The Effects on Parameter Estimation of Correlated Dimensions and a Distribution-Restricted Trait in a Multidimensional Item Response Model

Rose-Marie Batley and Marvin W. Boss

The University of Ottawa

This study was designed to assess the effects on parameter estimation of correlated dimensions and a distribution-restricted trait on one dimension using a two-dimensional item response theory model. Multidimensional analysis of simulated two-dimensional item response data fitting a multidimensional two-parameter logistic item response theory model (McKinley & Reckase, 1983a; Reckase & McKinley, 1991) was done using the program MIRTE (Carlson, 1987). Six datasets (2 trait distributions × 3 levels of correlation between dimensions) of 2,000 trait vectors over 104 items were generated. Each dataset was analyzed and replicated 100 times. Trait and item parameters generally were recovered adequately in the datasets in which both traits were normally distributed over the full range. In the datasets with a restricted range of trait level on the second dimension, recovery of the trait and item parameters was affected adversely. The results indicated that MIRTE recovers the structure of a multidimensional correlated space better than reported in earlier studies, especially when items are multidimensional. *Index terms: correlated traits, multidimensional item parameter estimates, multidimensional item response theory, multidimensional trait estimates, restricted traits.*

Early item response theory (IRT) models were based on the assumption of unidimensionality (i.e., only one trait is required to respond correctly to all items). When more than one trait accounts for test performance, the test is multidimensional and a multidimensional IRT (MIRT) model is appropriate.

Several researchers (Ackerman, 1987; Ansley & Forsyth, 1985; Bogan & Yen, 1983; Dorans & Kingston, 1985; Drasgow & Parsons, 1983; McCauley & Mendoza, 1985; McKinley & Reckase, 1984; Reckase, 1979, 1985; Reckase, Carlson, Ackerman, & Spray, 1986) have analyzed known multidimensional data with a unidimensional item response model. Unidimensional trait estimates have been found to be related to the average of the multidimensional traits (Ansley & Forsyth, 1985); unidimensional trait estimates also have been found to have different interpretations at different points on the unidimensional trait scale (Reckase et al., 1986). In general, unidimensional estimates from multidimensional data have been difficult to interpret and have not reflected well the original characteristics of the data.

Researchers who have used multidimensional IRT analysis (e.g., McKinley, 1983; McKinley & Reckase, 1983a, 1983b, 1984; Muraki & Englehard, 1985) have indicated that a multidimensional model more adequately describes both real and simulated multidimensional data than does a unidimensional model. However, because most of the simulation studies have not used replications, the stability of the estimates is difficult to determine. It is necessary to know how consistently these estimates are recovered.

Consider the situation in which items for a test are designed to measure one trait (e.g., mathematics ability) but require some amount of a second trait (e.g., verbal ability) for the examinee to respond correctly. This second required trait could be more crucial to success for some

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 2, June 1993, pp. 131-141

0146-6216/93/020131-11$1.80

examinees than others. Students of English as a second language (ESL) may have sufficient mathematics ability but lack the required amount of verbal ability to respond correctly.

It is reasonable to assume that the two traits are correlated. Therefore, it is appropriate to evaluate the effects of both correlated traits and a secondary trait, which is restricted in its distribution, on parameter estimation.

Purpose

This study was designed to determine the adequacy of multidimensional trait and item parameter estimates using a MIRT analysis. Three questions were addressed:

1. What is the effect of correlated trait dimensions on parameter estimation for a two-parameter, two-dimensional IRT model when both traits are normally distributed?
2. How is parameter estimation affected by a restricted secondary trait?
3. Are the effects of correlated dimensions similar when both traits are normally distributed, and when a restricted trait exists on the secondary dimension?

Method

Model Description

The data were generated to fit a multidimensional two-parameter logistic (M2PL) IRT model (McKinley & Reckase, 1983a; Reckase & McKinley, 1991). The M2PL model is defined as

$$P_{ij} = P(X_{ij} = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}_j) = \frac{\exp(\mathbf{a}_i\boldsymbol{\theta}_j + d_i)}{1 + \exp(\mathbf{a}_i\boldsymbol{\theta}_j + d_i)}, \quad (1)$$

$(i = 1, 2, \ldots, n; j = 1, 2, \ldots, N)$

where

P_{ij} is the probability of a correct response to item i by examinee j;

X_{ij} is the response (1 = correct; 0 = incorrect) of examinee j to item i;

$\mathbf{a}_i$ is a vector of m discrimination parameters (where m is the number of dimensions);

d_i is a parameter representing the difficulty of item i;

$\boldsymbol{\theta}_j$ is a vector of m trait parameters for individual j;

N is the number of examinees; and

n is the number of items.

This model is compensatory—it allows high levels on one dimension to compensate for low levels on other dimensions in arriving at a correct (or keyed) response to an item.

Reckase & McKinley (1991) defined a multidimensional discrimination parameter for item i as:

$$A_i = \left[\sum_{k=1}^{m} (a_{ik})^2\right]^{1/2} . \quad (2)$$

[Reckase & McKinley (1991) originally defined multidimensional discrimination as MDISC, rather than A.] A is proportional to the slope of the item response function at the point of steepest slope and it is, therefore, analogous to the unidimensional discrimination parameter (Carlson, 1987).

Reckase (1985) and Reckase & McKinley (1991) also defined a multidimensional item difficulty parameter, D_i, as:

$$D_i = -d_i/A_i . \quad (3)$$

[Reckase (1985) and Reckase & McKinley (1991) defined multidimensional item difficulty as MDIF, rather than D.] D represents the distance between the origin of the m-dimensional trait space and the point in the space at which the item information is a maximum. The line joining this point to the origin is at an angle of α_{ik} to the kth trait dimension where

$$\cos \alpha_{ik} = a_{ik}/A_i . \quad (4)$$

Data

Tests and responses. Six datasets were used. Datasets 1, 2, and 3 represented cases in which both underlying traits (θ_1 and θ_2) were normally distributed with mean 0 and standard deviation (SD) 1. These datasets differed in the degree of correlation between θ_1 and θ_2, which was 0.00, .25, and .50, respectively. In

Datasets 4, 5, and 6, θ_1 was normally distributed (mean = 0, SD = 1), but θ_2 was normally distributed with a lower mean and SD (-1 and .67, respectively). These datasets also varied in the degree of correlation (0.00, .25, and .50, respectively).

The simulated test contained 104 items—26 items required only θ_1 (i.e, only the first dimension had a nonzero discrimination), 52 items required predominantly θ_1, and 26 items required approximately equal amounts of θ_1 and θ_2. The item parameters are provided in Table 1. To meet the requirement that the items discriminate well on θ_1, four values of the angle, α_{i1}, were selected—0°, 15°, 30°, and 45°. 13 values of D (ranging from -3 to +3 at intervals of .5) were selected to cover the range of difficulties; two values of A (2.00, 1.70) were selected to simulate realistic discrimination conditions in which the items were designed to discriminate well on θ_1. a_1 and a_2 (one for each dimension) then were generated to fit the corresponding d_i and A. 104 items may be considered a long test, but it was necessary in order to have an item at each of the levels of α_{i1}, D, and A. The correlations between these item parameters were: $\rho(d, a_1) = .004$; $\rho(d, a_2) = -.004$; $\rho(a_1, a_2) = -.738$; and $\rho(D, A) = -.002$. Because of the dependency of a_1 and a_2, there was a larger correlation between a_1 and a_2. The same item parameters were used for each of the six datasets.

Analysis. Program M2PLGEN (Ackerman, 1985) was used to generate 2,000 θ vectors (θ_1, θ_2) satisfying the specified distributions of θ_1 and θ_2 for each dataset. M2PLGEN uses a random seed and the International Mathematical and Statistical Libraries (1979) subroutine GGNSM to generate random θ levels. These θ vectors and the item parameters (a_1, a_2, d) then were used to generate response vectors (0s and 1s) to each item for the 2,000 simulees according to the M2PL model. At the end of each replication, the random seed used by M2PLGEN to generate θ response vectors was increased by 2 and the process was repeated.

Program MIRTE Version 2.0 (Carlson, 1987) was used to analyze the two-dimensional data. MIRTE estimated θs, as, and ds for the M2PL model, estimates of the standard errors (SEs) for each of these parameter estimates, and estimates of A and D. The estimation method used is a variation of the joint maximum likelihood procedure using a modified Newton-Raphson iteration technique. The algorithm used is similar to that used in the unidimensional analysis program LOGIST (Wingersky, Barton, & Lord, 1982).

For each replication, the 2,000 × 104 matrix of response vectors was analyzed using MIRTE. The initial item parameter estimates for a_1 and a_2 were randomly generated according to the requirements of MIRTE: MIRTE requires that they be any positive number less than the maximum value defined by the user for the discrimination parameters, in this case 2.0. The same estimates were used for every replication and in every dataset to provide better control. Initial difficulty parameter estimates were computed by MIRTE. Means and SDs of $\hat{\theta}_1$, $\hat{\theta}_2$, the number-correct score, and item parameter estimates were calculated for each replication. For each parameter estimate, the average absolute deviation (AAD)—the average absolute deviation of the estimate from the true value of the variable—also was calculated.

Means of the SEs of $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{d}$, $\hat{a}_1$, and $\hat{a}_2$ also were determined, and correlations between parameters and estimates were calculated. Mean statistics then were calculated over the 100 replications for each of the six datasets.

Results

Adequacy of the Simulated Data

To assess the effects on parameter estimation of correlated θs and a restricted θ on one dimension, it was necessary to determine if suitable θs were generated to model the conditions specified. It was also important to determine whether MIRTE adequately estimated the parameters of the response vectors generated.

For Datasets 1–3, the θ vectors were both normally distributed over a full range. The correlation between θ_1 and θ_2 for data generated over

Table 1
True Item Parameters for the 104 Items

α_{i1}	D_i	d_i	Item Number	A	a_{i1}	a_{i2}	Item Number	A	a_{i1}	a_{i2}
0°	3.0	−6	1	2.00	2.00	0.00	53	1.70	1.70	0.00
0°	2.5	−5	2	2.00	2.00	0.00	54	1.70	1.70	0.00
0°	2.0	−4	3	2.00	2.00	0.00	55	1.70	1.70	0.00
0°	1.5	−3	4	2.00	2.00	0.00	56	1.70	1.70	0.00
0°	1.0	−2	5	2.00	2.00	0.00	57	1.70	1.70	0.00
0°	.5	−1	6	2.00	2.00	0.00	58	1.70	1.70	0.00
0°	0.0	0	7	2.00	2.00	0.00	59	1.70	1.70	0.00
0°	−.5	1	8	2.00	2.00	0.00	60	1.70	1.70	0.00
0°	−1.0	2	9	2.00	2.00	0.00	61	1.70	1.70	0.00
0°	−1.5	3	10	2.00	2.00	0.00	62	1.70	1.70	0.00
0°	−2.0	4	11	2.00	2.00	0.00	63	1.70	1.70	0.00
0°	−2.5	5	12	2.00	2.00	0.00	64	1.70	1.70	0.00
0°	−3.0	6	13	2.00	2.00	0.00	65	1.70	1.70	0.00
15°	3.0	−6	14	2.00	1.932	.518	66	1.70	1.642	.44
15°	2.5	−5	15	2.00	1.932	.518	67	1.70	1.642	.44
15°	2.0	−4	16	2.00	1.932	.518	68	1.70	1.642	.44
15°	1.5	−3	17	2.00	1.932	.518	69	1.70	1.642	.44
15°	1.0	−2	18	2.00	1.932	.518	70	1.70	1.642	.44
15°	.5	−1	19	2.00	1.932	.518	71	1.70	1.642	.44
15°	0.0	0	20	2.00	1.932	.518	72	1.70	1.642	.44
15°	−.5	1	21	2.00	1.932	.518	73	1.70	1.642	.44
15°	−1.0	2	22	2.00	1.932	.518	74	1.70	1.642	.44
15°	−1.5	3	23	2.00	1.932	.518	75	1.70	1.642	.44
15°	−2.0	4	24	2.00	1.932	.518	76	1.70	1.642	.44
15°	−2.5	5	25	2.00	1.932	.518	77	1.70	1.642	.44
15°	−3.0	6	26	2.00	1.932	.518	78	1.70	1.642	.44
30°	3.0	−6	27	2.00	1.732	1.00	79	1.70	1.472	.85
30°	2.5	−5	28	2.00	1.732	1.00	80	1.70	1.472	.85
30°	2.0	−4	29	2.00	1.732	1.00	81	1.70	1.472	.85
30°	1.5	−3	30	2.00	1.732	1.00	82	1.70	1.472	.85
30°	1.0	−2	31	2.00	1.732	1.00	83	1.70	1.472	.85
30°	.5	−1	32	2.00	1.732	1.00	84	1.70	1.472	.85
30°	0.0	0	33	2.00	1.732	1.00	85	1.70	1.472	.85
30°	−.5	1	34	2.00	1.732	1.00	86	1.70	1.472	.85
30°	−1.0	2	35	2.00	1.732	1.00	87	1.70	1.472	.85
30°	−1.5	3	36	2.00	1.732	1.00	88	1.70	1.472	.85
30°	−2.0	4	37	2.00	1.732	1.00	89	1.70	1.472	.85
30°	−2.5	5	38	2.00	1.732	1.00	90	1.70	1.472	.85
30°	−3.0	6	39	2.00	1.732	1.00	91	1.70	1.472	.85
45°	3.0	−6	40	2.00	1.414	1.414	92	1.70	1.202	1.202
45°	2.5	−5	41	2.00	1.414	1.414	93	1.70	1.202	1.202
45°	2.0	−4	42	2.00	1.414	1.414	94	1.70	1.202	1.202
45°	1.5	−3	43	2.00	1.414	1.414	95	1.70	1.202	1.202
45°	1.0	−2	44	2.00	1.414	1.414	96	1.70	1.202	1.202
45°	.5	−1	45	2.00	1.414	1.414	97	1.70	1.202	1.202
45°	0.0	0	46	2.00	1.414	1.414	98	1.70	1.202	1.202
45°	−.5	1	47	2.00	1.414	1.414	99	1.70	1.202	1.202
45°	−1.0	2	48	2.00	1.414	1.414	100	1.70	1.202	1.202

continued on the next page

Table 1, continued
True Item Parameters for the 104 Items

α_{i1}	D_i	d_i	Item Number	A	a_{i1}	a_{i2}	Item Number	A	a_{i1}	a_{i2}
45°	-1.5	3	49	2.00	1.414	1.414	101	1.70	1.202	1.202
45°	-2.0	4	50	2.00	1.414	1.414	102	1.70	1.202	1.202
45°	-2.5	5	51	2.00	1.414	1.414	103	1.70	1.202	1.202
45°	-3.0	6	52	2.00	1.414	1.414	104	1.70	1.202	1.202

the 100 replications was .001 for Dataset 1, .251 for Dataset 2, and .500 for Dataset 3. The means for θ_1 and θ_2 ranged from .002 to -.004, and SDs were within 1 ± .003. For Datasets 4–6, the correlations were equally well generated and similar results to Datasets 1–3 were found for θ_1 (.001 for Dataset 4, .251 for Dataset 5, and .499 for Dataset 6). For θ_2, the means ranged from -.999 to -1.001. The variance for the means and SDs for both θ_1 and θ_2 over the 100 replications was very small (less than .0005) in all datasets.

In agreement with the findings of Greaud (1988), the mean number of items correct appeared to be unaffected by changes in the degree of correlation between the θ dimensions. This may be because a compensatory model was used. For Datasets 1–3, all number-correct score means were approximately 52. As expected, the means for Datasets 4–6 were lower (approximately 47) because of the lower mean θ_2 values.

For all datasets over the 100 replications, $\hat{\theta}_1$ and $\hat{\theta}_2$ had means of 0.0 and SDs of 1.0. This is a function of MIRTE—it rescales the θ estimates to N(0,1) after each iteration. This presented problems for the estimation of item parameters for Datasets 4–6 because θ_2 was generated as N(-1.0, .67).

MIRTE did not always identify Dimensions 1 and 2 correctly in the estimated results. To avoid incorrect identifications of dimensions during the 100 replications, the first 13 item discrimination parameter estimates for each replication were analyzed, because they were pure on θ_1. If the sum of these a_1 estimates was less than the sum of the first 13 a_2 estimates, the âs and θ̂s for the estimates for the two dimensions were reversed (i.e., estimates identified as Dimension 1 were labeled as Dimension 2 and vice versa).

Recovery of θ

In Datasets 1–3, the degree of correlation did not seem to affect the recovery of θ_1 (see Table 2). This is shown by the mean AADs. θ_2 was better estimated (lower AADs) as correlations between θs increased.

For Datasets 4–6, both θ_1 and θ_2 were better estimated when θs were not correlated. The large AADs for $\hat{\theta}_2$ are a result of the rescaling of $\hat{\theta}_2$ to a mean of 0 and a SD of 1. The AADs were higher in Datasets 4–6 than Datasets 1–3. In each instance, the mean AADs were consistent over replications. The mean SEs of the θ̂s were slightly higher for Datasets 4–6 (.287) than for Datasets 1–3 (.259).

Table 2 also shows the correlations of the θs with their estimates, averaged over the 100 replications. For Datasets 3 and 6 (in which the θs were correlated .5), there was poorer recovery of θ_1 (.831 and .744) and better recovery of θ_2 (.865 and .721). When the θs were not correlated (Datasets 1 and 4), the correlations between $\hat{\theta}$ and θ were considerably lower in Dataset 4 for both θ_1 (.842 in Dataset 1 vs. .773 in Dataset 4) and θ_2 (.764 in Dataset 1 vs. .517 in Dataset 4).

The correlation between θ_1 and θ_2 was not well recovered. For Datasets 1 and 4, $\rho(\theta_1,\theta_2)$ was overestimated (r = .062 and .147, respectively). For the remaining datasets, $\rho(\theta_1,\theta_2)$ was underestimated (Table 2). These results agree with those reported by Carlson (1987).

It is interesting to note the high correlation between θ_1 and $\hat{\theta}_2$. For Datasets 4–6, these correlations were higher that those of θ_2 with $\hat{\theta}_2$. Clearly, higher correlations between dimensions

Table 2
Means Over 100 Replications of Statistics for Estimated θ

Dataset	$\rho(\theta_1,\theta_2)$	AAD($\hat{\theta}_1$)	AAD($\hat{\theta}_2$)	$r(\hat{\theta}_1,\hat{\theta}_2)$	$r(\theta_1,\hat{\theta}_1)$	$r(\theta_2,\hat{\theta}_2)$	$r(\theta_1,\hat{\theta}_2)$	$r(\theta_2,\hat{\theta}_1)$
1	0.00	.447	.544	.062	.842	.764	.505	-.295
2	.25	.446	.470	.179	.842	.824	.603	-.050
3	.50	.459	.412	.282	.831	.865	.699	.209
4	0.00	.463	.856	.147	.773	.517	.662	-.170
5	.25	.544	1.079	.201	.765	.623	.713	.052
6	.50	.566	1.047	.218	.744	.721	.755	.247

and restriction of range resulted in $\hat{\theta}_2$ becoming more like θ_1. It is also possible that the attempt to distinguish dimensions based on $\hat{a}$s for the items pure on θ_1 did not align the true and obtained dimensions.

In summary, a correlation between the θ dimensions seemed to result in slightly poorer estimation for θ_1 in all datasets. For θ_2, a correlation between dimensions resulted in better estimation for all datasets. A correlation between the θ dimensions and restriction of range resulted in $\hat{\theta}_2$ becoming more like θ_1. It seemed to be more difficult for the program to distinguish between the dimensions and there was a greater tendency to collapse the space.

Item Parameter Estimates

The maximum likelihood estimation procedures in MIRTE use θ estimates to improve item parameter estimates and vice versa. Hence, the final estimates are affected by each other. Thus, it was of interest to examine the effects of increases in $\rho(\theta_1,\theta_2)$ and restriction of range for θ_2 on the item parameter estimates.

Statistics on the item difficulty parameters are summarized in Table 3. In Datasets 1–3, $r(d,\hat{d}) = .997$, which indicated good recovery of d. As $\rho(\theta_1,\theta_2)$ increased, the mean and SD of d were increasingly overestimated but remained close to the original parameters. In Datasets 1–3, AAD($\hat{d}$) increased slightly as the correlation between θ dimensions increased, which indicated that d was being less well recovered. However, the SE of d decreased as $\rho(\theta_1,\theta_2)$ increased. The mean and SD of D were recovered well, although again D was less well recovered as $\rho(\theta_1,\theta_2)$ increased. D is a function of the discrimination parameters, and its estimate is therefore affected by the estimates of the a parameters.

Both d and D were less well recovered in Datasets 4–6 than in Datasets 1–3. The rescaling of θ_2 to mean 0 and SD 1 resulted in larger estimates for θ_2, which suggested that the simulees in Datasets 4–6 were more able than they actually were. In these datasets, there were larger SEs for $\hat{d}$ and larger SDs for $\hat{d}$ and $\hat{D}$ than in the other datasets. As $\rho(\theta_1,\theta_2)$ increased, SE($\hat{d}$) decreased but AAD($\hat{d}$) increased. However, the recovery of the relationships between the parameters and their estimates remained high—$r(d,\hat{d}) > .98$ and $r(D,\hat{D}) > .95$—and there was little change in these correlations as $\rho(\theta_1,\theta_2)$ increased. It is

Table 3
Means Over 100 Replications of Statistics for Item Difficulty

Dataset	$\hat{d}$	SD($\hat{d}$)	SE($\hat{d}$)	AAD($\hat{d}$)	$\hat{D}$	SD($\hat{D}$)	$r(d,\hat{d})$	$r(D,\hat{D})$
True	.009	3.771			-.005	2.058		
1	.009	3.929	.112	.224	.006	2.079	.997	.995
2	.010	3.936	.109	.228	.005	2.028	.997	.994
3	.030	.936	.106	.232	.012	1.999	.997	.991
4	-.726	3.995	.137	.811	.460	2.535	.984	.958
5	-.734	4.001	.132	.827	.434	2.459	.982	.956
6	-.716	4.044	.123	.834	.397	2.247	.982	.969

interesting to note the high correlation of D with $\hat{D}$ because D is a function of A (which was not well recovered) and d. Obviously the range of d (−6 to +6) affected estimates of D much more than did the two values (2.00 and 1.70) of the A parameters.

Discrimination parameter estimates were less well recovered than difficulty parameter estimates. The mean of $\hat{a}_1$ was lower than the true mean, and its SD was higher than the true SD for all six datasets (see Table 4). The means of $\hat{a}_2$ were much higher than the true mean of .678. In fact, the mean of $\hat{a}_2$ was higher than the mean estimates of a_1 and approached the true mean of a_1 as $\rho(\theta_1, \theta_2)$ increased. The means of $\hat{a}_2$ and $\hat{a}_1$ both increased slightly as $\rho(\theta_1, \theta_2)$ increased. The SD of $\hat{a}_2$ was higher than the true SD but there was not as large a difference as with $\hat{a}_1$. SEs of estimation of $\hat{a}_1$ and $\hat{a}_2$ were approximately .09 for Datasets 1–3, but were somewhat higher for Datasets 4–6.

For Datasets 4–6, the SEs for $\hat{a}_2$ were smaller than for $\hat{a}_1$. The SEs for $\hat{a}_2$ tended to decrease as $\rho(\theta_1, \theta_2)$ increased, but the SEs for $\hat{a}_1$ increased. For Datasets 1–3, the AAD($\hat{a}_2$) was much larger than AAD($\hat{a}_1$). For Datasets 4–6 it was only slightly larger.

A was recovered with a higher mean and higher SD in Datasets 1–3. Over the 100 replications, the mean $\hat{A}$ was closer to the true value for Datasets 4–6; however, within replications it was much more variable than for Datasets 1–3.

There appeared to be a rotational indeterminacy in the recovery of each of the a parameters and a tendency to spread the a parameter estimates over the entire space even though they originally did not cover the entire space. This was supported by the estimates of the angles α_1 and α_2 (the complement of α_1). α_1 had a mean of 22.50°. This mean was recovered in all datasets at over 49°. Similarly, the mean of α_2 was 67.50°, and it was recovered in all datasets at just over 40°. The original SD of 16.85° increased for the estimates to approximately 20°. Estimates of α_1 and α_2 ranged from very close to 0° to almost 90°. For all parameters relating to discrimination (a_1, a_2, A, α_1, α_2), their estimates seemed to cover the entire θ_1,θ_2 space.

Table 5 shows parameter recovery of the discrimination parameters as correlations. In all cases, a_1 correlated more highly with $\hat{a}_1$ than with $\hat{a}_2$, and a_2 correlated more highly with $\hat{a}_2$ than a_1 did with $\hat{a}_2$. $\hat{a}_2$ correlated more highly with a_2 than a_1 did with $\hat{a}_1$ (except for Dataset 4). a_2 correlated more highly with $\hat{a}_2$ than it did with $\hat{a}_1$ (except for Dataset 4). The anomaly in the correlations was the similarity of the relationship of a_1 and a_2 with $\hat{a}_1$. The correlation between A and its estimate was low (<.6) in Datasets 1–3 and extremely low (<.3) in Datasets 4–6. The high correlation between α_1 and its estimate for Datasets 1–3 was not expected after obtaining such poor estimates of the multivariate discrimination parameters. $r(\alpha_1, \hat{\alpha}_1)$ was much smaller in Datasets 4–6.

Restriction of range of the second trait, θ_2, drastically affected the recovery of the discrimination parameters (a_1, a_2, and A) in Datasets 4–6. For all three variables, the squared correlations between the parameter and its estimate indicate that they had less than half as much variance

Table 4
Means Over 100 Replications of Statistics for Item Discrimination

Dataset	$\hat{a}_1$	SD($\hat{a}_1$)	SE($\hat{a}_1$)	AAD($\hat{a}_1$)	$\hat{a}_2$	SD($\hat{a}_2$)	SE($\hat{a}_2$)	AAD($\hat{a}_2$)	$\hat{A}$	SD($\hat{A}$)	$\hat{\alpha}_1$
True	1.637	.251			.678	.496			1.850	.151	22.50°
1	1.195	.569	.099	.500	1.379	.512	.096	.707	1.957	.288	49.07°
2	1.201	.528	.095	.486	1.448	.582	.094	.775	2.013	.319	49.40°
3	1.202	.502	.093	.490	1.510	.628	.093	.836	2.057	.381	49.98°
4	1.076	.551	.119	.623	1.228	.449	.112	.653	1.736	.398	49.09°
5	1.094	.557	.138	.620	1.298	.501	.108	.708	1.803	.449	49.76°
6	1.139	.599	.134	.624	1.408	.534	.103	.791	1.922	.495	50.94°

Table 5
Mean Correlations (Over 100 Replications) for Item Discriminations

Dataset	$r(a_1, \hat{a}_1)$	$r(a_2, \hat{a}_2)$	$r(\hat{a}_1, \hat{a}_2)$	$r(a_1, \hat{a}_2)$	$r(a_2, \hat{a}_1)$	$r(A, \hat{A})$	$r(\alpha_1, \hat{\alpha}_1)$
True	–	–	-.738	–	–	–	–
1	.834	.893	-.765	-.572	-.865	.600	.943
2	.818	.899	-.769	-.587	-.830	.565	.933
3	.760	.895	-.735	-.587	-.747	.502	.907
4	.530	.523	-.428	-.309	-.543	.296	.630
5	.460	.511	-.401	-.306	-.455	.269	.586
6	.431	.514	-.459	-.285	-.403	.285	.564

in common for Datasets 4–6 than for Datasets 1–3.

Discussion and Implications

This study was designed to determine how well multidimensional IRT trait and item parameters would be estimated under different degrees of correlation between the two trait dimensions and the existence of a restricted trait on the second dimension. The best estimates of θ_1 and the poorest estimates of θ_2 occurred at 0 correlation between traits. This relationship was shown in two ways—the size of the AADs and the correlation of each $\hat{\theta}$ with its parameter. Except for AAD($\hat{\theta}_2$), the same pattern was found for Datasets 4–6 in which θ_2 had a restricted range.

The θ_2 estimates were affected by the rescaling of the estimates to a 0,1 distribution at each iteration. Evidence of this is that values of AAD($\hat{\theta}_2$) were almost double those in the restricted datasets. With a restricted trait on the second dimension, traits were not as well estimated. As the correlation between dimensions increased, $\hat{\theta}_2$ became more highly correlated with θ_1, and with restriction of range these correlations exceeded those of $\hat{\theta}_2$ with its parameter. Thus, it seems that as traits become more highly correlated and the range of θ is restricted there is a tendency for the space to collapse and become more unidimensional.

The difficulty parameters were well recovered in all datasets. SEs of $\hat{d}$ were small, and correlations of both d and D with their estimates were quite high ($r \geq .956$). The AADs for the restricted trait datasets (Datasets 4–6) were quite large, which resulted from forcing θs to have a 0,1 distribution. Thus, the data indicate that recovery of multidimensional difficulties was not greatly affected by correlated traits.

The recovery of discrimination parameters was not good. There seemed to be a tendency for θ_2 to become the dominant dimension. This was evidenced by the mean $\hat{a}$s and the mean $\hat{\alpha}$s. The variance of a_2 in the original items was almost four times as great as that for a_1. This seemed to be a likely explanation of the dominance of θ_2. The estimates also seemed to be scattered throughout the space.

Carlson (1987) reported that estimates of a parameters are sensitive to their distribution in the generated data. In the present study, the a_i parameters were not distributed over the entire latent space. In addition, the variance of the a_2 parameters was almost four times as large as for the a_1 parameters. This restricted the recovery of the a parameters which, in turn, affected the recovery of the θ and d parameters. It is also possible that MIRTE might not function as well as previous pilot testing had suggested.

Some Problems

Rescaling of the θ_2 estimates. The rescaling of the θ_2 estimates in the restricted datasets seemed to affect estimates of difficulty as well as estimates of θs and discriminations. Perhaps they should have been rescaled to the mean and SD of the true θ_2, but this would require having information that would not be known with nonsimulated data. Thus, operation of MIRTE was examined here as well as the concept of a restricted θ_2 trait. The $\hat{d}$ and $\hat{D}$ means were adversely affected in the restricted datasets. However,

correlations between d and $\hat{d}$, and D and $\hat{D}$ were high ($r > .95$). The estimates of the a_2 means improved in these datasets. From the results reported here, the extent of the effects of rescaling cannot be determined, but it appears that the rescaling problem affects all parameter estimates.

Recovery of the two-dimensional space. There was a tendency for the space to collapse as the traits became more correlated. Recovery of the structure of the trait space may relate to a rotational indeterminacy in the recovery of the traits. Correlations of θ_1 with $\hat{\theta}_1$ were moderately high ($r > .744$) for the six datasets. This would indicate that θ_1 had been recovered relatively well. However, as $\rho(\theta_1, \theta_2)$ increased, the correlation between θ_1 and $\hat{\theta}_2$ increased and, in Dataset 6, $r(\theta_1, \hat{\theta}_2) > r(\theta_1, \hat{\theta}_1)$ (.755 vs. 744). This higher correlation of θ_1 with $\hat{\theta}_2$ suggests that $\hat{\theta}_1$ was off by a rotational factor. This would, in part, explain the AAD for Dimension 1 and the modest correlation between θ_1 and $\hat{\theta}_1$. This correlation may underestimate the recovery of θ_1 in the two-dimensional space. The correlations of θ_2 and $\hat{\theta}_1$ were low, which indicated that θ_2 was recovered at approximately the level indicated by the correlation between θ_2 and $\hat{\theta}_2$. The rotational issue is important in considering whether a dimension is recovered poorly or matched poorly in the particular rotation. Dimension 2 was recovered modestly whereas Dimension 1 was recovered far better, but the rotational position of $\hat{\theta}_1$ masked this fact.

In the initial research design, some items pure on Dimension 2 were included to anchor the traits in an attempt to improve the recovery of all parameters. Because such a test would not simulate the desired conditions, these items were not included. This might be reconsidered in a future design.

The collapsing of the space as $\rho(\hat{\theta}_1, \theta_2)$ increased not only affected the θ estimates but also the discrimination estimates. In Datasets 4–6, the structure of the latent space was recovered less well than in the other datasets. In retrospect, combining corresponding datasets of the two types prior to analysis of the number-correct score vectors would provide a sample that might more typically represent the situation in which ESL students would likely be placed and would have allowed for better coverage of the θ_1, θ_2 space. This might improve the estimation of some parameters and it would also reduce the effects of the problems caused by the rescaling of $\hat{\theta}_2$.

Dimensionality of the item space. 26 of the items were unidimensional (pure on Dimension 1). The remaining 78 were two-dimensional—52 required primarily θ_1, and 26 required equal amounts of θ_1 and θ_2. The latent structure of the data was more complex than a two-dimensional test composed of two sets of unidimensional items. There were serious concerns with respect to the recovery of the item space—the most serious being the apparent dominance of $\hat{\theta}_2$ over $\hat{\theta}_1$, or $\hat{\alpha}_2$ over $\hat{\alpha}_1$. The poor recovery of the discrimination parameters also affected recovery of the difficulty and trait parameters. The item space seemed to become somewhat unidimensional. The estimates of the a_is were more alike and the size of the α angle estimates moved toward 45° with $\hat{\alpha}_2$ becoming dominant. Because the range of a_2 was greater than that of a_1, this could have affected the dominance of $\hat{a}_2$ over $\hat{a}_1$.

Interpretation of parameter estimates appears to depend on the model, $\rho(\theta_1, \theta_2)$, and the characteristics of the dataset. From the results presented here, there is every indication that there are indeed three components of multidimensionality (examinee dimensionality, test dimensionality, and the interaction of the two) as suggested by McKinley & Reckase (1984). Although the population may be multidimensional, if the test is largely unidimensional, resulting scores will tend to unidimensionality as well. It may be expecting too much of the model and MIRTE to have better recovery of the parameters relating to the second dimension when few items measured that dimension and when the populations in the restricted datasets were low on trait level in the second dimension. As a result, little information could be obtained on items of higher difficulty measuring the second dimension.

Future Directions

Several questions remain that suggest future studies. First, are the results affected by the estimation procedures and/or the model selected? Replication of the research using different models [perhaps the multidimensional three-parameter logistic IRT model of Bogan & Yen (1983) or a noncompensatory model] would indicate the extent to which the model affected results. Inclusion of a guessing parameter in the model also would provide additional information. A version of MIRTE (Carlson, 1987) allows for inclusion of the guessing parameter.

It also would be useful to estimate item parameters only while holding the given trait parameters fixed and vice versa to determine further the efficiency of MIRTE. These results could be compared with those obtained when item and trait parameters are estimated simultaneously. Presumably, both item and trait parameters would be better estimated when the other parameter is held fixed.

Corresponding restricted and nonrestricted datasets could be combined in a suitable ratio to present the ESL group as part of a large sample of non-ESL examinees. This would be more typical of real data. This should solve some of the rescaling and space problems.

The test design could be altered to allow for better distribution of the discrimination parameters. The discrimination and difficulty parameters also could be generated randomly to cover the space. The test would then not simulate the condition that it primarily measure one of the two dimensions. However, valuable information could be gained on parameter recovery.

It would be useful to determine the means of trait estimates at different trait levels rather than just reporting the mean level over all trait levels. This could be ascertained by examining the θ vectors in different sections of the θ_1,θ_2 space and comparing the original θs with θ estimates. It also would be useful to know how influential the second trait dimension becomes as the items require more of this trait for a correct response.

An important finding of this research is the capability of MIRTE to retain the structure of the data and the examinees. Although there was some tendency to collapse the latent space as $\rho(\theta_1, \theta_2)$ increased, estimates provided by MIRTE recovered two dimensions. It would be appropriate to further develop estimation programs so that rotational solutions could be produced that might alleviate the tendency to collapse a two-dimensional space as the correlation between the dimensions increases.

Conclusions

When a compensatory model is used, correlated dimensions affect recovery both of the dominant and less dominant dimensions. Restriction of range seems also to affect recovery. This seems likely even without a rescaling of the θs to 0,1. Of the parameter estimates, difficulty is least affected. The fact that the difficulty estimates are influenced by multiple trait and discrimination estimates may assure that difficulty is well estimated.

References

Ackerman, T. A. (1985). *M2PLGEN: A computer program for generating thetas and response strings corresponding to the M2PL model.* Iowa City IA: American College Testing.

Ackerman, T. A. (1987, April). *The use of unidimensional item parameter estimates of multidimensional items in adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, Washington D.C.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9,* 37–48.

Bogan, E. D., & Yen, W. M. (1983, April). *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, Canada.

Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program* (ACT Research Rep. No. 87-19). Iowa City IA: American College Testing Program.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estima-

tion of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22,* 249–262.

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

Greaud, V. A. (1988, April). *Some effects of applying unidimensional IRT to multidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

International Mathematical and Statistical Libraries. (1979). *IMSL Library* (7th ed.). Houston TX: Author.

McCauley, C. D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9,* 389–400.

McKinley, R. L. (1983, April). *A multidimensional extension of the two-parameter logistic latent trait model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada. (ERIC Document Reproduction Service No. ED 228 326)

McKinley, R. L., & Reckase, M. D. (1983a). *An application of a multidimensional extension of the two-parameter logistic latent trait model* (ONR-83-3). (ERIC Document Reproduction Service No. ED 240 168)

McKinley, R. L., & Reckase, M. D. (1983b). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation, 15,* 389–390.

McKinley, R. L., & Reckase, M. D. (1984). *An investigation of the effect of correlated abilities on observed test characteristics* (Research Report). Iowa City IA: American College Testing Program, Test Development Division. (ERIC Document Reproduction Service No. ED 249 249)

Muraki, E., & Englehard, G. (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9,* 417–430.

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. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986, June). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data.* Paper presented at the annual meeting of the Psychometric Society, Toronto, Ontario.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361–373.

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

Author's Address

Send requests for reprints or further information to Marvin W. Boss, Faculty of Education, University of Ottawa, 145 Jean-Jacques Lussier, Ottawa, Ontario K1N 6N5, Canada.