Unidimensional IRT Calibration of Compensatory and Noncompensatory Multidimensional Items

Terry A. Ackerman
American College Testing Program

The characteristics of unidimensional ability estimates obtained from data generated using multidimensional compensatory models were compared with estimates from noncompensatory IRT models. Reckase, Carlson, Ackerman, and Spray (1986) reported that when a compensatory model is used and item difficulty is confounded with dimensionality, the composition of the unidimensional ability estimates differs for different points along the unidimensional ability (θ) scale. Eight datasets (four compensatory, four noncompensatory) were generated for four different levels of correlated two-dimensional θs. In each dataset, difficulty was confounded with dimensionality and then calibrated using LOGIST and BILOG. The confounding of difficulty and dimensionality affected the BILOG calibration of response vectors using matched multidimensional item parameters more than it affected the LOGIST calibration. As the correlation between the generated two-dimensional θs increased, the response data became more unidimensional as shown in bivariate plots of the mean $\hat{\theta}_1$ as opposed to the mean of $\hat{\theta}_2$ for specified unidimensional quantiles. *Index terms: BILOG, compensatory IRT models, IRT ability estimation, LOGIST, multidimensional item response theory, noncompensatory IRT models.*

One of the underlying assumptions of unidimensional item response theory (IRT) models is that a person's ability can be estimated in a unidimensional latent space. However, some researchers and educators doubt that the response process to any one item requires only a single latent ability. Traub (1983), for example, suggested that many cognitive variables are brought to the testing task and that the number used varies from person to person. Likewise, the combination of latent abilities that individuals use to obtain a correct response may vary from item to item. Reservations about applying unidimensional IRT estimation to multidimensional response data have been expressed by Ansley and Forsyth (1985), Reckase, Carlson, Ackerman, and Spray (1986), and Yen (1984), among others.

Reckase (1985) calibrated item response data from a form of the American College Test (ACT), Assessment Mathematics Usage Test, using the two-dimensional IRT estimation program MAXLOG (McKinley & Reckase, 1983) and found that easy items in the beginning of the test measured primarily one dimension and the more difficult items at the end of the test measured mostly along a second dimension. Thus difficulty appeared to be confounded with dimensionality. Using an updated version of MAXLOG, called MIRTE (Carlson, 1987), these results have been verified for several forms of the math usage test.

Using a compensatory multidimensional IRT model, Reckase et al. (1986) demonstrated that when dimensionality and difficulty are confounded (i.e., easy items discriminate only on θ_1, difficult items discriminate only on θ_2), the unidimensional θ scale has a different meaning at different points on the scale. Specifically, for their two-dimensional dataset, upper deciles of the unidimensional θ continuum differed mainly on θ_2 while the lower deciles differed mostly on θ_1. Thus, Reckase et al. suggested that the univariate calibration of two-dimensional response data can be explained by the interaction between multidimensional test information and the distribution of the two-dimensional abilities. Reckase et al. examined the condition in

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 2, June 1989, pp. 113–127
© Copyright 1989 Applied Psychological Measurement Inc.
0146-6216/89/020113-15$2.00

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

which θ estimates were uncorrelated. Such an approach may not be very realistic, however, because most cognitive abilities tend to be correlated.

Ansley and Forsyth (1985) examined the unidimensional estimates from two-dimensional data generated using a noncompensatory model (Sympson, 1978). They selected item parameters so that generated response data would match item difficulty parameters as taken from a "real" test. They examined situations in which θs were correlated 0.0, .3, .6, .9, and .95. Although dimensionality might have been confounded with difficulty, the issue was not addressed. Ansley and Forsyth found that the $\hat{a}$ values were "best considered" as averages of the true a_1 and a_2 values, that the $\hat{b}$ values were "overestimates of b_1," and that $\hat{\theta}$s were "highly related" to the average of the true θ_1 and θ_2 values.

Way, Ansley, and Forsyth (1988) compared the effects of using a unidimensional IRT model to estimate two-dimensional data generated by both noncompensatory and compensatory multidimensional IRT models. Results for the noncompensatory datasets were similar to those of the Ansley and Forsyth (1985) study. For the generated compensatory datasets, the $\hat{a}$ values were "best considered as an estimate of the sum of a_1 and a_2"; $\hat{b}$ values were close to the "average of the b_1 and b_2 values"; and $\hat{\theta}$ was "highly related to the average" of the true θ_1 and θ_2 parameters.

This paper extends previous work by introducing new methodology to match compensatory and noncompensatory items, using simulated data based on multidimensional IRT calibration of real data using the compensatory model, and providing a graphical perspective from which the two models can be compared. Three issues are examined:

1. Do systematic differences exist between unidimensional calibrations of compensatory and noncompensatory items when difficulty is confounded with dimensionality? That is, would the results of the Reckase et al. (1986) study hold for both models?
2. Do different levels of correlation between two-dimensional abilities affect the confounding of difficulty and dimensionality under each model? It was hypothesized that as the correlation between θ_1 and θ_2 increased, the response data would become essentially unidimensional, thus reducing the effect of confounding difficulty and dimensionality.
3. Do systematic differences exist between unidimensional item parameters and θ estimates obtained from the item calibration programs LOGIST and BILOG when these programs are applied to either compensatory or noncompensatory multidimensional items?

Model Definitions

A multivariate logistic model introduced by Reckase (1985) was used to specify compensatory items. This model defines the probability of a correct response as

$$P(x_{ij} = 1|\mathbf{a}_i, d_i, \boldsymbol{\theta}_j) = \frac{1}{1+\exp\left(d_i - \sum_{k=1}^{n} a_{ik}\theta_{jk}\right)} \,, \qquad (1)$$

where x_{ij} is the response to item i by person j,
θ_{jk} is the ability parameter for person j on dimension k,
a_{ik} is the discrimination parameter for item i on dimension k, and
d_i is the difficulty parameter for item i.

The probability of a correct response for the noncompensatory logistic model introduced by Sympson (1978) is

$$P(x_{ij} = 1|\mathbf{a}_i, \mathbf{b}_i, \mathbf{c}_i, \boldsymbol{\theta}_j) = c_i + \frac{1-c_i}{\prod_{k=1}^{n}\{1+\exp[-1.7a_{ik}(\theta_{jk}-b_{ik})]\}} \,, \qquad (2)$$

where x_{ij}, θ_{jk}, and a_{ik} are defined as above,
b_{ik} is the difficulty parameter for item i on dimension k, and
c_i is the guessing parameter for item i.

Method

Data Generation

To test the effects of correlated θ dimensions, four levels of correlation were selected ($\rho_{\theta_1\theta_2}$ = 0.0, .3, .6, and .9). Parameters for a set of 40 two-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

dimensional compensatory items were selected with difficulty and dimensionality confounded. Discrimination parameters ranged from $a_1 = 1.8$, $a_2 = .2$ to $a_1 = .2$, $a_2 = 1.8$. Difficulty was confounded with dimensionality such that the difficulty parameters ranged from $d = -2.4$ (for $a_1 = 1.8$, $a_2 = .2$) to $d = 2.4$ (for $a_1 = .2$, $a_2 = 1.8$). Thus as the items became more difficult, they discriminated less along θ_2 and more along θ_1. This systematic shift in dimensionality with increasing difficulty was established to mimic the pattern found in real data by Reckase (1985). Although perhaps not very realistic, the guessing parameter was set to 0 because of concern over how much "noise" would be added to the multidimensional data with a nonzero guessing parameter.

Instead of the trial-and-error methodology used by Way et al. (1988) to match compensatory items to noncompensatory items, a new methodology was developed. For each compensatory item, a corresponding noncompensatory item (same probability of a correct response) was created using a least-squares approach to minimize the quantity

$$\sum_{j=1}^{1000} [(P_C|\boldsymbol{\theta}_j,\mathbf{a},d) - (P_{NC}|\boldsymbol{\theta}_j,\mathbf{a},\mathbf{b})]^2 \quad , \qquad (3)$$

where P_C is a given compensatory item's probability of correct response, and P_{NC} is the noncompensatory item's probability of correct response (which varies as a function of $\mathbf{a}$ and $\mathbf{b}$, given $\boldsymbol{\theta}_j$). The summation was taken over 1,000 simulated examinees sampled from a bivariate normal population with $\mu_1 = \mu_2 = 0$, $\sigma_{\theta_1} = \sigma_{\theta_2} = 1$, and $\rho_{\theta_1\theta_2}$ fixed at selected values. In all, four noncompensatory item sets were created with $\rho_{\theta_1\theta_2} = 0.0$, .3, .6, and .9. Different starting values were tried to ensure that the function did have unique local minima where P_C and $\rho_{\theta_1\theta_2}$ were fixed.

The mean, standard deviation, and minimum and maximum values of the difficulty and discrimination parameters for the compensatory and noncompensatory item sets are presented in Table 1. Because the difficulty parameters for the compensatory model were symmetrical (i.e., $d_1 = -d_{40}$, $d_2 = -d_{39}$, etc.), the mean difficulty is 0. In each correlational condition, the noncompensatory b_1 is greatest and positive for item 1 and decreases steadily as the item number increases. However, noncompensatory difficulties for Dimension 2 are negative for all items for all correlational conditions.

The a_1 parameters for each model are greatest for item 1 and decrease with item number. The a_1 parameters are greater for the noncompensatory model and decrease at a slower rate than their compensatory counterparts. The a_2 parameters for each model are lowest for the first item and increase with item number. The a_2 parameters are greater for the noncompensatory model and increase at a slower rate than their compensatory counterparts.

Eight response datasets were then produced. Using the compensatory item parameters, 1,000 response vectors were simulated for each of four correlational values ($\rho_{\theta_1\theta_2} = 0.0$, .3, .6, .9) from a bivariate normal distribution. For each set of noncompensatory item parameters, 1,000 response vectors were generated using the same (θ_1,θ_2) combinations that produced the compensatory response datasets.

Graphical Comparisons

The generated compensatory item set is represented graphically in an item vector plot (Reckase, 1985), shown in Figure 1. In this plot the vector for each item represents the distance and direction from the origin to the point of maximum slope or discrimination. The $p = .5$ equiprobability line runs orthogonal to the tip of the item vector. Thus, the longer a vector extends into the third quadrant, the easier the item, and the longer a vector extends into the first quadrant, the more difficult the item.

To clarify how the probability of a correct response changes as a function of $\boldsymbol{\theta}$ in each model, item response surfaces (IRSs) and corresponding contour plots for three pairs of matched items are presented in Figures 2 through 4. The IRS and its corresponding contour plot are shown for compensatory and noncompensatory items 1, 20, and 40. Little difference exists among the IRSs for the matched items when the items discriminate only along θ_2 (Figure 2) or only along θ_1 (Figure 4). However, when both items discriminate equally along θ_1 and θ_2 (Figure 3), the noncompensatory equiprobability curves contrast sharply with the

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1
Mean, Standard Deviation, Minimum and Maximum for Generating Compensatory and Noncompensatory Item Parameters

Parameter and Model	Mean	SD	Min	Max
Compensatory				
a_1	1.00	.48	.20	1.80
a_2	1.00	.48	.20	1.80
d_1	.00	1.43	−2.39	2.39
Noncompensatory: r = 0.0				
a_1	1.35	.34	.91	1.93
a_2	1.33	.32	.75	1.86
b_1	−1.28	1.93	−5.30	1.20
b_2	−1.29	.56	−3.22	−.85
Noncompensatory: r = .3				
a_1	1.47	.31	1.01	1.96
a_2	1.44	.33	.81	1.93
b_1	−1.08	1.67	−4.59	1.15
b_2	−1.11	.42	−2.58	−.75
Noncompensatory: r = .6				
a_1	1.59	.26	1.18	1.97
a_2	1.56	.33	.89	1.99
b_1	−.90	1.44	−3.86	1.13
b_2	−.94	.31	−1.98	−.63
Noncompensatory: r = .9				
a_1	1.68	.17	1.38	1.95
a_2	1.64	.28	1.08	1.99
b_1	−.73	1.26	−3.28	1.12
b_2	−.78	.26	−1.35	−.49

Note. A complete set of all item parameters can be obtained from the author on request.

parallel equiprobability lines of the compensatory item.

In the compensatory model, the $\theta_1\theta_2$ combinations with an equiprobability of correct response form parallel lines. In contrast, for the noncompensatory model the $\theta_1\theta_2$ combinations with equiprobability of correct response are curvilinear. Also, unlike the compensatory model, in which the direction of the equiprobability lines is a function of the discrimination parameters, the direction of the equiprobability lines for the noncompensatory model is a function of the difficulty parameters. Thus, no constraints were placed on the noncompensatory difficulty parameters so that the equiprobability curves could be "shifted" to match compensatory items which measure mostly θ_1 or θ_2.

Another way to compare the two matched item sets is to examine the amount of test information each provides over the $\theta_1\theta_2$ ability plane. Multidimensional test information plots ("INFLINE" plots; Reckase, 1985) for one set of matched items ($\rho_{\theta_1\theta_2}$ = 0.0) are shown in Figures 5a (compensatory item set) and 5b (noncompensatory item set). In each plot, the amount of test information at each of 49 selected points is illustrated by the length of the respective vectors for 10 different directions from 0° to 90° in 10° increments. The maximum amounts of test information for the compensatory and noncompensatory item sets with $\rho_{\theta_1\theta_2}$ = 0.0 were 18.85 and 16.88, respectively. For examinees with extremely high or extremely low θ on both dimensions, little information was provided for either

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

item set. In general, for each level studied, more information was provided by the set of compensatory items relative to the matched noncompensatory set of items. However, for each item set, in the two-dimensional **θ** plane some isolated regions existed where this was not true.

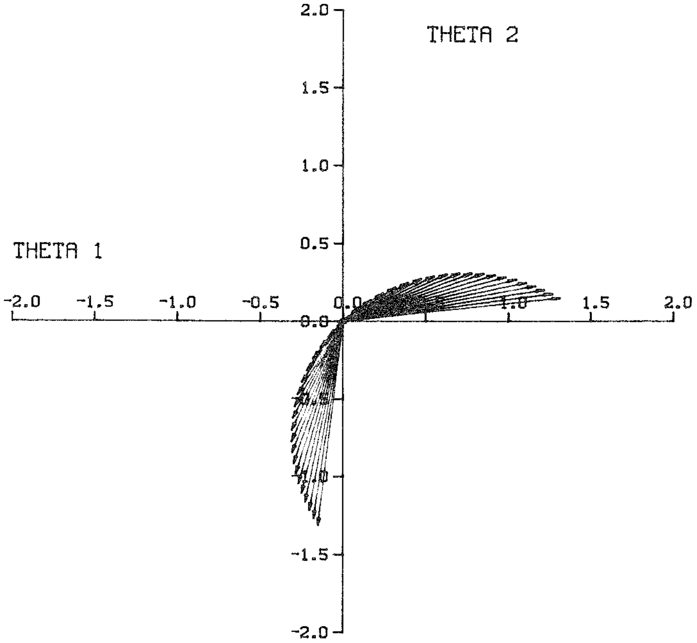
Ability and Item Parameter Estimation

Each dataset was then calibrated using both LOGIST (Wingersky, Barton, & Lord, 1982) and BILOG (Mislevy & Bock, 1982). A two-parameter logistic IRT model was calibrated by both programs. The IRT calibration programs use different estimation procedures. LOGIST uses bounded joint maximum likelihood (JML) estimation. For the calibrations in this study, $\hat{a}$ was constrained to ≤2 and the c parameter was not estimated. For all BILOG computer runs the default method of scoring examinees was selected. The default method of θ estimation was expectation a posteriori using a normal 0,1 Bayesian prior. The default priors were also used in the item parameter calibration: a log-normal prior on the discrimination estimates and no prior on the difficulty estimates (Mislevy & Stocking, 1989).

To estimate the LOGIST and BILOG orientation in the two-dimensional θ plane, the θ estimates from each calibration run were first rescaled to the compensatory θ estimates for the $\rho_{\theta_1\theta_2} = 0.0$ case. The $\hat{\theta}$ values for each calibration run were rank-ordered and divided into 20 quantiles. The means of the θ_1 and θ_2 parameters for each $\hat{\theta}$ quantile were then calculated and plotted. These "centroid" plots were then examined for curvilinearity, which would suggest that the composite (θ_1,θ_2) combination was not uniform across the univariate scale as predicted by the Reckase et al. (1986) study.

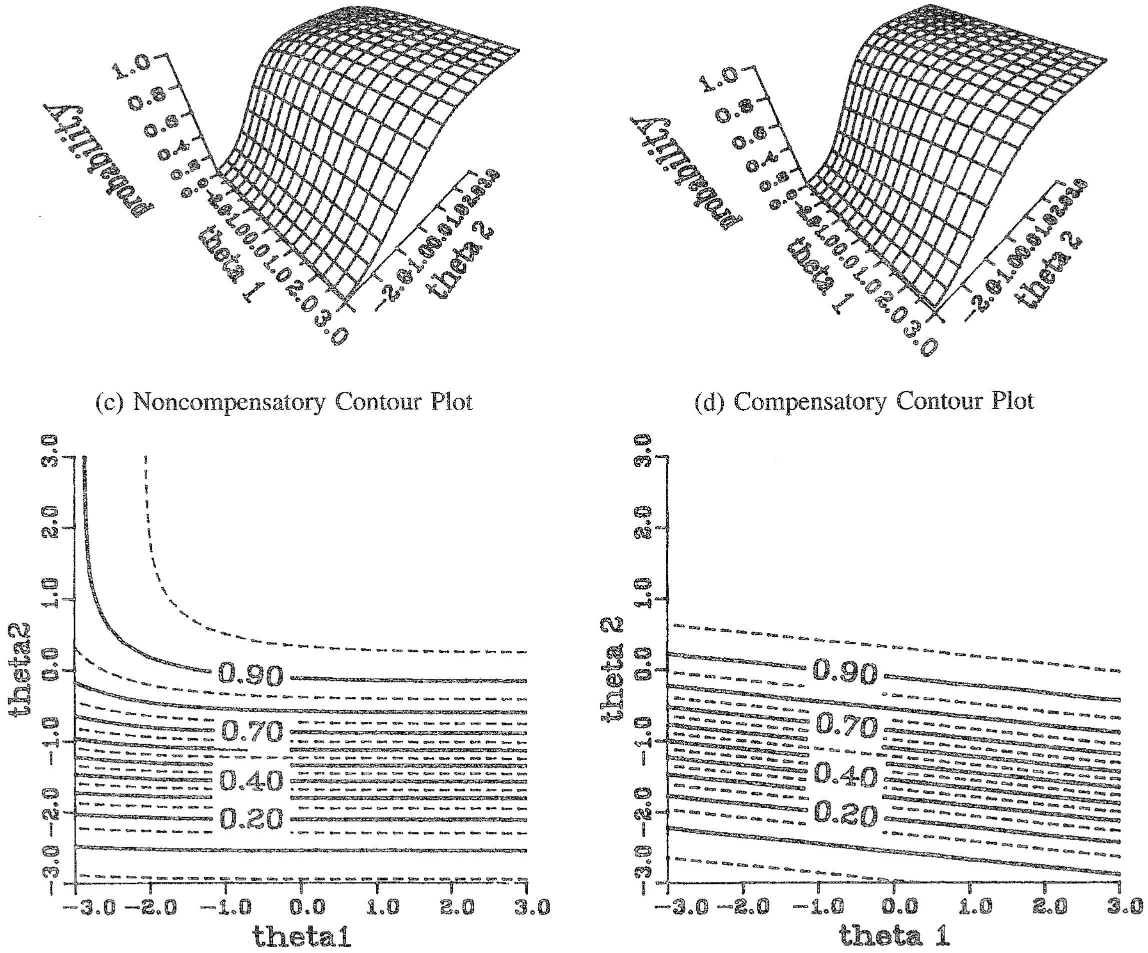
Results from the 16 sets of item calibrations were also evaluated by examining the correlations between $\hat{\theta}$ and θ_1 $(r_{\hat{\theta}\theta_1})$, between $\hat{\theta}$ and θ_2 $(r_{\hat{\theta}\theta_2})$, and

Figure 1
Vectors Representing the Distance and Direction From the Origin
to the Point of Maximum Discrimination for the 40 Generated Compensatory Items



Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Figure 2
The Item Response Surfaces and Contour Plots for Item 1



(a) Noncompensatory IRS
($a_1 = .91$, $a_2 = 1.85$, $b_1 = -5.29$, $b_2 = -1.35$)

(b) Compensatory IRS
($a_1 = .20$, $a_2 = 1.80$, $d = 2.39$)

(c) Noncompensatory Contour Plot

(d) Compensatory Contour Plot

between $\hat{\theta}$ and the average of θ_1,θ_2 ($r_{\theta,\theta_{ave}}$). Mean absolute differences (MADs) were computed between $\hat{\theta}$ and θ_1 $[(\sum|\theta_1 - \hat{\theta}|)/k]$, between $\hat{\theta}$ and θ_2 $[(\sum|\theta_2 - \hat{\theta}|)/k]$, and between $\hat{\theta}$ and the average of θ_1 and θ_2 $[(\sum|\theta_{ave} - \hat{\theta}|)/k]$ for the k simulated examinees.

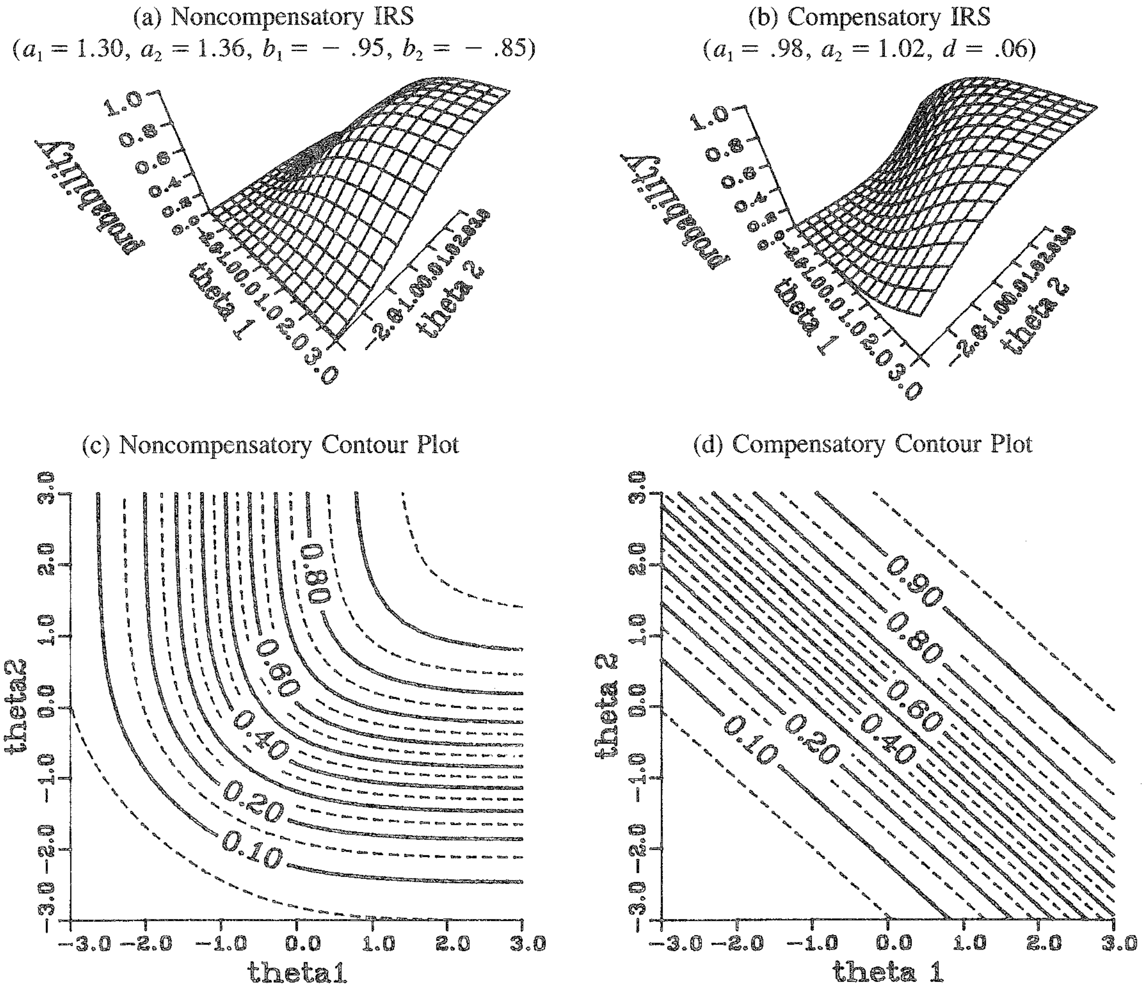
Similar statistics were computed for the item discrimination and difficulty parameters and their unidimensional estimates to determine the effect of confounding difficulty with dimensionality for both the compensatory and noncompensatory item sets.

Results

Descriptive statistics for difficulty indices, item-total biserial correlations, KR-20 reliabilities, and the eigenvalues from a principal components analysis of the tetrachoric correlation matrix were then obtained for each of the eight datasets, to validate

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Figure 3
The Item Response Surfaces and Contour Plots for Item 20



(a) Noncompensatory IRS
($a_1 = 1.30$, $a_2 = 1.36$, $b_1 = -.95$, $b_2 = -.85$)

(b) Compensatory IRS
($a_1 = .98$, $a_2 = 1.02$, $d = .06$)

(c) Noncompensatory Contour Plot

(d) Compensatory Contour Plot

the similarities in item difficulty and to show the dimensionality of the data. These results are displayed in Table 2.

The eight item response sets have the same mean difficulty, with the range of p values also similar. The mean biserials for compensatory and noncompensatory item sets appear more similar as the correlation between θs increases. As the mean biserials increase, the KR-20 reliability coefficients also increase. Evidence of multidimensionality can be seen by forming a ratio of the first to the second eigenvalue, λ1/λ2 (Hambleton & Murray, 1983). As the correlation between the θs increases, the ratio increases, which suggests a more dominant first principal component; at $\rho_{\theta_1\theta_2} = .9$ the data are almost unidimensional.

The LOGIST calibration program converged to a JML solution for each dataset. The maximum number of stages that LOGIST took to converge was 12 for the compensatory datasets ($\rho_{\theta_1\theta_2} = 0.0$) and 15 for the noncompensatory datasets ($\rho_{\theta_1\theta_2} = .3$). The number of stages required for convergence was

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

always greater for the noncompensatory dataset for each correlational level of the two-dimensional θs. However, no pattern emerged between the correlational structure of the generating θs and the number of stages until convergence.

The centroid plots for the LOGIST calibration of the four compensatory and four noncompensatory datasets are shown in Figures 6a and 6b, respectively. The BILOG counterparts are presented in Figures 7a and 7b.

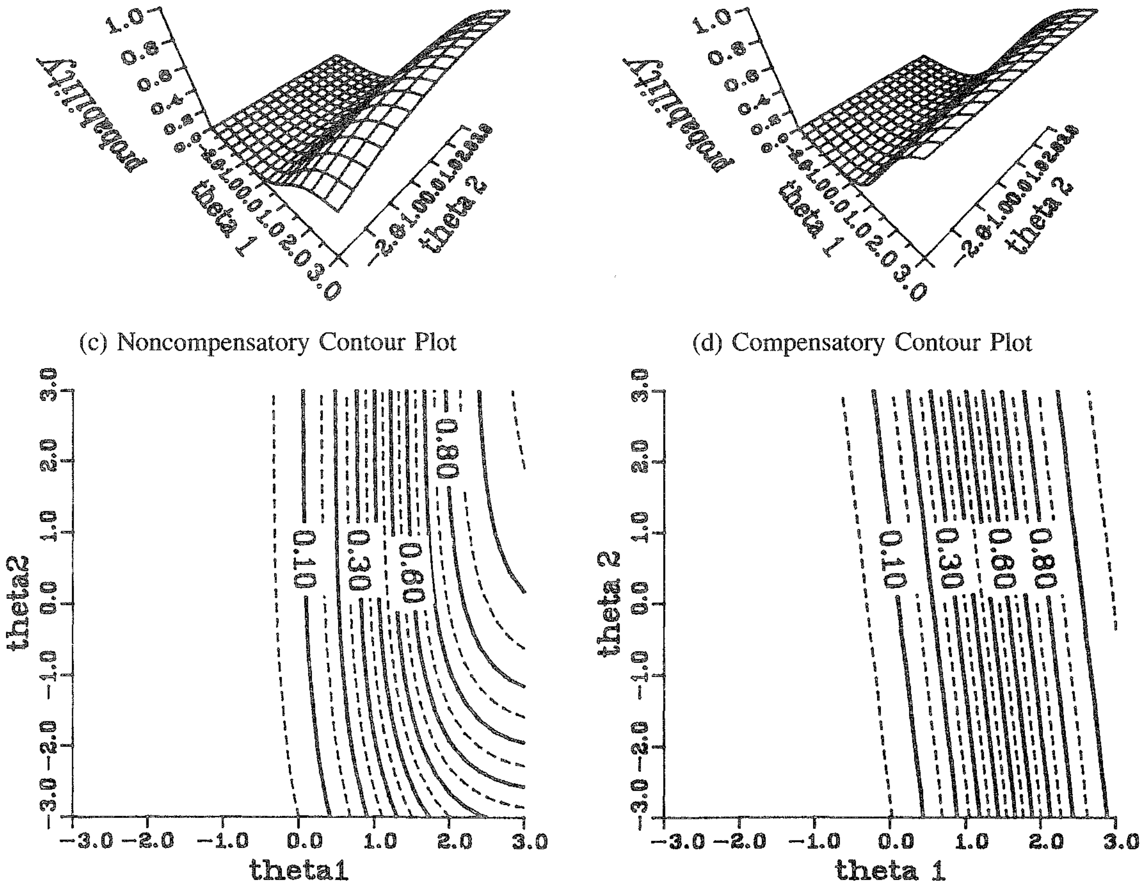
The LOGIST orientation is similar for each level of correlation and for each type of multidimensional model. The BILOG centroids are noticeably more variable. For the BILOG centroids, as r_{θ,θ_1} approaches 0, the plot of the centroids increases in curvature. Thus, BILOG is more sensitive to the confounding of difficulty and dimensionality. When the θ correlation is .9, the centroids for both calibration programs are almost linear.

The correlations and MAD values for θ are shown in Table 3. Compared to the centroid plots, the data are much more alike for compensatory and noncompensatory datasets and for LOGIST estimates compared to BILOG's estimates. It is inter-

Figure 4
The Item Response Surfaces and Contour Plots for Item 40

(a) Noncompensatory IRS
($a_1 = 1.92$, $a_2 = .75$, $b_1 = 1.19$, $b_2 = -3.22$)

(b) Compensatory IRS
($a_1 = 1.80$, $a_2 = .20$, $d = -2.39$)

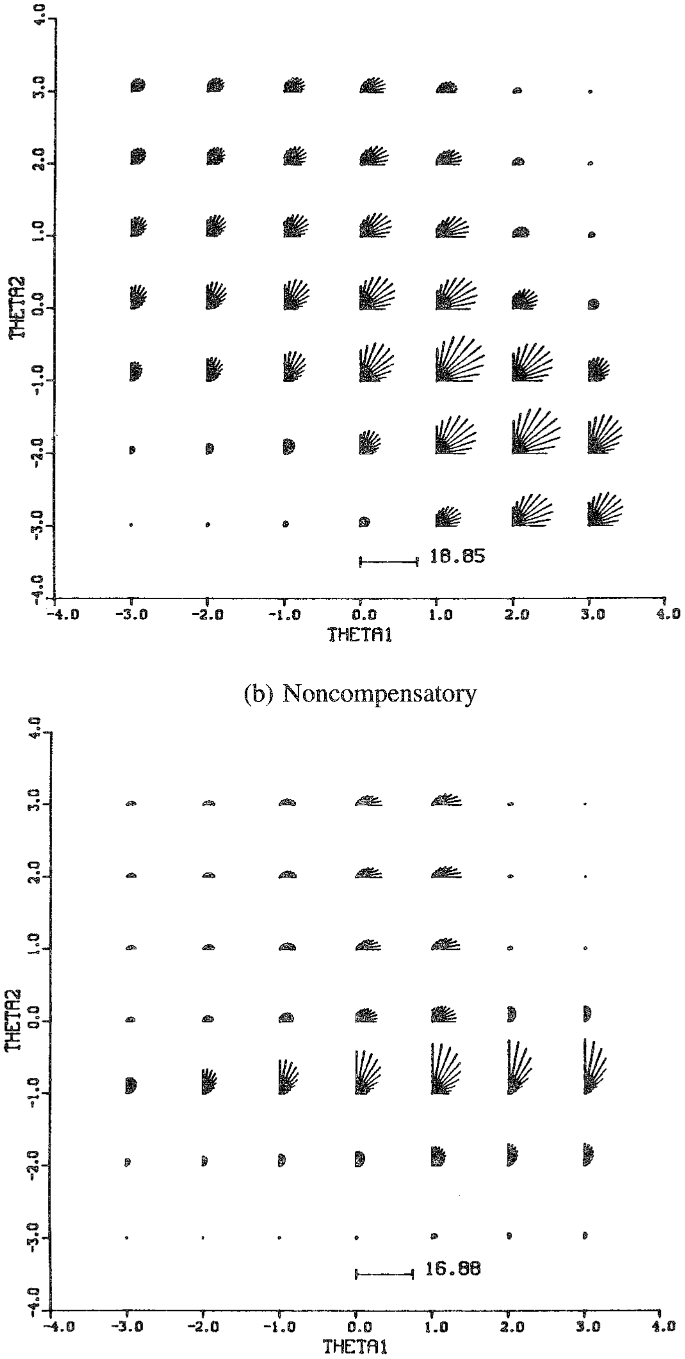


(c) Noncompensatory Contour Plot

(d) Compensatory Contour Plot

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Figure 5
Test Information Vectors at Selected Points in the Ability Plane



(a) Compensatory

(b) Noncompensatory

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 2
Eigenvalues of the First and Second Principal Components
of the Inter-Item Tetrachoric Correlation and Descriptive Statistics
of the Multidimensional Datasets ($N = 1{,}000$, $i = 40$)

Data Type	Eigenvalues			Proportion Correct			Biserial Correlation			Raw Score	
and $\rho(\theta_1\theta_2)$	λ_1	λ_2	KR-20	Mean	Lo	Hi	Mean	Lo	Hi	Mean	SD
Compensatory											
0.0	9.24	2.94	.91	.50	.16	.85	.64	.50	.71	20.15	8.64
.3	10.84	2.59	.93	.50	.17	.84	.69	.57	.75	20.18	9.41
.6	12.17	2.27	.94	.50	.18	.84	.73	.59	.79	20.15	10.04
.9	13.38	2.00	.95	.50	.18	.83	.76	.61	.82	20.18	10.61
Noncompensatory											
0.0	7.22	3.17	.88	.50	.16	.84	.56	.47	.64	20.03	7.64
.3	9.52	2.69	.92	.50	.17	.83	.65	.57	.72	20.08	8.84
.6	11.64	2.25	.94	.50	.17	.84	.71	.65	.76	20.13	9.82
.9	13.53	1.98	.95	.50	.18	.83	.77	.69	.80	20.00	10.67

esting to note that the univariate θ estimates correlate about equally with θ_1 and θ_2 for all levels of θ correlation and for each model. The correlations between θ_1 and $\hat{\theta}$ and between θ_2 and $\hat{\theta}$ range from .59 ($\rho_{\theta,\theta_1} = 0.0$) to .95 ($\rho_{\theta,\theta_1} = .9$). The correlations between $\hat{\theta}$ and the average of θ_1,θ_2 are all quite high, ranging from .89 to .97.

These results parallel the MADs: As the correlation between θ dimensions increases, the MAD values decrease. Thus as the data become more unidimensional, the MAD and correlational values support the conclusion that both programs appear to align the univariate scale about equidistant from the two θ axes.

For the compensatory datasets, correlations and MAD values for a (univariate discrimination) are displayed in Table 4. As the correlation between θs increases, the correlation between $\hat{a}$ and a_1 and between $\hat{a}$ and a_2 approaches 0 for both LOGIST and BILOG. MAD values between the discrimination estimates and parameters were slightly higher for BILOG in all correlational conditions. For both programs, the correlation between $\hat{b}$ and d was $-.99$ for all datasets. This suggests that the pattern of difficulty between the individual items is recoverable to a high degree.

Correlations and average MAD values for the discrimination and difficulty parameters and their estimates for the noncompensatory datasets are in Table 5. The pattern of correlations between discrimination parameters and estimates is similar to that of the compensatory data. The correlations between $\hat{b}$ and b_1 are all .99, while the correlations between $\hat{b}$ and b_2 range from $r = .38$ to $r = .42$ for both LOGIST and BILOG. This suggests that a stronger relationship exists between Dimension 1 and the estimated unidimensional ability for the noncompensatory data. This may also be due to the restricted range of b_2 values.

In both the compensatory and noncompensatory datasets, the $\hat{a}$s correlated positively with a_1 and negatively with a_2, except for the $\rho_{\theta_1\theta_2} = .9$ case. Noticeable differences exist between the MAD values for the noncompensatory discrimination parameters and estimates for BILOG and LOGIST. For LOGIST the average absolute differences of both $a_1 - \hat{a}$ and $a_2 - \hat{a}$ range from .80 to .86, while the range is .32 to .38 for BILOG. For both calibration programs the correlations between $\hat{a}$ and a_2 are negative, except for the $\rho_{\theta_1\theta_2} = .9$ case in which the pattern reverses.

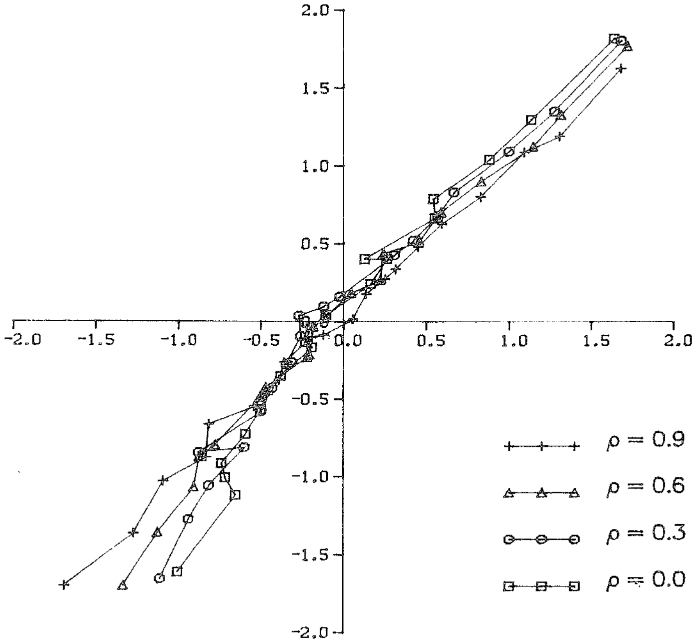
Conclusions

Differences between the item response surfaces for each model when the item parameters are matched

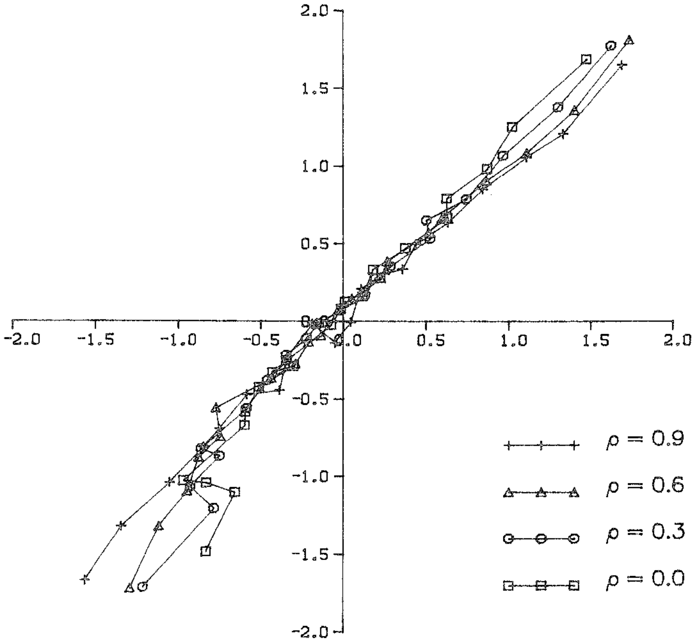
Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Figure 6
Centroids for the LOGIST Calibrated Response Sets Among the Two-Dimensional Abilities

(a) Compensatory Response Sets

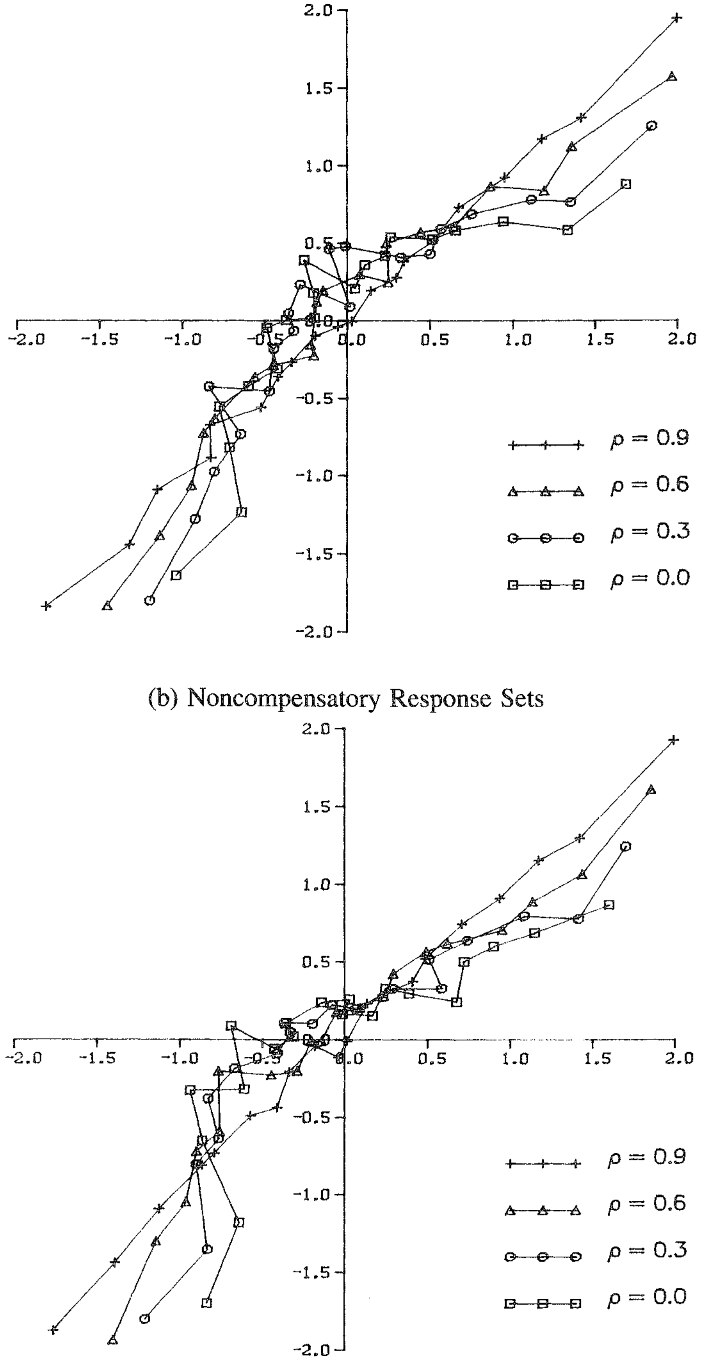


(b) Noncompensatory Response Sets



Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Figure 7
Centroids for the BILOG Calibrated Response Sets Among the Two-Dimensional Abilities


(a) Compensatory Response Sets

(b) Noncompensatory Response Sets

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 3
Correlations and Mean Absolute Differences
Among $\hat{\theta}$, θ_1 and θ_2 By Levels of Correlation for
Compensatory and Noncompensatory Datasets

				$\Sigma\lvert\theta_1-\hat{\theta}\rvert$	$\Sigma\lvert\theta_2-\hat{\theta}\rvert$	$\Sigma\lvert\theta_{avg}-\hat{\theta}\rvert$
$\rho(\theta_1,\theta_2)$	$r(\hat{\theta},\theta_1)$	$r(\hat{\theta},\theta_2)$	$r(\hat{\theta},\theta_{avg})$	k	k	k
Compensatory Data: LOGIST						
0.0	.67	.64	.94	.65	.67	.34
.3	.76	.76	.95	.53	.53	.27
.6	.85	.85	.96	.42	.42	.23
.9	.94	.94	.97	.26	.27	.20
Compensatory Data: BILOG						
0.0	.68	.64	.94	.63	.65	.32
.3	.78	.76	.95	.53	.54	.28
.6	.87	.86	.96	.43	.44	.25
.9	.95	.95	.97	.28	.28	.21
Noncompensatory Data: LOGIST						
0.0	.65	.60	.89	.66	.70	.40
.3	.76	.72	.92	.54	.58	.32
.6	.85	.84	.95	.42	.43	.25
.9	.94	.94	.96	.27	.28	.21
Noncompensatory Data: BILOG						
0.0	.67	.59	.90	.62	.67	.35
.3	.77	.73	.93	.53	.56	.29
.6	.86	.85	.95	.42	.44	.26
.9	.94	.94	.97	.28	.29	.22

appear to be minimal, and they exist at places on the θ_1,θ_2 plane where very few examinees would be expected to be found. Mean *p* values for the eight sets were identical, and the matches on biserial correlations were almost identical for the $\rho_{\theta_1\theta_2} = .9$ case. Thus, the least-squares matching procedures appear to be an excellent method of matching the two multidimensional IRT models.

Table 4
Correlations and Mean Absolute Differences Between LOGIST
and BILOG Estimates and Parameters Under the Compensatory Model

				$\Sigma\lvert a_1-\hat{a}\rvert$	$\Sigma\lvert a_2-\hat{a}\rvert$
$\rho(\theta_1,\theta_2)$	$r(\hat{a},a_1)$	$r(\hat{a},a_2)$	$r(\hat{b},d)$	k	k
LOGIST					
0.0	.30	−.30	−.99	.40	.45
.3	.26	−.26	−.99	.41	.45
.6	.17	−.17	−.99	.41	.44
.9	−.07	.07	−.99	.42	.43
BILOG					
0.0	.26	−.26	−.99	.48	.52
.3	.18	−.18	−.99	.49	.50
.6	.19	−.19	−.99	.48	.50
.9	−.04	.04	−.99	.48	.48

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 5
Correlations and Mean Absolute Differences Between LOGIST and BILOG Estimates and Parameters Under the Noncompensatory Model

$\rho(\theta_1,\theta_2)$	$r(\hat{a},a_1)$	$r(\hat{a},a_2)$	$r(\hat{b},b_1)$	$r(\hat{b},b_2)$	$\frac{\Sigma\lvert a_1-\hat{a}\rvert}{k}$	$\frac{\Sigma\lvert a_2-\hat{a}\rvert}{k}$	$\frac{\Sigma\lvert b_1-\hat{b}\rvert}{k}$	$\frac{\Sigma\lvert b_2-\hat{b}\rvert}{k}$
LOGIST								
0.0	.31	−.23	.99	.42	.86	.82	.67	.87
.3	.27	−.22	.99	.41	.85	.81	.67	.87
.6	.19	−.14	.99	.40	.84	.80	.67	.86
.9	−.07	.06	.99	.38	.84	.80	.67	.85
BILOG								
0.0	.28	−.21	.99	.42	.35	.38	.67	.86
.3	.19	−.17	.99	.41	.33	.37	.67	.86
.6	.19	−.20	.99	.40	.32	.35	.67	.86
.9	−.04	−.01	.99	.39	.32	.35	.67	.85

However, confounding of difficulty with dimensionality, which was reported in Reckase et al. (1986), was replicated only for the BILOG calibration of response data in which $\rho_{\theta_1\theta_2}$ was closer to 0. The "wrap-around" effect of the θ_1,θ_2 centroids did not occur for any of the LOGIST estimation runs. However, in the Reckase et al. study the items only measured θ_1 and θ_2, whereas in this study each item measured a combination of θ_1 and θ_2 to varying degrees. Thus the confounding was not as great as in the Reckase et al. study. Another possible explanation may be the method of estimation. Perhaps the marginal maximum likelihood (MML) procedure of BILOG is more sensitive to the confounding of difficulty and dimensionality.

Confounding of difficulty appeared to have the same effect on θ estimates for both the compensatory and noncompensatory datasets. Despite different item information patterns, as seen in the INFLINE plots, the orientation of the centroids appeared to be the same for both calibration programs.

The correlations among θ parameters and estimates suggest that as the relationship between the two ability dimensions becomes more linear, the data become unidimensional in a sense. As $\rho_{\theta_1\theta_2}$ approached .9, $\hat{\theta}$ correlated in the mid-.90s with θ_1 and θ_2. This was confirmed by the plots of the θ_1,θ_2 centroids and the correlations of the discriminating parameters with their estimates. The correlations between $\hat{a}$ and a_1 and between $\hat{a}$ and a_2 became closer as the correlation between θ_1 and θ_2 increased. Similarly, as $\rho_{\theta_1\theta_2}$ approached .9, the centroids appeared to align themselves along a 45° line. Both of these results suggest that θ_1 and θ_2 were being measured equally.

These results differ slightly from those of Way et al. (1988) and Ansley and Forsyth (1985), in which the relationship between $\hat{\theta}$ and θ_1 was found to be stronger than the relationship between $\hat{\theta}$ and θ_2 for low levels of $\rho_{\theta_1\theta_2}$. However, the relationship Way et al. and Ansley and Forsyth found between the estimated θ values and the average of the true θ_1,θ_2 values was found to be even stronger in this study. Both of these findings may be a function of the method used to generate item parameters in these prior studies, which differed from the method used in this study.

The plots of the θ_1,θ_2 centroids for the 20 $\hat{\theta}$ quantiles revealed differences between the two estimation programs. The centroid plot for LOGIST revealed only a slight confounding effect as $\rho_{\theta_1\theta_2}$ became closer to 0. However, θ_1,θ_2 centroids for BILOG's $\hat{\theta}$ display a much sharper wrapping around about the negative θ_2 axis and the positive θ_1 axis, especially when $\rho_{\theta_1\theta_2} = 0.0$. Thus, BILOG appears

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

to be more sensitive to the confounding of difficulty with dimensionality for both multidimensional models.

Several directions for future research are suggested by this study. One area would be to systematically vary test information with different two-dimensional θ distributions to determine how the interaction affects the orientation of the univariate θ scale in the two-dimensional plane. Also, differences between maximum likelihood and MML estimation of multidimensional response data need to be further explored. Perhaps for the datasets used in this study, BILOG mapped different $\theta_1\theta_2$ combinations than LOGIST onto the univariate θ scale because the BILOG combinations accounted for more variation in the item responses from examinees in that neighborhood.

References

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.

Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program* (Research Report 87-19). Iowa City IA: American College Testing Program.

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

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

Mislevy, R. J. & Bock, R. D. (1982). *BILOG, maximum likelihood item analysis and test scoring: Logistic model*. Mooresville IN: Scientific Software, Inc.

Mislevy, R. J., & Stocking, M.L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57–75.

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.

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 consideration in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57–70). Vancouver BC: Educational Research Institute of British Columbia.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12,* 239–252.

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

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125–145.

Acknowledgments

The author thanks Judy Spray, Mark Reckase, Jim Carlson, and Brad Sympson for their useful comments.

Author's Address

Send requests for reprints or further information to Terry A. Ackerman, ACT, P.O. Box 168, Iowa City IA 52243, U.S.A.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/