) would be large enough for accurate estimation of item parameters when the unidimensionality assumption was satisfied. Table 3, which presents item parameter RMSDs for Data Sets 1 through 5, shows that N and n were indeed sufficiently large. For Data Set 1 the RMSD for a is .136 and the RMSD for b is .220, both of which are small enough for most practical purposes.

The effects of the decreasing importance of the general factor in determining item responses across Latent Structures 2 through 5 are clear in Table 3. The RMSD of a increases only a small amount from the baseline value of .136 for Data Set 1 to .172 for Data Set 3. Then, there is a moderate increase to .223 in Data Set 4 and large increase to .455 in Data Set 5. The first group factor v_1, which underlies responses to the largest number (15) of items, has a RMSD for $\tilde{a}_1$ that is markedly lower than any of the other group factors in Data Set 4. In Data Set 5, the RMSD for $\tilde{a}_1$ is much lower than the RMSD for the general factor. Thus, in Data Set 4 it appears that LOGIST has been drawn to a latent trait that is a *composite* of the general factor and the most influential group factor. It is clear that LOGIST has been drawn to the first group factor in the fifth data set.

These conclusions are further supported by examining the estimates of b. In Data Set 5, values of $\hat{b}$ seem to be approaching the values that would result from applying Equation 7 to the loadings of items on the first group factor. Values of $\hat{b}$ for the 15 items with nonzero loadings on the first group factor are quite close to the factor loadings on the first group factor transformed by Equation 7: the RMSD of b for these 15 items is .324. In contrast, values of $\hat{b}$ are excessively large for the 35 items with zero loadings on the first group factor. Application of Equation 7 to an item with a zero loading produces an infinite value of b. Thus, the large values of $\hat{b}$ obtained for items with zero loadings on the group factor support the conclusion that LOGIST has been drawn to the first group factor in Data Set 5.

The results for the five data sets with guessing are also shown in Table 3. Default convergence criteria were again used for LOGIST, but $\hat{c}_i$ was free to vary as would be appropriate if LOGIST were used to estimate parameters of multiple-choice test items.

Although the RMSDs for a increase more quickly for data sets with guessing than for data sets without guessing, the pattern of results is generally similar. In Data Set 8 the RMSD for a (.316) is much smaller than the RMSD for $\tilde{a}_1$ (.648). These two RMSDs are much more similar in magnitude in Data Set 9: .402 and .491. In Data Set 10 the

[1] Birnbaum (1968), p. 399) showed that appropriately scaled logistic models are virtually indistinguishable from corresponding normal ogive models.

Table 3
Root Mean Squared Differences Obtained by Equations 9, 10, and 11

	RMSD for					
Data Set	a	$\tilde{a}_1$	$\tilde{a}_2$	$\tilde{a}_3$	$\tilde{a}_4$	$\tilde{a}_5$	b
Data Sets With No Guessing							
1	.136	----	----	----	----	----	.220
2	.149	.775	.811	.778	.782	.796	.292
3	.172	.519	.638	.603	.598	.594	.474
4	.223	.377	.569	.544	.561	.538	.755
5	.455	.225	.628	.696	.678	.668	4.093
Data Sets With Guessing							
6	.209	----	----	----	----	----	.194
7	.247	.847	.890	.864	.867	.873	.368
8	.316	.648	.758	.721	.697	.699	.735
9	.402	.491	.702	.688	.720	.698	.817
10	.535	.319	.704	.761	.743	.721	9.374

RMSD for $\tilde{a}_i$ is substantially smaller than the RMSD for a.

The RMSD for b in Data Set 8 is .735, a value that is considerably larger than the corresponding RMSDs for Data Sets 6 and 7. This large value results in part from two sources. First, it is evident that as loadings on the general factor decrease, values of b_i in Equation 7 increase. For example, Item 11 has b_{11} values of 3.25, 3.42, and 4.64 in Data Sets 6, 7, and 8. Second, Hulin et al. (1982) found that it is very difficult to estimate parameters of three-parameter logistic items with values of b that are large in magnitude. This finding is not particularly surprising. An unexpected result obtained by Hulin et al., however, is that the problems encountered with two-parameter logistic item responses are much less severe. For example, with a simulated test of 1,000 examinees and 30 items, Hulin et al. found that the correlation between estimated and actual b_i values was .995 for two-parameter logistic item responses but only .623 for three-parameter logistic item responses.[2] Interestingly, the correlation between estimated and actual b_i values was .949 for three-parameter logistic items with $|b_i| < 2.0$. Thus, it is extreme values of b_i that are particularly difficult to estimate for the three-parameter logistic model. In analysis of variance terminology, there is an interaction between item response model and magnitude of b_i in determining the accuracy of estimation of b_i.

In sum, the large Data Set 8 RMSD for b could be due to the interaction of item response model and b_i magnitude (which is relevant because multidimensionality has caused many b_i values to become large in magnitude) or due to LOGIST being drawn to the first group factor. From the results concerning the RMSD for a, it appears that the former interpretation should be adopted.

Person Parameters

The correlations between estimates $\hat{\theta}$ of ability computed by LOGIST and factor scores appear in Table 4. These correlations were computed from the actual factor scores used in Equation 3, not from factor score *estimates*. The results in Table 4 are orderly and compelling: As the prepotency of the general factor decreases, $r(\hat{\theta},z)$ values de-

[2]Hulin et al. (1982) used exactly the same population values of a and b when simulating two- and three-parameter logistic item responses.

Table 4
Correlations Between Ability Estimates and Factor Scores

Data Set	z	Factor v_1	v_2	v_3	v_4	v_5
Data Sets With No Guessing						
1	.965	----	----	----	----	----
2	.939	.044	.071	.173	.076	.053
3	.843	.322	.078	.154	.129	.153
4	.736	.461	.061	.139	.212	.225
5	.376	.774	.064	.039	.062	.083
Data Sets With Guessing						
6	.939	----	----	----	----	----
7	.909	.119	.029	.078	.097	.096
8	.828	.278	.044	.118	.136	.206
9	.722	.444	.126	.145	.085	.127
10	.348	.709	-.030	.015	.056	.153

Note: Correlations between group factors and ability estimates are zero for Data Sets 1 and 6.

crease and $r(\hat{\theta},v_1)$ values increase. For both data sets with guessing and data sets without guessing, $\hat{\theta}$ is virtually unrelated to v_1 when item responses are generated from the second latent structure shown in Table 2. The correlations between v_1 and $\hat{\theta}$ are only about .3 when item responses are generated from the third latent structure. Here $r(\hat{\theta},z)$ is about .83, which is large enough for many practical applications. In Data Sets 4 and 9, $\hat{\theta}$ values are more highly related to the general factor than to the first group factor, but it is clear that LOGIST's ability estimates are strongly influenced by both z and v_1. It is apparent that LOGIST has been drawn to v_1 in Data Sets 5 and 10.

Discussion

The types of multidimensionality studied here have several effects on the estimation techniques programmed in LOGIST. Perhaps most important is that as the prepotency of the general factor decreases, LOGIST is gradually drawn to the strongest group factor. RMSDs for a and RMSDs for b increase, slowly at first and then more rapidly, as the latent structure varies across Levels 1 through 5 in Table 2. As the prepotency of the general factor decreases, the effects on the RMSDs are paralleled by decreasing correlations between $\hat{\theta}$ and z and increasing correlations between $\hat{\theta}$ and v_1.

Estimates of item difficulty occasionally become excessively large in magnitude when actual data sets are analyzed by LOGIST[3]. The results obtained here indicate that this phenomenon may partially be due to multidimensional item pools. Of course, these items may be poorly written, too easy or too difficult. Nonetheless, Equation 7 and Table 3 indicate that decreasing the prepotency of the general factor (for example, by increasing the number of content areas of an achievement test) may cause some items to have values of b that are very large in magnitude.

Latent Structure 2 of Table 2 was originally designed to simulate a very homogeneous test such

[3]The most recent version of LOGIST (Wingersky, in press) has options that may reduce this problem.

as a test of verbal ability that uses several types of items. From Tables 3 and 4 it is clear that LOGIST is robust to the minor violations of multidimensionality seen in Data Sets 2 and 6.

The third latent structure in Table 3 was designed to simulate more heterogeneous measurement, such as an instrument measuring overall job satisfaction or algebra achievement. Here about 70% of the variance in $\hat{\theta}$ is due to the general factor and only about 10% is due to the strongest group factor. Moreover, comparing RMSDs of a for data sets based on the third latent structure to baseline RMSDs of a shows that LOGIST still recovers the item discrimination parameter implied by the general factor. Consequently, it appears reasonable to conclude that use of LOGIST is justified in item pools with multidimensionality of the type seen in Data Sets 3 and 8. The common factors in Equation 1 that underlie Data Sets 3 and 8 item responses have correlations from .46 to .60. This means that these simulated item pools are quite heterogeneous. Note that factor analyzing the *dichotomous* item responses of Data Sets 3 and 8 may yield factors with intercorrelations smaller than .46 to .60 and that guessing may further decrease such factor correlations.

Use of LOGIST in Data Sets 4 and 9 leads to parameter estimates with interpretations that are ambiguous at best. In Data Sets 5 and 10, LOGIST is clearly drawn to a group factor. In sum, it appears that LOGIST should not be used in data sets with the degree of multidimensionality seen in the fourth and fifth latent structures of Table 2.

The results obtained here indicate that retaining the null hypothesis that an item pool is unidimensional when conducting a significance test, such as the ones developed by Christoffersson (1975) and Muthén (1978), is not a prerequisite for applications of IRT. A powerful significance test would always reject the null hypothesis of unidimensionality for large samples of actual examinees. Nonetheless, results for Data Sets 2, 3, 7, and 8 indicate that LOGIST parameter estimates will have many practically useful applications in multidimensional item pools.

One criticism of the present research is that unidimensional IRT models are improperly applied to multidimensional item pools; instead, it would be argued that multidimensional IRT models (see Reckase & McKinley, in press; Sympson, 1978) should be used for multidimensional item pools. If a single dominant latent trait is not sufficiently prepotent, the results presented here clearly show that a unidimensional model is inadequate. However, it is important to note that unidimensional models *do* provide a good description of multidimensional data sets when the dominant latent trait is sufficiently prepotent. Moreover, it appears that robustness studies of the type described here will be necessary when workable estimation methods for multidimensional IRT models become available: A small number of interpretable latent traits will not span the entire latent trait space underlying item responses. This point is illustrated by the results of Christoffersson's (1975) significance tests (with $\alpha = .01$) for the number of common factors underlying a pool of 12 items. His significance tests showed that there were more than four common factors, of which only two factors were interpretable. Of course, more items could be written in an attempt to study these uninterpretable common factors, but it seems likely that further significance tests would then show that even more common factors were required and some of these additional factors would be uninterpretable. Furthermore, the additional factors seem unlikely to provide valuable contributions to substantive theory. Thus, in the present context it seems clear that researchers should be more concerned with the robustness of estimation techniques to minor violations of dimensionality assumptions than with the possibly never-ending task of measuring all latent variables that underlie responses in a particular content domain.

It is important for researchers to investigate the dimensionality of an item pool before applying IRT. One obvious indication that multidimensionality is too severe for LOGIST is relatively many items with $\hat{b}$ values that are excessively large. A second, more sophisticated, approach for examining the latent structure of dichotomous item responses with or without guessing and with possibly nonnormal ability distributions is currently under investigation (Drasgow & Lissak, in press).

It is important to note a number of limitations

on the results obtained here. First, item responses were simulated using exactly one second-order general factor. The effects of two or more second-order general factors on parameter estimates computed by LOGIST are unknown. A multidimensional IRT model may be essential to model adequately data sets with two or more second-order general factors. Second, the magnitudes of the factor loadings presented in Table 1 are comparable across the five first-order common factors. Presumably, relatively smaller loadings on one common factor would reduce the influence of its underlying second-order group factor on parameter estimates; however, the details of the relations between LOGIST's parameter estimates and magnitudes of loadings on common factors are not clear. Third, only one pattern of factor loadings with 15, 5, 10, 10, and 10 items per factor was examined. The effects of a wider range than the 5 to 15 items per factor and of different distributions than 5, 10, 10, 10, and 15 require further investigation. A fourth limitation lies in the use of the idealized factor loading matrix shown in Table 1. Actual items are likely to have small but nonzero loadings on several factors, and it is common to find that some items have large loadings on more than a single factor.

References

Birnbaum, A. Some latent trait models. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley, 1968.

Christoffersson, A. Factor analysis of dichotomized variables. *Psychometrika*, 1975, *40*, 5–32.

Drasgow, F., & Lissak, R. I. Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, in press.

Hulin, C. L., Lissak, R. I., & Drasgow, F. Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, 1982, *6*, 249–260.

IMSL Library 1 (5th ed.). Houston TX: International Mathematical and Statistical Libraries, 1975.

Lord, F. M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (Research Bulletin 75-33). Princeton NJ: Educational Testing Service, 1975.

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

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

Muthén, B. Contributions to factor analysis of dichotomous variables. *Psychometrika*, 1978, *43*, 551–560.

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

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

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

Swaminathan, H., & Gifford, J. A. *Estimation of parameters in the three parameter latent trait model* (Report No. 90). Amherst MA: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluation Research, 1979.

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

Thurstone, L. L. *Multiple-factor analysis*. Chicago: University of Chicago Press, 1947.

Wingersky, M. S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *ERIBC monograph on applications of item response theory*. Vancouver BC: Educational Research Institute of British Columbia, in press.

Wood, R. L., & Lord, F. M. *A user's guide to LOGIST* (Research Memorandum 76-4). Princeton NJ: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., & Lord, F. M. *LOGIST—A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum 76-6). Princeton NJ: Educational Testing Service, 1976.

Acknowledgments

The authors thank Charles L. Hulin for his comments on an earlier draft. The order of authorship was determined by a computer simulation of the toss of a fair coin.

Author's Address

Send requests for reprints or further information to Fritz Drasgow, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign IL 61820, U.S.A.