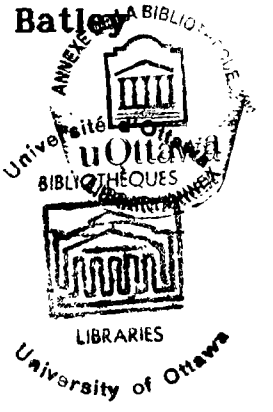
The Effects on Parameter Estimation of Correlated Dimensions and a Differentiated Ability in a Two-dimensional, Two-parameter Item Response Model

Rose-Marie Batley



Thesis submitted to
the School of Graduate Studies and Research
in partial fulfillment of the requirements for the Ph.D.
degree in Education

University of Ottawa

Acknowledgements

Several persons have played integral roles in the completion of this thesis. I am grateful to Dr. Mark Reckase and Dr. Terry Ackerman of American College Testing and to Dr. Marc Gessaroli of the University of Ottawa for their interest and assistance. I am indebted to Dr. James Carlson of Auburn University for the use of his computer program, MIRTE, and for his sustained involvement. I am fortunate to have worked with Jackson Shaw who gave me extensive time and provided expertise in the computer aspects of the project.

I am particularly grateful to Dr. Marvin W. Boss for his constant support, advice, and guidance throughout. His encouragement, coupled with the understanding of my family, have allowed the completion of the research.

Table of Contents

List of Figures

List of Tables

ABSTRACT

The purpose of this study was to assess the effects of correlation between dimensions and differential ability on one dimension on parameter estimation when using a two-dimensional IRT model. Past research has shown the inadequacies of unidimensional analysis of multidimensional item response data. However, few studies have reported multidimensional analysis of multidimensional data and, in those which used simulated data, results were usually based on one replication.

Multidimensional analysis of simulated two-dimensional item response data fitting the M2PL model of McKinley and Reckase (1982a) was done using the analysis program, MIRTE (Carlson, 1987).

Six data sets (2000 ability vectors by 104 items) were generated to satisfy two conditions of the distributions of the ability dimensions and three different degrees of correlation between the two abilities. The six data sets (2 distributions x 3 correlations) and analyses were replicated 100 times each. Summary statistics on the 100 replications were used to assess the effects of degree of correlation between ability dimensions and differential ability on the second dimension.

With the exception of the discrimination parameter on the second dimension and the multidimensional discrimination parameter, ability and item parameters were adequately recovered in the data sets in which both abilities were normally distributed over the full range. In the data sets with a restricted range of ability on the second dimension, recovery of the ability and item parameters was adversely affected. There were some interaction effects noticed in the recovery of the ability and item parameters. As the correlation between the dimensions increased and there was less ability on the second dimension, the dimensions

appeared to become less distinguishable. The latent space seemed to be collapsing into a more unidimensional space when the ability dimensions were correlated 0.50.

Results of the research indicate that MIRTE recovers the structure of a multidimensional correlated space better than previous estimation programs have done, especially in the cases in which the items were multidimensional in themselves. Because of the limitations imposed on any single piece of research in terms of research design, some alternative situations need to be studied. There remains further investigation to be done on the accuracy of estimation procedures when there is inclusion of a guessing parameter as well as with different latent space structures both in terms of population and items.

Chapter 1

INTRODUCTION

Research studies related to Item Response Theory (IRT) have been prominent in the literature over the past two decades. IRT, an alternative to classical testing theory (CTT), is defined by a mathematical model. This model allows for the estimation of item characteristics (e.g., difficulty, discrimination) from the examinees' responses independent of the sample of people and estimation of people characteristics (e.g., ability on the underlying trait) independent of the set of items.

Benefits of IRT rest in its potential to solve many measurement problems which measurement specialists have been unable to solve using CTT (Hambleton & Cook, 1977; Lord, 1980). The theory is applicable at the item level as well as the test level making its application to test construction and equating more direct than that of CTT. If the item response model fits the data, this results in three main advantages of IRT over CTT. (1) The estimate of an examinee's ability is independent of the particular sample of test items administered to the examinee (i.e., item-free ability statistics). (2) The estimation of item parameters is independent of the particular sample of examinees used to calibrate the items (i.e., sample-free item statistics). Item parameter estimates obtained separately for two different groups and put on the same scale should be the same except for sampling error (Linn & Harnisch, 1981). (3) Parallel form reliability of CTT is replaced by the concept of statistical estimation and associated standard errors. Rather than an overall test standard error, there is a statistic indicating the precision with which each examinee's ability is estimated (Bock & Aitkin, 1981; Lord & Novick, 1968). There are also standard errors for each of the estimated item parameters.

The original item response models have stringent assumptions. The two

main assumptions are those of local independence and unidimensionality. Satisfying local independence requires that only the examinee's ability and the characteristics of the test item related to the trait being measured by the test influence performance (Hambleton, 1982). The test item responses of a given examinee are statistically independent.

Unidimensionality requires that all the items measure only one ability or latent trait. If more than one ability accounts for test performance, the test is not unidimensional. The unidimensionality assumption may be the most limiting aspect of IRT.

Consider the situation in which items for a test are designed to measure predominantly one ability (e.g., mathematics) but require some amount of a second ability (e.g., verbal). This second, or interfering, ability could be more crucial to some 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 these abilities are correlated to some extent (particularly in an educational situation). The effect of degree of correlation on the parameter estimates is not known. What would happen to ability estimates on mathematics for these students if a unidimensional IRT model were used to analyze their responses?

Several authors (e.g., Ackerman, 1987a, 1987b; Ansley & Forsyth, 1987, 1985; Bogan & Yen, 1983; Dorans & Kingston, 1985; Drasgow & Parsons, 1983; McCauley & Mendoza, 1985; McKinley & Reckase, 1984; Reckase, 1979, 1985c; Reckase, Carlson, Ackerman, & Spray, 1986) have considered the effects of analyzing known multidimensional data with a unidimensional item response model. The resulting estimates in most cases were not acceptable unless there was clearly one dominant dimension.

Some of the problems that result from analyzing multidimensional data with

a unidimensional IRT model include difficulties with interpretation of the ability at different ends of the unidimensional ability scale (Reckase, Carlson, Ackerman & Spray, 1986) and unidimensional ability estimates which appear to be an average of the multidimensional abilities (Ansley & Forsyth, 1985). For the hypothetical situation mentioned above in which a test is designed to measure mathematics ability but success requires some amount of verbal ability, this would be an unacceptable result. If the item responses were analyzed using a unidimensional model and the results were used for placement or selection, students with high mathematics ability but low verbal ability would be penalized if the model provided an estimate which was the average of the two true abilities. The results of these studies indicate that it is unreasonable to assume a unidimensional item response model will adequately fit multidimensional data in many cases.

In spite of findings that unidimensional models are not often robust to violations of unidimensionality, few researchers have made use of multidimensional models to analyze multidimensional data. Multidimensional IRT (MIRT) models are relatively recent but are being developed and tested. They are more complex than their unidimensional counterparts and require more computer capacity and processing time for analysis of data than the unidimensional models. This makes research using MIRT models expensive.

Researchers who have used multidimensional analysis (e.g., McKinley, 1983; McKinley & Reckase, 1983a, 1983d, 1984; Muraki & Englehard, 1985) have indicated that a multidimensional model more adequately models both real and simulated data than does a unidimensional model. Nevertheless, problems do exist with the MIRT models. As with the unidimensional models, there is no known effective measure of dimensionality (Hattie, 1985). Researchers do not agree on the relationship between factor analysis (FA) dimensions and MIRT dimensions. In general either FA models or MIRT models are used to generate and/or analyze

multidimensional data. Results of studies using the two different models for assessing multidimensional data are difficult to compare. Also, unidimensional procedures for estimation of parameters appear to provide inaccurate estimates as dimensionality increases (McKinley & Reckase, 1984).

Two aspects related to dimensionality which need further study are restricted ranges of ability of the examinees on one or more ability dimensions and the correlation between the ability dimensions. In a two-dimensional test, if a group of examinees had a normal distribution over the full range of the primary ability but a narrower range and lower mean on the secondary ability, how would this affect the parameter estimates? In addition, how well do the current estimation procedures for multidimensional models recover parameter estimates when the underlying abilities are correlated? McKinley and Reckase (1984) considered the effects of correlated abilities and the effects of multidimensional items using estimation procedures which did not accommodate correlated abilities. As the degree of correlation increased, the ability parameters were less well recovered. The question of how a restricted secondary ability affects both item and ability estimates has not been addressed in a multidimensional analysis. The combinations of correlated abilities and restricted secondary ability need to be evaluated in a comprehensive, systematic manner.

The objectives of this research are to determine how parameter estimates are affected by: (a) different degrees of correlation between two latent abilities; (b) responses of examinees with ability on the dominant dimension normally distributed over the full range but with lower mean and narrower ability range on the second dimension; and (c) the interaction of differential ability on the second dimension and degree of correlation between dimensions.

In the next chapter a review of the literature relevant to this research is presented.

Chapter 2

REVIEW OF THE LITERATURE

Multidimensional Item Response Theory (MIRT) models are a relatively recent development. The justification for these models rests in the shortcomings of unidimensional IRT. Therefore, a brief description of unidimensional IRT is provided as background for the theory underlying MIRT models. The unidimensionality assumption is considered in more detail and it is shown that unidimensional models have not been found to be robust to multidimensional data. Their failure in more complex, realistic situations explains the creation and testing of the multidimensional models which are presented. Related studies involving multidimensional analysis of multidimensional data are critiqued in some detail followed by a summary and statement of the research problem.

Item Response Theory and Models

Measurement theory can be divided into two main categories: (1) Classical Testing Theory (CTT) which is founded on Spearman's conception of the observed score as a composite of true score and error score components; and (2) Item Response Theory (IRT) which links the probability of the outcome when a single person attempts a single item to the characteristics of the person and the item (Choppin, 1982). Most of the research and development in testing has been based on CTT models. In the past twenty years, however, IRT has become an increasingly popular area for research and numerous tests are being developed based on IRT.

IRT was initially developed by Lawley (1943, 1944) and Lazarsfeld (1950), but it was Lord (1952, 1953) who stimulated interest in the theory.

The objectives when using IRT are to obtain from the observed response patterns of a sample of examinees either estimates that characterize the items in a test (e.g., item difficulty or discrimination) or estimates that characterize the examinees (e.g., a latent ability) or estimates of both.

There are assumptions underlying the theory. One of these is the assumption that each examinee in a population may be characterized by her/his score on one or more unobserved latent variables, such that in a population of examinees, each with the same score on each latent variable, responses to the items in the test are mutually statistically independent. This is known as the assumption of local independence, an assumption which is stronger than the uncorrelated errors assumption of CTT. Local independence requires that only the examinee's ability and the characteristics of the test item related to the trait being measured by the test influence performance (Hambleton, 1982).

As well as the assumption of local independence, the original IRT models were based on an assumption concerning the dimensionality of the latent space. Dimensionality refers to the number of latent traits that underlie test performance. A latent space is defined by a set of dimensions corresponding to the latent traits (Hambleton & Cook, 1977). The examinee's position in the latent space is determined by latent trait scores. The most simplistic latent space is unidimensional (i.e., all items measure only one ability or latent trait). The original IRT models are unidimensional; they are based on the assumption that only one underlying ability accounts for performance on the test. If there is more than a single ability accounting for test performance, the latent space is then considered to be multidimensional.

A third notion common to all item response models is the assumption of a mathematical function which describes the relationship between the unobserved traits and the observed item responses (Gamache, 1983). Several models have

been proposed and many different mathematical functions have been employed. The function used is referred to as the item characteristic function (ICF). It is a non-linear curve in the unidimensional IRT models, known as an item characteristic curve (ICC). In multidimensional models, it is known as an item response surface (IRS) or hyperplane. Each ICF is identified by certain parameters, whose values define the shape of the ICC or IRS. Different models require different numbers of parameters to define the ICF, which represents the regression of item scores on the latent abilities. Hence, given the ICF for an item, the probability of a correct response for any given ability(ies) can be determined.

Item response models may be classified by various criteria such as mathematical function, dimensionality, or number of parameters. Item response models are readily identifiable by the type of mathematical function used to describe the ICF. The two most commonly used functions are the normal ogive and the logistic function. Estimation procedures for the normal ogive function are very costly. The logistic function is mathematically more tractable and more robust to careless mistakes. As it so closely matches the normal ogive curve and is easier to use, it is preferred by many researchers.

Models may also be categorized according to dimensionality (i.e., unidimensional or multidimensional). As well, the multidimensional models may be either compensatory or noncompensatory. Compensatory models permit high ability on one dimension to compensate for low ability on another dimension in terms of the probability of a correct response. The compensatory models are additive in their mathematical functions. The mathematical formulae for noncompensatory models proposed to date are multiplicative. No amount of one ability can compensate for a deficit in a second ability being tested.

Different item response models have different numbers of item parameters. Typically these parameters are known as the item difficulty, discrimination, and

pseudo-chance level or guessing parameter. Figure 1 shows a typical ICC for a three-parameter logistic (3PL) unidimensional IRT model whose mathematical equation is given by Formula (1).

$$P_i(\theta) = c_i + (1 - c_i) / \{ 1 + \exp[-1.7\, a_i(\theta - b_i)] \} , \qquad (1)$$

The three parameters refer to item i: a_i represents item discrimination; b_i represents item difficulty; and c_i represents the item pseudo-chance level or guessing. $P_i(\theta)$ is the probability of a correct response for item i given ability level θ.

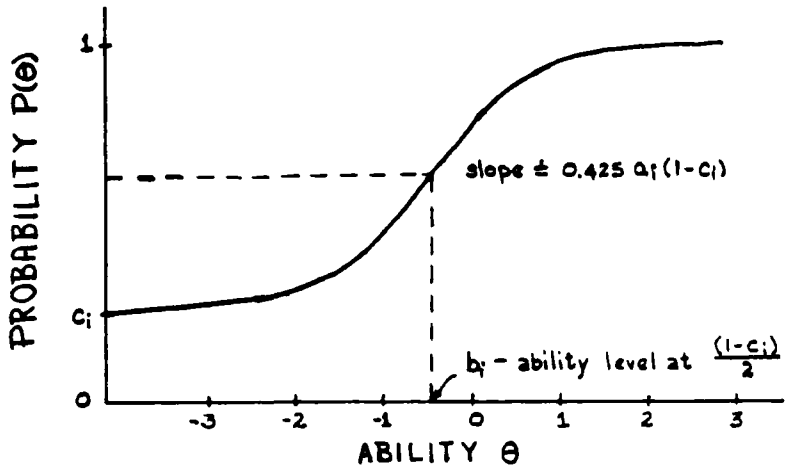


Figure 1. The logistic form of an ICC with three item parameters.

Item difficulty is the point on the ability scale at the point of inflection of the curve. The difficulty parameter is known as a location parameter because it locates the curve on the ability scale. For a more difficult item requiring a higher level of ability to have the same probability of a correct response, the curve in Figure 1 would be shifted to the right moving the point of inflection higher up the ability scale.

The slope of the ICC at the point of inflection is related to the item discrimination. The point of inflection separates the examinees into two groups: those examinees with ability lower than the difficulty level would have a higher probability of responding incorrectly to the item; those examinees with higher ability would have a higher probability of getting the item correct. The slope of the curve at the point of inflection indicates how well the item discriminates between these two groups. If the curve is steep at that point, the item is said to discriminate well because a small increase in ability results in a large increase in probability of a correct response. For flatter ICC s, a large change in ability is required in order to detect much change in probability of a correct response.

This ICC has an upper asymptote of unity and a lower asymptote greater than zero, indicating that examinees of any ability (θ) level have some probability of correctly responding to the item. Lord (1974) characterized this nonzero asymptote as a pseudo-chance level parameter. It is related to guessing but it is usually less than what it would be if it were a result of random guessing.

Not all item response models have the same number of parameters. Often a pseudo-chance level parameter is not included resulting in what is known as a two-parameter logistic (2PL) model. Figure 2 shows a typical ICC for a 2PL model.

For this model the item parameters are difficulty (b_i) and discrimination (a_i) which remain as defined for the ICC shown in Figure 1. Note that the lower asymptote for the ICC in Figure 2 is at zero, which means that c_i is zero for the item. The upper asymptote remains at unity.

One-parameter logistic (1PL) models do not include an item discrimination parameter. Use of these models requires that all items have a constant discrimination parameter (usually unity). Applications of this model are then restricted to situations in which this is a viable assumption. This model is sometimes referred to as the Rasch model although it is in fact a specific case of

the generalized Rasch model (discussed below).

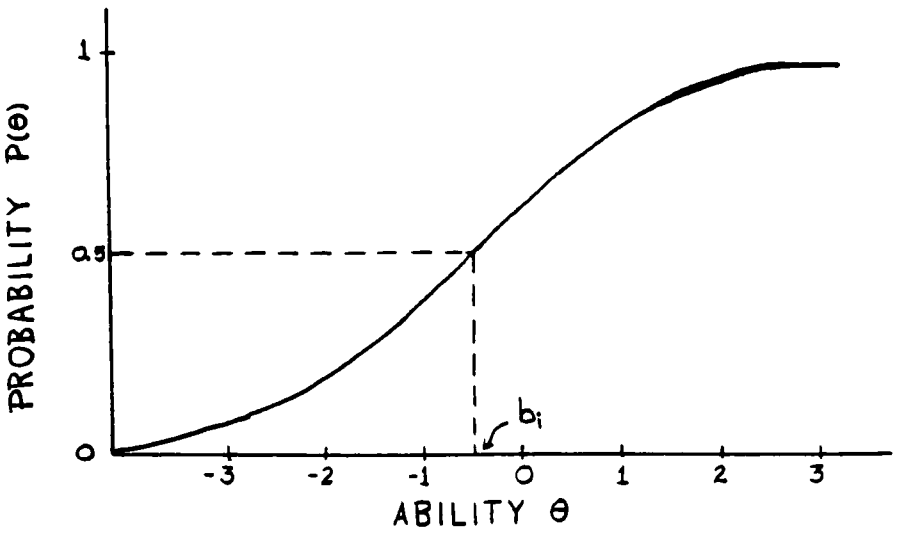


Figure 2. The logistic form of an ICC with two item parameters.

Multidimensional Item Response Theory (MIRT) models are those models for which more than one underlying trait is assumed. The size of the latent space for these models is defined in terms of the ability dimensions of the examinees. To determine the characteristics of the test item, it is necessary to consider the surface describing the relationship between probability of a correct response and the examinee's position in the $\underline{\theta}$-space, where $\underline{\theta}$ is an m-dimensional vector describing the subject's abilities on the *m* dimensions or traits. An item response surface (IRS) replaces the ICC of unidimensional IRT. An example of an IRS for a two-dimensional, two-parameter logistic model is given in Figure 3.

Items are still characterized by location parameters (difficulty of the item) and slope parameters (discrimination of the item). These parameters are described in terms of the derivatives of the function *P* which is monotonically increasing for all m dimensions, is asymptotic to zero as $\underline{\theta}$ approaches $-\infty$ and to one as $\underline{\theta}$ approaches $+\infty$. The difficulty of the item is then defined as the value of $\underline{\theta}$ for which the second derivative of $P(\underline{\theta})$ with respect to $\underline{\theta}$ is equal to zero. This is

equivalent to the unidimensional point of inflection of the ICC. When the class of models is based on the logistic function, the solution gives a difficulty function rather than a difficulty value, i.e., the locus of points in the $\underline{\theta}$-space that yields a 0.5 probability of a correct response to the item.

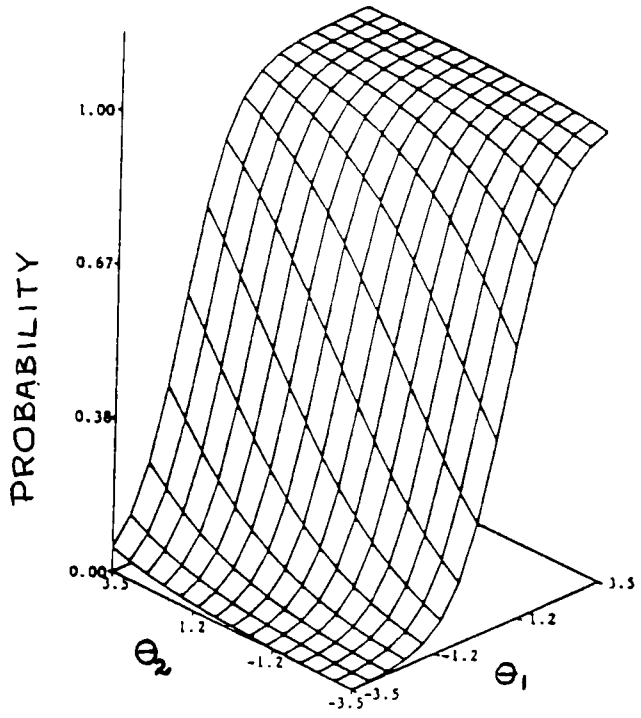


Figure 3. A two-dimensional, two-parameter logistic IRS.

The discrimination is also a generalization from the unidimensional case. In the case of a two-dimensional latent space, the discrimination is the slope of the IRS along the line defining the difficulty of the item. In the m-dimensional case with $m \geq 2$, discrimination is the slope of the IRS at the difficulty hyperplane.

The pseudo-chance level parameter can also be included to make a multidimensional three-parameter logistic (M3PL) model. A non-zero value of this parameter raises the IRS vertically on the probability axis.

There may be values of each item parameter for each ability dimension.

Therefore, as the number of dimensions increases the number of item parameters can increase rapidly making estimation of these parameters more complicated and costly. It is essential then to determine if there is a need for the multidimensional models. This is discussed in the next section.

The Unidimensionality Assumption

First of all, is it reasonable to assume unidimensionality of the abilities underlying performance on the test items? Researchers have different opinions about this. Lord (1968) recognized that the unidimensionality assumption was not strictly true for most tests but he claimed that good approximations to unidimensionality were true in some cases. Harper (1972) suggested the assumption that a set of test items measuring one attribute free from other contamination or extraneous factors was unlikely to happen in practice. So many factors affect dimensionality of a test that Hambleton (1982) contended that the assumption of unidimensionality could never be met. There would always be some additional cognitive, personality, and test-taking factors which would contribute to performance.

Dimensionality has been shown to be affected by test factors such as length of the test (Jones, Wainer, & Kaplan, 1984), time during instructional sequence at which data are collected (Bejar, 1980; Kingsbury & Weiss, 1979), or content or position of items within a test (Yen, 1980). Sources of multidimensionality also include examinee sources such as fatigue and study habits (Bejar, 1980). Traub (1983) pointed out that in the educational environment several factors including curriculum, instruction, and individual differences can affect the unidimensionality if the test is used over time. He emphasized the need to check for unidimensionality of the same test at different points in time. This argument would surely apply as well to checking for unidimensionality in the case where

test items are used over a wide range of age or grade levels. What might be unidimensional at age 12 years could be multidimensional at 6 years. As well, unidimensionality may be more serious in some content areas than in others (Hoover, 1983). Limitations of unidimensional IRT models in practical settings which typically involve multidimensional data have been noted (Loyd & Hoover, 1980; Slinde & Linn, 1979).

Lumsden (1961) suggested unidimensional tests be constructed by choosing items with high inter-item correlations. However, Cattell (1964, 1978), McDonald (1981), and Hattie (1985) have argued that highly correlated items do not necessarily mean unidimensional tests. Cattell noted that requiring homogeneous items was similar to asking the same question in several different ways. Even if an established method existed such that tests could be constructed to be unidimensional, being restricted to this type of test ignores the appealing characteristics of MIRT, one of which is the acceptability of heterogeneous items.

There appears to be enough evidence to suggest the unidimensionality assumption is often not realistic. This does not in itself establish the need for multidimensional models. If unidimensional models fit multidimensional data reasonably well, dimensionality may not be a serious concern. Several studies have been carried out testing the robustness of unidimensional models with multidimensional data. These studies are discussed below.

Unidimensional Analysis of Multidimensional Data.

The multidimensional data for most of these studies have been identified using one of two different techniques. Researchers have used either factor analysis (FA) procedures or a MIRT model to identify or simulate multidimensional data.

The relationship between the linear factor analytic model and the item

response model is not clear and there are varied opinions on the validity of FA techniques when applied to IRT. McDonald (1982), in a review of the relationship between IRT and nonlinear FA, reported that IRT is a special case of nonlinear Common Factor Theory. The confusion in IRT and linear FA seems to centre around the definition of a dimension. In factor analysis a dimension is a common factor. In latent trait situations a dimension is defined by the latent trait itself. A two-dimensional test in the latent trait sense is a test which requires two separate abilities simultaneously to answer correctly. In the FA sense a two-dimensional test as defined by two common factors is more apt to be a test composed of two groups of highly correlated items. Factor analysis doesn't necessarily identify simultaneously necessary abilities (Ansley, 1984). Green, Lissitz, and Mulaik (1977) claimed the existence of only one common factor meant the items were homogeneous in the CTT sense but did not mean there was one latent dimension. Conversely, a set of items may be unidimensional in the latent trait sense but still result in several factors in a FA solution (Hambleton, 1980). Other problems with the use of FA to define dimensions occur depending on the choice of correlation coefficient used (Lord & Novick, 1968, p. 382). The common factor model assumes there is a linear relationship between the latent variables and the observed variables. Because the responses are dichotomous, they cannot be linearly related to continuous latent variables (Hulin, Drasgow, & Parsons, 1983). Phi correlation coefficients, with size contingent upon difficulty, tend to yield a factor known either as a difficulty factor (Wherry & Gaylord, 1944) or a factor due to nonlinearity (McDonald & Ahlawat, 1974). Tetrachoric correlations, although somewhat free of the difficulty factor problem, do tend to decrease as the items becomes less similar in difficulty (Carroll, 1945). Computationally, the sample matrix of tetrachoric correlation coefficients is seldom positive definite necessitating a smoothing process before a maximum likelihood FA can be applied

(Muraki & Englehard, 1985). Lord (1980) stated that tetrachoric correlations were not useful where guessing was prevalent or if the trait was not distributed normally. As conventional linear FA of phi or tetrachoric correlations appears to be less than a satisfactory way of investigating dimensionality, Christoffersson (1975) and Muthén (1978) developed generalized least squares methods which alleviate the difficulties associated with using tetrachoric correlation but their methods are limited to about 25 items because of the computational requirements.

Efforts have been made to combine item characteristic curve theory and FA in order to create a model for determining dimensionality. Bock and Aitken (1981) developed a FA procedure for use with dichotomous data based on multidimensional item response models. This method uses the distinct item response vectors of the examinees and is known as full-information factor analysis (FIFA) because it uses all the information available in the matrix of the response patterns. The FIFA model is described in Zwick (1988). Researchers (e.g., Divgi, 1980; Kingston, 1986; Muraki, 1985; Muraki & Englehard, 1985) who have used FIFA procedures are beginning to make progress in further determining the relationship between IRT and FA.

The results of research using FA procedures are interesting and provide some insights into the need for multidimensional models. Table 1 lists recent studies in which the researchers used unidimensional analysis of multidimensional data where dimensionality is defined in the linear factor analytic sense.

Factor analysis has been used to establish dimensionality, both in cases of simulated data (e.g., Weiss & Suhadolnik, 1982) and real data (e.g., Reckase, 1979). Weiss and Suhadolnik (1982) studied the effects of multidimensionality during the administration of computerized adaptive tests with simulated data generated using FA to define the dimensions. The authors reported that, as the number and

Table 1. Studies Involving Unidimensional Analysis of Multidimensional Data (dimensionality defined by FA procedures)

AUTHOR (YEAR)	FOCUS of STUDY	DATA INFORMATION	ANALYSIS of DATA
McCauley & Mendoza (1985)	Eight Item Bias Indices	Simulated: N = 1000; n = 50; two groups differing on θ_2	LOGIST (2PL)
McKinley & Mills (1985)	Four Goodness-of-Fit Indices	Simulated: N = 500, 1000, 2000; n = 75; three ability levels	LOGIST (1-, 2-, 3PL)
Dorans & Kingston (1985)	Consequences of violation of unidimensionality on equating	Real: N = 2579, 4351; n = 45, 47, 108	LOGIST (2PL)
Drasgow & Parsons (1983)	When can a set of data be considered sufficiently unidimensional?	Simulated: N = 1000; n = 50	LOGIST (3PL)
Weiss & Suhadolnik (1982)	Effects of multidimensionality during administration of CAT	Simulated. N = 1700; n = 5 to 30	3PL weighted by factor loadings; Root Mean Square Error
Reckase (1979)	How factor structure affects latent trait model estimates	Simulated: N = 1000, n = 50 Real: N = 176 to 3126; n = 50	BICAL (1PL) LOGIST (3PL)

strength of factors increased (from one to four), the estimated unidimensional ability deviated further from true ability and the individuals were rank ordered differently. As the overall degree of multidimensionality (as measured by the sum of the eigenvalues for each factor) in the generated item responses increased, the estimated theta values deviated further from the first factor theta values. Both the number and the strength of factors in the underlying factor structure affected theta estimates.

In one of the first studies of this type, Reckase (1979) reported that factor structure does affect estimates for an item response model. When there was a dominant first factor, both the BICAL (Wright & Panchapakesan, 1969; a 1PL program) and LOGIST 3PL (Wingersky, Barton, & Lord, 1982) programs gave unidimensional ability estimates which were highly correlated with scores from that factor. When there were two or more equally potent factors, the one-parameter unidimensional ability estimate from BICAL appeared to be a combination of the factors while LOGIST appeared to estimate only one of the factors. These results held for both real and simulated data. That the number of parameters in the unidimensional model used can affect the ability estimates is a cause for concern.

Dorans and Kingston (1985) considered the consequences of violation of unidimensionality on equating using real data from the Graduate Records Examination Aptitude Test. Two verbal dimensions were defined by FA of interitem tetrachoric correlations and these were considered as two subtests. The full test was then considered two-dimensional. The authors reported that dimensionality had an effect on equating via its effect on the magnitude of the item discrimination estimates. The estimates for discrimination were higher in calibration of homogeneous items on the single subtests. While the results of this study are based on the acceptance of the notion of common factors as dimensions,

it seems reasonable that multidimensionality would affect discrimination estimates.

Drasgow and Parsons (1983) found that the LOGIST analysis program, based on a unidimensional model for the data, was drawn to a general factor if one existed in the multidimensional data or to the strongest group factor otherwise. Even though the authors used items which loaded on one dimension only (i.e., the test was composed of unidimensional subtests), they concluded that unidimensional models were inadequate for multidimensional data unless there was a sufficiently strong single dominant trait.

McKinley and Mills (1985), in an evaluation of four goodness-of-fit indices, concluded that a unidimensional model doesn't adequately fit multidimensional data. The authors investigated fit for three different groups of ability levels (low, centred, and high) and found only minor fluctuations in the fit for the three groups. The conclusion reached was that unidimensional models do not fit multidimensional data regardless of the ability level being tested.

The FA study most relevant to the proposed research is that of McCauley and Mendoza (1985) in which eight item bias indices were being evaluated. Data were generated based on a linear factor analysis model for 20 data sets with one general factor and zero to two secondary factors. Two groups of simulees differing on mean level of secondary required ability were used. Lowering the mean secondary ability of one group affected the unidimensional estimates of difficulty more so than discrimination. However, McCauley and Mendoza didn't examine the effects of different loadings on the first factor which are directly related to the discrimination parameter. As unidimensional analysis of multidimensional data has not produced satisfactory estimates in many instances, it is not clear whether results reported by McCauley and Mendoza would be valid in a multidimensional analysis of the same multidimensional data.

Four results relevant to this research are evident from these studies. Firstly, as the number of dimensions increases, the unidimensional ability estimates are less accurate in terms of size and rank order. Also, the unidimensional item discrimination estimates are less accurate. Secondly, unidimensional analysis programs appear to be drawn most often to a general factor or the strongest factor or some combination of the factors. This is not always acceptable even if it were predictable which factor or combination of factors was being represented. Thirdly, the unidimensional models do not fit multidimensional data even when applied to groups of examinees with smaller ranges of ability for certain dimensions. Lastly, unidimensional analysis of two-dimensional data from grouping with a lower mean on the second ability dimension does affect the unidimensional difficulty estimate, an item parameter which is usually very stable. These results are indicative of a lack of fit of unidimensional models to multidimensional data both for item and ability parameter estimates.

In Table 2 twelve studies have been listed in which the data were generated by a MIRT model. There is a consistency here not evident in the previous studies in that both the generation programs and the analysis programs are based on IRT although the data are generated according to a multidimensional model and the analysis program is based on a unidimensional model. In determining the robustness of unidimensional models to multidimensional data, the results are not confounded by having the dimensions defined by factor analysis. Results of these studies also demonstrate that unidimensional models are not robust to multidimensional data except under somewhat restricted conditions such as tests with a single dominant underlying dimension.

Yen (1985), in a study of 1000 simulated responses to 30 items (using the multidimensional three-parameter logistic (M3PL) compensatory model for generating data and LOGIST for analysis of data) found that the unidimensional

Table 2. Studies Involving Unidimensional Analysis of Multidimensional Data (dimensionality defined by MIRT)

AUTHOR (YEAR)	FOCUS of STUDY	DATA INFORMATION		ANALYSIS of DATA
Greaud (1988)	Effects of varying degrees of correlation on IRT estimation	Simulated	M2PL Compensatory Model, N = 2000, n = 98, 5 levels of correlation, 5 replications	Lord's M L E (items) Bayes Modal Est (θ)
Ackerman (1987a)	A comparison of unidimensional estimation of compensatory and noncompensatory multidimensional data when difficulty and dimensionality are confounded	Simulated	M2PL Compensatory Model, N = 1000, n = 40, Sympson's Noncompensatory Model, N = 1000, n = 40, 4 levels of correlation between dimensions	LOGIST, BILOG
Ackerman (1987b)	The use of unidimensional item parameter estimates of multidimensional items in adaptive testing	Simulated	M2PL Compensatory Model, N = 2000, n = 100, N = 1000, n = 96	LOGIST (1-, 2PL) MIRTE
Ansley & Forsyth (1987)	Comparison of the effect of using the unidimensional one-parameter logistic model with various types of multidimensional data	Simulated	M2PL Compensatory Model, M2PL Noncompensatory Model, N = 2000, n = 60, 4 levels of correlation between dimensions	LOGIST (1PL)
Way, Ansley, & Forsyth (1986)	The nature of unidimensional estimates from two-dimensional data	Simulated	M2PL Compensatory Model M2PL Noncompensatory Model	LOGIST
Reckase, Carlson, Ackerman & Spray (1986)	The meaning of unidimensional estimates from multidimensional data	Simulated	M2PL Compensatory Model, N = 1000, 1500, n = 20	LOGIST
Ansley & Forsyth (1985)	The nature of unidimensional estimates from two-dimensional data	Simulated	Sympson's Noncompensatory Model, N = 1000, 2000, n=30,60	LOGIST (3PL)
Doody (1985)	Effects of correlations between abilities and among item parameters on estimation	Simulated	Bogan & Yen Compensatory 3PL Model, N = 2000, n = 30, 3 ability groupings, 2 replications	PCA (r_{tet}), PFA (r_{phi}), LOGIST (3PL)
Reckase (1985c)	Effect of applying unidimensional models to items that are multidimensional	Simulated	M2PL Compensatory Model, N = 5000, 3000 , n = 19	LOGIST (2PL)
		Real ACT	Mathematics Test Data, n = 40	MAXLOG, LOGIST (3PL
Yen (1985)	Scale shrinkage caused by analyzing multidimensional data with a unidimensional program	Simulated	M3PL Compensatory Model, N = 1000, n = 30	LOGIST
Hattie (1984)	Dimensionality Indices	Simulated	Hattie's M3PL Compensatory Model, N = 500, 36 data sets, MONTE CARLO, 24 replications	BICAL, NOHARM, FADIV, NOFA
Bogan & Yen (1983)	Effect of two-dimensional data on use of 3PL model in equating	Simulated	M3PL Compensatory Model, N = 2000, n = 30, Three ability groupings	LOGIST

estimate of ability was a weighted combination of the underlying dimensions with weights proportional to the relative discrimination. Bogan and Yen (1983) generated data using a compensatory M3PL model and analyzed the data with LOGIST. They were primarily interested in detection of multidimensionality and assessing its effects on vertical equating. While the authors did consider different correlations of ability dimensions, the role of the correlation of abilities in determining potency of the two dimensions was not clearly defined and there was no attempt to create data that systematically varied in dimensionality. The research does provide another illustration of the inadequacy of unidimensional analysis techniques for multidimensional data.

Doody (1985) (formerly Bogan) seems to have extended this study. She generated data to fit the Bogan and Yen (1983) model such that there were 8 degrees of correlation between the two ability dimensions and various item configurations from very easy to very difficult. Doody considered three groups of simulees (2000 of low ability [$\theta_L \sim (-0.57, 1)$]; 2000 of medium ability [$\theta_M \sim (0, 1)$]; 2000 of high ability [$\theta_H \sim (0.57, 1)$]). Doody wanted to investigate the robustness of item and ability parameter estimation using the 3PL model to violations of the unidimensionality assumption. She analyzed the data with FA methods and using LOGIST 3PL. The results of the FA indicated that as the correlation between the ability dimensions increased, the size of the first eigenvalue increased and as the test got harder, the size of the first eigenvalue decreased. The correlations between the factors determined in FA did not follow the pattern of correlations between traits. This is similar to the results of McKinley & Reckase (1984) (reported later in thesis) although they used a different model and a multidimensional IRT analysis as well as FA. In the IRT analysis, the LOGIST estimate for difficulty correlated well with both multidimensional difficulty parameters when the two difficulty parameters were highly correlated. As they

became less correlated the unidimensional difficulty estimate appeared to be recovering the first difficulty parameter. The LOGIST estimate of discrimination did not represent either of the true discrimination parameters supporting the results of Drasgow & Parsons (1983) that for some multidimensional data, LOGIST is not drawn to a general factor. The LOGIST unidimensional ability estimate improved as the two dimensions became more highly correlated (i.e., approached unidimensionality). As the test became harder, the ability dimension was less well recovered. Doody concluded that the unidimensional model describes two-dimensional data well when the correlation between the dimensions is above 0.5.

Reckase (1985c) considered the effect of applying unidimensional models to multidimensional items (i.e., items that required more than one ability in order to respond correctly). Both real and simulated data (generated to fit a M2PL model) were poorly fit by a unidimensional model. The unidimensional estimate of ability (provided by LOGIST) could be given different interpretations at different points on the unidimensional ability scale. At the lower end of the unidimensional ability scale the LOGIST ability estimate was related to the easier items which primarily measured the first dimension. At the upper end of the scale, the LOGIST ability estimates were related to the harder items measuring predominantly the second dimension. In this case the LOGIST ability scale had a different meaning depending on the part of the scale being considered. This makes it difficult to interpret test scores if different parts of the scale measure different constructs.

Ansley and Forsyth (1985) generated data using Sympson's (1978) noncompensatory, two-dimensional logistic model. The authors selected item parameters so that the generated data would match item difficulty parameters as taken from a "real" test. Situations were examined in which abilities were correlated 0.0, 0.3, 0.6, 0.9, 0.95. The data were analyzed with LOGIST. The authors determined that violations of the unidimensionality assumption had an

effect on the parameter estimation. The estimated unidimensional discrimination parameter was most highly related to the average of the two multidimensional discriminations. The estimated item difficulty appeared to be an overestimate of the difficulty of the items on the first dimension. Estimated unidimensional ability was most highly related to the average of the two true ability values. As the correlation between the two dimensions increased, the relationships described above between the estimates and the true parameters became stronger.

In a similar study using a compensatory model to generate data, Way, Ansley, and Forsyth (1986) reported findings similar to those of Ansley and Forsyth (1985) regarding the nature of the ability estimates. The unidimensional discrimination estimates, however, appeared to be most similar to the sum of the two discrimination parameters. The unidimensional difficulty estimates were most highly related to the average of the two difficulty parameters.

Ansley and Forsyth (1987) hypothesized that the number of item parameters and the complexity of their interrelationships were factors which affected the results of studies in which multidimensional data are analyzed with unidimensional programs. The authors generated data to fit four types of two-dimensional, two-parameter logistic models (two compensatory models, one with one difficulty parameter per item and the second with two difficulty parameters per item; and two noncompensatory models with the same difficulty parameter distinctions) in which the discrimination parameters were set to 1.0. The correlation between θ_1 and θ_2 was also varied (0.0, 0.3, 0.6, 0.9) to determine the effects of different correlations between dimensions. The data were analyzed using LOGIST (1PL) and the parameter estimates were correlated with the corresponding parameter values. The unidimensional difficulty parameter estimates were more similar to the difficulty parameters for the compensatory data sets than for the noncompensatory data sets. For the compensatory data sets

the estimated difficulty values were very similar to their corresponding parameter values or to the average of the two difficulty parameter values for the model with two difficulty parameters per item. For the noncompensatory data sets the difficulty estimates appeared to be overestimates of the true difficulty values. Regardless of the type of two-dimensional data set, the unidimensional ability estimates were best considered as the average of the true ability parameters. The more highly correlated the theta vectors, the stronger this relationship became. The authors concluded that simplifying the nature of the model did not yield clearer results. The interpretation of LOGIST estimates for two-dimensional data appears to depend on the nature of the data. As well it appears that different interpretations of unidimensional estimates are possible depending on whether a compensatory or noncompensatory model is used to fit the multidimensional data.

Reckase, Carlson, Ackerman, and Spray (1986) discussed the interpretation of unidimensional estimates for simulated multidimensional data generated using a multidimensional two-parameter logistic (M2PL) compensatory model. Two data sets were generated. The first data set was for 1000 examinees responding to 20 items (10 easy items forming Dimension 1 and 10 harder items forming Dimension 2). For the second data set, 1500 response vectors were generated for 20 items which were multidimensional. Estimates were calculated using LOGIST. The authors determined that when difficulty and dimensionality of the items in a test are confounded (e.g., easy items measure only one ability and difficult items measure the other ability) the unidimensional ability estimate scale may have a different meaning at different points on the scale. This is similar to the findings of Reckase (1985c). It is difficult to imagine a situation in which it would be acceptable to have different interpretations of ability from the same test for examinees at different ability levels.

Ackerman (1987b) investigated the use of unidimensional item parameter

estimates of two-dimensional items in adaptive testing situations. Using 2000 randomly generated bivariate normally distributed ability vectors and item parameters for 100 items (which tended to be primarily measuring one of the two dimensions), response vectors were simulated to satisfy a M2PL compensatory model. These response vectors were analyzed with LOGIST (2PL). Ability estimates were made under the conditions of a computer adapted test (CAT) situation and for the entire sample of simulees and items. Ackerman suggested that while a large item pool might provide the same amount of information in each situation at all points in the $\theta_1\theta_2$-plane, smaller subsets of the pool are not guaranteed to provide the same amount of uniform information. In other words, different points on the ability plane did not receive parallel tests (i.e., tests that provide the same amount of information on each point through the ability scale (Samejima, 1977)).

In the same study Ackerman repeated this procedure using 3000 real response vectors to the ACT Assessment Math Usage Test Form 26A. He analyzed these responses using MIRTE (Carlson, 1987), a multidimensional analysis program, to obtain estimates for the multidimensional item parameters. Using these estimates as the multidimensional item parameters, Ackerman expanded the 40 item test by simulating additional items to obtain parameters for 96 items (16 for each of six content areas) and generated 3000 response vectors using a compensatory M2PL model for these items. These response vectors were analyzed with LOGIST in order to obtain unidimensional parameter estimates. The author then ran two simulations, a CAT situation and entire test situation, each for 1000 simulees. In this part of his study, Ackerman was attempting to determine if different abilities received different compositions of items in the CAT simulation. Based on his results, the author reported that this is a valid concern. The orientation of the univariate ability scale in the two-dimensional ability plane appears to be a function of multidimensional composition of the items administered

in the test. Hence, strictly parallel tests will not be administered at all (θ_1,θ_2) points in the ability plane.

To further investigate the differences between the compensatory and the noncompensatory MIRT models, Ackerman (1987a) used 1000 (θ_1,θ_2) combinations from the bivariate normal distribution and generated response vectors for 40 items for both the compensatory M2PL model and Sympson's (1978) noncompensatory model. Difficulty and dimensionality were confounded as in the Reckase, Carlson, Ackerman, and Spray(1986) study although each item in this study was multidimensional in varying degrees rather than unidimensional on one of two dimensions. Ackerman generated eight data sets, four for each model at different levels of correlation between the dimensions (0.0, 0.3, 0.6, 0.9). The response vectors were analyzed twice, once with LOGIST and then with BILOG (Mislevy & Bock, 1982). While Ackerman found BILOG to be more sensitive to the confounding of difficulty and dimensionality for both compensatory and noncompensatory data, both programs tended to measure one dimension more strongly, particularly so with the noncompensatory data. As the correlation between the ability dimensions increased, the data tended to become more unidimensional as the estimated θ aligned itself more with one of the two ability dimensions as evidenced by the correlations. However, centroid plots of (θ_1,θ_2) and the correlations of the discrimination parameters and estimates suggested that both θ_1 and θ_2 were being measured equally. It is still not clear then what in fact is happening when unidimensional estimates are made of multidimensional data.

An interesting observation of studies reviewed thus far is that there have been no replications for the simulated data situations. Only Hattie (1984) used a Monte Carlo approach with 24 replications. Replications are expensive and this alone would prohibit extensive use of Monte Carlo research. Nevertheless, conclusions based on results from one set of simulated data leave room for some

doubt regardless of the numbers of items or subjects chosen. Hattie used data generated to fit his multidimensional three-parameter logistic compensatory model. The data sets consisted of one, two, or five dimensions (with intercorrelations of 0.1 or 0.5 between the factors). Unidimensional analysis of the data was used in order to assess different measures of dimensionality. Although the emphasis of Hattie's research differs from the proposed research, Hattie did report some findings relevant to this study. He found that when the dimensions were correlated 0.5, many of the indices of dimensionality did not distinguish between one and more than one dimension. His results also indicated that indices based on component or factor analysis do not aid in determining unidimensionality. When data are multidimensional and the dimensions are even moderately correlated, it became more difficult to detect multidimensionality.

In a more recent study, Greaud (1988) generated data to fit the M2PL compensatory two-dimensional model for 2000 simulees with abilities selected from a bivariate normal distribution. There were five levels of ability correlation (0.2, 0.4, 0.6, 0.8, or 1.0). The 98 items in her simulated test were such that one half were pure on the first dimension and the other half were pure on the second dimension, making the test a composite of two unidimensional subtests of equal difficulty and discrimination levels. Lord's (1980) maximum likelihood estimation was used to obtain discrimination and difficulty parameter estimates from the response data; Bayes modal estimation was used for calculation of the ability estimates. Greaud found that as the correlation between the ability dimensions decreased, the unidimensional ability estimate was closer to one of the true ability parameters, i.e., the estimation procedure appeared to pick up a dominant dimension even though none existed. The data sets for which the ability dimensions were correlated 0.2 and 0.6 were replicated four more times. High multiple correlation coefficients indicated that the ability estimates were strongly

related to some combination of the true abilities. However, comparison of the simple correlations indicated an increasingly stronger relationship between the unidimensional ability estimate and one or the other true ability parameters as the correlation between the true abilities decreased. Greaud also noted that conventional ability estimates (as determined by the raw scores) were unaffected by differing correlations between the abilities. When the correlation between the ability dimensions was low, some item subsets appeared to be more readily identifiable as distinct domains. Greaud recommended the further development of multidimensional models.

The results of these studies with respect to unidimensional analysis of MIRT data are similar to those reported previously for the FA data. Unidimensional ability estimates from multidimensional data appear to be combinations of the true abilities related to the item discrimination values and they may be different combinations at different points on the unidimensional ability scale. Item parameter estimates are also affected. Interpretation of unidimensional item parameter estimates for multidimensional data seems to depend on the model chosen, the degree of correlation between the dimensions, and the characteristics of the data set. When the data are multidimensional and the dimensions are moderately correlated, multidimensionality becomes difficult to detect and recovery of parameters assuming a unidimensional model is inconsistent.

Dimensionality is an important issue in IRT. If unidimensional models are used, there needs to be a means of identifying the data as unidimensional. Factor analysis is not universally accepted as a means of assessing dimensionality (Kim & Mueller, 1978). Breaking a test into individual unidimensional subtests when dichotomously scored items are used or when guessing can be a factor is not easily or adequately done (Reckase, 1981). According to Hattie (1985), in a methodology review of assessments of unidimensionality, there is still no accepted and effective

index of dimensionality of a set of items. Yet unidimensionality remains axiomatic for the IRT models. The simplicity of a unidimensional space almost ensures its absence in the real world, particularly in educational situations. Because of the restrictions this assumption places on the development and applications of IRT, it seems natural to remove the restriction and develop more general models for use with multidimensional data.

The need for multidimensional item response theory is apparent. There is no denying that when a unidimensional model fits the data, there are substantial savings in computer analysis time and costs. However, the estimates obtained for multidimensional data using unidimensional models are acceptable only in very restricted cases. Multidimensional models and analysis programs exist and it is necessary that these models be developed, tested, refined, and used for analysis of multidimensional data in order to obtain accurate and valid estimates of item and ability parameters and to avoid loss of information. It is important to remember that these models have not undergone extensive testing and as a result, little can be said about their use and application in most cases. In this next section, multidimensional models will be presented. Unless stated otherwise, the models are compensatory.

Multidimensional Item Response Models

Generalized Rasch Model.

Rasch (1961) introduced this model initially as a unidimensional model but pointed out that it could be extended to a multidimensional model by replacing the item and person parameters with vectors. The model (Formula 2) is extremely general allowing for both dichotomous and polychotomous responses and including both the 1PL and 2PL models as special cases (Reckase & McKinley, 1982a).

$$P(x = j \mid \underline{\theta}, \underline{\Omega}) = \frac{\exp\{\underline{W}'_j\underline{\theta} + \underline{u}'_j\underline{\Omega} + \underline{\theta}'\underline{V}_j\underline{\Omega} + z_j\}}{\sum_{j=1}^{r}\{\exp[\underline{W}'_j\underline{\theta} + \underline{u}'_j\underline{\Omega} + \underline{\theta}'\underline{V}'_j\underline{\Omega} + z_j]\}}, \quad (2)$$

where x is one of the r possible item responses; $\underline{\theta}$ is the vector of person parameters with elements θ_k, $k = 1, 2, \ldots, m$; $\underline{\Omega}$ is the vector of item parameters with elements Ω_k, $k = 1, 2, \ldots, m$; $\underline{W}_j$ and $\underline{u}_j$ are vectors of weights for each item response; $\underline{V}_j$ is a matrix of weights for each item response; and z_j is a scalar constant for the item response.

The generality of the model makes it difficult to determine the form of the difficulty and discrimination functions. Rasch never attempted to apply this model but McKinley and Reckase (1982a) worked on a specific case for dichotomous data. They set $\underline{W}$ and $\underline{u}$ as unit vectors for correct responses and zero vectors for incorrect responses, $\underline{V}$ became the identity matrix for correct responses and the zero matrix for incorrect responses and z was zero for all responses. The resulting model is much more tractable yielding a difficulty which is a hyperplane and the conditional slopes of the surface intersecting the 0.5 plane are functions of $\underline{\Omega}$.

By generating and analyzing data for several versions of the Rasch model, McKinley and Reckase found the model (Formula 3) for the two-dimensional case with dichotomous data to be most useful. The model has three item parameters and two person parameters.

$$P(x = 1 \mid \underline{\theta}, \underline{a}_i, d) = 1 / \{1 + \exp[-(a_1\theta_1 + a_2\theta_2 + d)]\}, \quad (3)$$

where $P(x=1 \mid \underline{\theta}, \underline{a}_i, d)$ is the probability of a correct response given the ability vector ($\underline{\theta}$), the discrimination vector for item i ($\underline{a}_i$), and a measure of difficulty for

item i (d). An example of the probability surface was given for $a_1 = 1.5$, $a_2 = 0.5$, and $d = 0.65$ in Figure 3 (see p. 11).

McKinley and Reckase determined this version of the Rasch model to be the most mathematically tractable and to yield the most realistic data sets (McKinley & Reckase, 1982b; Reckase & McKinley, 1982b). It is an extension of Birnbaum's (see Lord, 1968) unidimensional two-parameter logistic model to the two-dimensional space. (Many of the multidimensional models are extensions of unidimensional counterparts.) Reckase and McKinley have done extensive testing and development with this model (McKinley & Reckase, 1983a, 1983b, 1983c, 1983d, 1984; Reckase, 1985b, 1985c, 1986; Reckase & McKinley, 1983). A description of the most recent version follows in Chapter 3.

Mulaik's Model.

Mulaik (1972) proposed an extension of the unidimensional 1PL model to the multidimensional space. The equation for the probability of a correct response is given by:

$$P(x = 1 \mid \mathbf{t}, \boldsymbol{\pi}) = \sum_{k=1}^{m} \gamma_k \pi_k \Big/ \left\{ 1 + \sum_{k=1}^{m} \gamma_k \pi_k \right\}, \qquad (4)$$

where the γ_ks are item parameters and the π_ks are person parameters and m is the number of dimensions. The surface defined by Formula 4 doesn't have a point or a line of inflection. (See Figure 4 for a graph of an IRS satisfying this model.) The conditional slope of the IRS at the intersection of the 0.5 plane is a function of both the item and the ability parameters. Ability is defined on a range from 0 to ∞. Mulaik developed a procedure for estimation of the parameters which doesn't appear to have been applied, possibly because of the excessive computation requirements. The model has received little attention from other researchers.

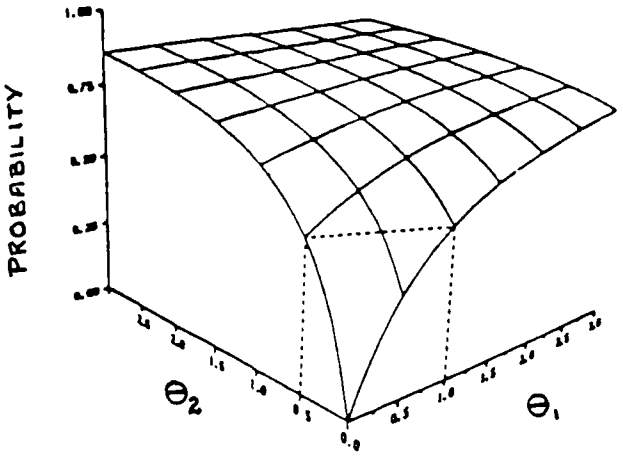


Figure 4. An item response surface for Mulaik's model

Samejima's Model.

Samejima (1974) developed the model (Formula 5) which has some similarities to linear factor analysis in that it expresses item responses as a function of the item and person parameters on a set of latent ability factors. She extended her unidimensional continuous response model (Samejima, 1973a, 1973b) to the multidimensional latent space to achieve this model.

$$P_z(\underline{\theta}) = \int_{-\infty}^{\underline{a}'(\underline{\theta} - \underline{b})} \Lambda(u)\, du \quad , \tag{5}$$

where z is the point of dichotomy of the continuous trait measured by an item; $P_z(\theta)$ is the probability of a correct response; $\underline{a}$ is a vector of discrimination parameters; $\underline{\theta}$ is a vector of ability parameters; $\underline{b}$ is a vector of difficulty parameters; and $\Lambda(u)$ is a twice-differentiable function.

When Λ is the logistic function, the model is a special case of the general Rasch model (Reckase & McKinley, 1982a). When Λ is the normal density function, the model given in Formula 5 becomes equivalent to Bock and Aitken's (1981) multidimensional model (described below). Samejima never applied this model but suggested how the parameters might be estimated. Further use of the model was not found.

Sympson's Model.

Sympson (1978) extended the 3PL model to give:

$$P(x=1|\underline{\theta}, \underline{a}, \underline{b}, c) = c + (1-c) \Big/ \prod_{k=1}^{m} [1 + \exp(-1.7a_k(\theta_k - b_k))], \qquad (6)$$

where x is the item response; $\underline{\theta}$ is the vector of ability parameters; $\underline{a}$ is the vector of discrimination parameters; $\underline{b}$ is the vector of difficulty parameters; c is the pseudo-chance level parameter; and m is the number of dimensions.
(See Figure 5 for a graph of the two-dimensional case in which $c = 0.2$, $a_1 = 0.7$, $a_2 = 1.2$, $b_1 = -0.6$ and $b_2 = 0.5$.)

This is the first noncompensatory model presented. It expresses item response probabilities as the continued product of several logistic latent trait models. Sympson specifies a logistic function for each dimension. Unlike the models of Rasch (1961) and Mulaik (1972), this model yields a separate difficulty value for each dimension rather than a difficulty function. The difficulty is a vector of b_k values which defines a point in the multidimensional space.

The slope of the IRS at the point of inflection for Sympson's model is a function of the item discrimination parameters and the c parameter. Satisfactory procedures for parameter estimation were not found in the literature for this

model. It has been used to generate multidimensional data (e.g., Ackerman, 1987a; Ansley & Forsyth, 1985) which were then analyzed according to another model.

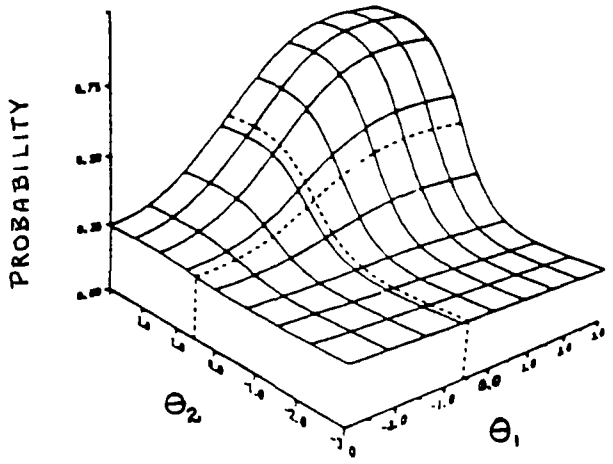


Figure 5. An item response surface for Sympson's model.

Whitely's Model.

In order to use the models described thus far, it is assumed that the dimensions underlying test performance represent unknown hypothetical constructs whose properties are also unknown. Whitely (1980) developed her model from a different perspective of multidimensionality. She proposed that the dimensions be considered as predefined components in a cognitive model of performance. Whitely called the model the multicomponent latent trait model (MLTM) and the form is:

$$P(x = 1 \mid \underline{\theta}, \underline{b}) = \prod_{k=1}^{m} [\exp(\theta_k - b_k)] / [1 + \exp(\theta_k - b_k)], \qquad (7)$$

where $\underline{\theta}$ is a vector of ability parameters; $\underline{b}$ is a vector of difficulty parameters; and m is the number of dimensions.

This noncompensatory model for dichotomous data is a special case of Sympson's model with c = 0 and $a_k = -1/1.7$ for all k. It is therefore another extension of the 1PL model. Whitely (1981) justified the use of one item parameter only by arguing that discrimination and guessing don't usually have direct meaning in terms of information processing although she did insert a guessing parameter (not shown in Formula 7) in the model to accommodate multiple-choice tests (Whitely, 1980, p. 486).

In her model, Whitely unifies information processing and latent trait measurement. She requires separate response variables to measure each latent component, i.e., two kinds of data are required: the responses to the standard psychometric item; and the responses to a series of subtasks that represent an exhaustive set of information processing components. These information processing components must be identified a priori and explicitly postulated in order to estimate the cognitive demands in an item set. Thus, both multidimensionality of ability and multiple component process factors are proposed.

Whitely has done considerable research with this model. She combined Fischer's (1973) Linear Logistic Test Model and her MLTM to create a general multicomponent latent trait model (GLTM) for response processes (Embretson (Whitely), 1984). The GLTM contains both component outcomes and complexity factors in a single model. The multicomponent latent trait models are applicable to complex intelligence test items such as verbal analogies and mathematical reasoning items (Embretson, 1985). Since the model places emphasis on identifying the different cognitive skills required by an item, the application would seem to be limited to data collected under restricted experimental conditions.

Bock and Aitken Model.

This model (Bock & Aitken, 1981) is a multidimensional extension of the two-parameter normal ogive (2PN) model. The form of the model is:

$$P(x=1 \mid \underline{\theta}, \underline{a}, b) = (2\pi)^{-0.5} \int_{-\infty}^{z(\underline{\theta})} \exp(-t^2/2)\, dt, \qquad (8)$$

where $z(\underline{\theta}) = b + \sum_{k=1}^{m} a_k \theta_k$; $\underline{\theta}$ is a vector of ability parameters; $\underline{a}$ is a column vector of discrimination parameters; b is the difficulty parameter; and m is the number of dimensions. The difficulty function for the model defines a hyperplane represented by $z(\underline{\theta}) = 0$. The item discrimination is represented by the $\underline{a}$ vector since the conditional slope of the IRS at points on the difficulty function is given by $(2\pi)^{0.5} a_k$. As noted above, this model is very similar to Samejima's (1974) model.

Bock and Aitken developed a marginal maximum likelihood (MML) estimation procedure for this model which they applied to the Law School Admissions Test data assuming a two-dimensional solution. The solution yielded a good fit to the data. Otherwise, the model does not appear to have been applied or tested extensively.

Hattie's Model.

Hattie's (1981) model is a multidimensional extension of the 3PL model. He developed the model for use in an extensive look at decision criteria for determining dimensionality. The form is:

$$P_{ij}(\underline{\theta}_j) = c_i + (1 - c_i)/\left[1 + \exp\left(-1.7 \sum_{k=1}^{m} a_{ik}(\theta_{jk} - b_i)\right)\right], \qquad (9)$$

where P_{ij} is the probability of being correct for person j on item i; θ_{jk} is the ability

parameter for person j for dimension k, k = 1, 2, ..., m; a_{ik} is the discrimination parameter for item i on dimension k; b_i is the difficulty parameter for item i; and c_i is the pseudo-chance level parameter for item i.

Bogan and Yen Model.

Bogan and Yen (1983) modified Hattie's model to give:

$$P_{ij}(\underline{\theta}_j) = c_i + (1 - c_i)/ [1 + \exp(-1.7 \sum_{k=1}^{m} a_{ik}(\theta_{jk} - b_{ik}))], \qquad (10)$$

where b_{ik} is the difficulty parameter for item i on dimension k and the other parameters remain as above. This change provides a separate b-value for each item on each dimension similar to Sympson's model. Hattie's version has only one b-value per item. Unlike Sympson's model, these last two models are compensatory. Neither model has undergone extensive research or testing, although unpublished research of Huynh & Ansley (1988) did use the Bogan & Yen model to generate two-dimensional data which were analyzed using LOGIST for the purpose of studying grouping effects with chi-square statistics in fitting the 3PL model in IRT settings.

Stegelmann's Model.

Stegelmann (1983) extended Whitely's MLTM model using vector-valued ability parameters which he considered suitable for theoretically handling more complex cognitive processes involving subprocesses. Stegelmann reasoned that it was difficult to identify pure items involving just one psychological component. Components required for each item must be specified a priori. Then only the item response is necessary to calculate the ability score for this model. A disadvantage of the model (see Formula 11) is that it requires an ability parameter for each examinee on each subprocess and application of the model therefore requires

observations of the component events which are almost impossible to obtain.

The form of the model is given by:

$$P(x_{ji} \mid \underline{\theta}_j, \underline{\sigma}_i) = \frac{\left[\prod_{k=1}^{t} \left(\theta_{jk}\, \sigma_{ik} \right)^{b_{ik}}\right]^{x_{ji}}}{1 + \prod_{k=1}^{t} \left(\theta_{jk}\, \sigma_{ik} \right)^{b_{ik}}}, \qquad (11)$$

where $x_{ji} = 1$ for a correct response and 0 for an incorrect response to item i by examinee j, θ_{jk} is the ability of examinee j on component k, σ_{ik} is the difficulty of component k in item i, $b_i = 1$ if a correct answer to item i requires ability on dimension k or 0 otherwise, and t is the number of components.

Other Multidimensional Models and Approaches.

Other researchers have considered various approaches to multidimensional item response models. Mislevy (1983) and Reiser (1983) developed item response models for grouped data in order to determine population parameters without requiring an individual estimate for each examinee. Mislevy (1984, 1985) continued work in this area but extensive work by other researchers has not been found. These types of models are not relevant to the questions being considered in this research.

Andersen (1985) proposed a model for longitudinal latent structure analysis in which the item parameters were considered equal over time but the two measures of $\underline{\theta}$ over time were considered as a two-dimensional latent density. The fit of data to this model has not been good.

Takane and de Leeuw (1986) developed a MIRT model for multicategory data in which both examinees and item categories are represented in a multidimensional Euclidean space. The authors developed a maximum likelihood estimation (MLE)

procedure to determine both the examinee and category points. This procedure has been used on both real and simulated data but the sample sizes were small. Further testing is required.

Stout (1984) described a procedure for assessing dimensionality in which he uses a non-parametric MIRT model which assumes local independence and retains monotonicity of the IRS. The model described is useful if a unidimensional subtest can be established a priori. Stout suggests using the factor loadings of the second extracted factor in factor analysis to select the items for this subtest. The approach appears useful in some situations although this procedure for selection of a unidimensional subtest remains somewhat controversial. No further research was found using this model.

Jannarone (1986) described a conjunctive item response theory for serial process models, i.e., models in which all components are necessary for the process to occur. He described methods in detail for a Conjunctive Rasch Kernel Theory model which allows for local dependence as well, thus making the model potentially more useful for cognitive process measurement. Jannarone identified the following problems with the model. The number of possible conjunctive item parameters can become prohibitively large depending on the number of cross-products allowed between items. The model incorrectly classifies items as conjunctive or additive in too many instances and does not provide for guessing. It has yet to be determined how useful in practice the model will be. Jannarone's work has clarified somewhat the notion of local dependence. If IRT models are no longer required to be locally independent, Jannarone's kernel theory may be useful in construction of more efficient tests and in revealing cognitive structure in large samples.

Kolen and Harris (1987) have proposed a multivariate test theory model based on IRT and generalizability theory. Initial research with this model

indicates it may be useful for studying test equating of alternate forms. More rigorous testing is needed on this model.

Summary and Comparison of MIRT Models.

The models described fall into two basic classes, compensatory and noncompensatory. With one exception, the compensatory models are of the form:

$$P(x=1 \mid \underline{\theta}, \underline{\Omega}) = \int_{-\infty}^{z(\underline{\theta})} \Lambda(u)du \quad , \tag{12}$$

where $\underline{\theta}$ is a vector of person parameters, $\underline{\Omega}$ is a vector of item parameters, $z(\underline{\theta})$ is a linear function, and $\Lambda(u)$ is a twice-differentiable function, usually the normal ogive or logistic function. This class includes the General Rasch (1961) model, Samejima's (1974) model, and the Bock and Aitken (1981) model. Each of these models allows high ability on one dimension to compensate for low ability on another dimension. They all have linear difficulty functions if dichotomously scored items are used and they have conditional slopes at points on the difficulty function that are functions of the corresponding discrimination parameters. Hattie's (1981) model, Bogan and Yen's (1983) model, and the M2PL model of McKinley and Reckase (1983b) are versions of the general Rasch model which make changes in the definitions of difficulty and discrimination but still remain compensatory. Hattie's model has only one difficulty parameter for each item. Bogan and Yen define a difficulty parameter for each item on each dimension. Reckase defines a multidimensional item difficulty (MID) (Reckase, 1985b) and a multidimensional item discrimination (MDISC) (Reckase, 1986) for each item of the M2PL model proposed by McKinley and Reckase (1982a, 1983c).

Mulaik's (1972) model differs from the other models in this class. The ability metric is defined from 0 to $+\infty$. The IRS defined by this model doesn't have a

point or line of inflection and thus the concepts of difficulty and discrimination are less clear.

Noncompensatory models do not allow a high ability on one dimension to compensate for lower ability on a second dimension. The noncompensatory models reviewed include Sympson's (1978) model, Whitely's (1980) model (and its extensions by herself and Stegelmann (1983)), and Jannarone's (1986) conjunctive kernel theory model. The models of Sympson and Whitely take the form:

$$P(x=1 \mid \underline{\theta}, \underline{a}, \underline{b}, c) = c + (1 - c) \prod_{k=1}^{m} P_k(\theta_k) , \qquad (13)$$

where c is a lower asymptote parameter and $P_k(\theta_k)$ is the probability of response with respect to a specific dimension. In Sympson's model the dimensions are hypothetical traits based on commonalities among items. Whitely defines the dimensions as specific cognitive processes required to solve the problem proposed in the item and the lower asymptote for her model is usually set at zero. Some work has been done on the estimation of parameters but no generally accepted algorithm for estimation is known to exist. Hattie (1984) maintains there are serious problems, both theoretically and computationally, with trying to generalize the latent trait model to take account of noncompensatory assumptions. R. L. McKinley (personal communication, November 13, 1986) considers the noncompensatory models to be less tractable than the compensatory models. For the noncompensatory models, the IRS loses the monotonicity characteristic that makes IRT so logically appealing. All of this must be considered when selecting a model for study.

Multidimensional Analysis of Multidimensional Data

Several models have been proposed. However, for many of these models

estimation procedures do not exist or are not well developed. A MIRT model is of limited use unless there is some corresponding estimation procedure available. Until recently only two analysis programs have been readily available for analysis of multidimensional data: TESTFACT (Wilson, Wood, & Gibbons, 1984); and MAXLOG (McKinley & Reckase, 1983c). Neither has undergone exhaustive testing and both have some restrictions in terms of application. The computer costs involved in using the multidimensional analysis programs are high.

Estimates from TESTFACT are made using full-information factor analysis (FIFA) which does not require the computation of correlation coefficients, but operates instead on the set of distinct item response vectors to provide factor scores. Bayesian estimation rather than MLE is used. The discrimination values are not constrained to be positive. They range from -1 to +1 similar to factor loadings. Intuitively, a negative discrimination value makes little sense. The monotonicity of IRT is not retained by the FA approach.

The second program, MAXLOG, performs joint MLE on multidimensional data. Abilities are assumed to be uncorrelated for the model analyzed in this program.

A third program, MIRTE (Carlson, 1987), has recently been written for a compensatory M2PL model. The preliminary results available indicate the estimation procedures, which are patterned after those used in LOGIST, yield more accurate results than MAXLOG.

Five studies have been published in which a multidimensional analysis of multidimensional data has been performed. These studies are listed in Table 3 and are reviewed here.

TESTFACT was used in only one of the five studies. Muraki and Englehard (1985) applied FIFA procedures to both simulated item response data and to real data. The purpose of the study was to demonstrate the properties of expected a

Table 3. Studies Involving Multidimensional Analysis of Multidimensional Data

DIMENSIONALITY DEFINED BY:	AUTHOR (YEAR)	FOCUS of STUDY	DATA INFORMATION	ANALYSIS of DATA
MIRT	Reckase & Ackerman (1986)	Unidimensionality as a composite of abilities	Real: N = 1000; n = 40;	MAXLOG
Factor Analysis	Muraki & Englehard (1985)	Application of FIFA	Simulated M2PN & M3PN; N = 500, 1000, 2000, n = 25, 22	TESTFACT
MIRT	McKinley & Reckase (1984)	Effects of correlated abilities	Simulated N = 2000, n = 50	Item Analysis, PC Analysis; MAXLOG, Correlational Anal
MIRT	McKinley & Reckase (1983a)	Do multidimensional models explain multidimensional data better than unidimensional models?	Simulated M2PL, 1, 2, 3 dimensions, N = 1000; n = 50 Real. N = 1000, n = 10, 15, 30, 1, 2, 3 Subtests	MML for item parameters; MAXLOG; LOGIST
MIRT	McKinley & Reckase (1983d)	Determination of dimensionality of data	Real: N = 2794, n = 50	LOGIST (3PL), MAXLOG (2PL) MAXLOG (M2PL)

posteriori (EAP) scores and to determine whether these EAP scores are preferable to maximum likelihood estimates or to factor loadings from FA of tetrachoric correlations.

The simulated data were generated to fit a multidimensional two-parameter normal ogive model and a multidimensional three-parameter normal ogive model, both MIRT models. The number of dimensions was set at two and sample sizes ranged from 500 to 2000 examinees for tests of 22 or 25 items. The authors reported that FIFA recovered the original factor loadings in all cases better than FA of tetrachoric correlations and that FIFA gave better estimates for items that were very easy or very difficult. Estimating the parameter for guessing caused the other parameter estimates for difficulty and discrimination to be less stable. Muraki and Englehard concluded that MLE was less reliable than EAP and, since the EAP scores seemed to be orthogonal, they were more easily interpreted. This last conclusion would seem to indicate either that orthogonality of scores should be forced to aid in interpretation or that some other analysis of the same data might not provide orthogonal scores. In either case, the estimation procedures should not determine whether or not orthogonality exists. The relationship between MIRT and linear factor analytic procedures is not clear. Researchers (e.g., Ansley, 1984; Green, Lissitz,& Mulaik, 1977; Hambleton, 1980; Tatsuoka & Tatsuoka, 1982) have agreed that linear factor analysis and item response theory do not share a common definition of dimensionality. For this reason, it is difficult to generalize the results of this study. As mentioned earlier, McDonald (1982) has shown a similarity between nonlinear FA and MIRT.

McKinley and Reckase (1983a) undertook a large scale study (by multidimensional standards) using both simulated and real data in order to determine whether a multidimensional two-parameter logistic (M2PL) model more adequately explained multidimensional data than did a 2PL model and whether the

results of M2PL analysis were consistent with the results of FA. Simulated data were generated to fit a M2PL model for 1000 examinees and 50 items. Data sets were generated to satisfy one, two, and three underlying dimensions. All three data sets were then analyzed using LOGIST to estimate unidimensional item and ability parameters, using MAXLOG to estimate multidimensional ability parameters, and using marginal maximum likelihood (MML) procedures similar to those used by Bock and Aitkin (1981) to estimate multidimensional item parameters. (MML was used for the estimation of the item parameters because it produces more stable estimates.) The authors then compared the fit of the M2PL model and the 2PL model to the data.

Similar analyses were performed on real data from Form 16 of the Texas Grammar, Spelling and Punctuation Test (from the University of Texas, 1978). Three data sets were established: a unidimensional spelling test of 30 items for 1000 examinees; a two-dimensional spelling/grammar test consisting of 15 spelling items and 15 grammar items for 1000 examinees; and a three-dimensional spelling/grammar/punctuation test with 10 items for each dimension and 1000 examinees. Items were selected for inclusion in the subtests on the basis of having high factor loadings on the first factor. While the test was considered multidimensional, the items themselves were not considered multidimensional.

For the unidimensional simulated data, the difficulty parameter was recovered more accurately than the discrimination parameter by both LOGIST and MAXLOG ($r = 0.99$ between difficulty parameters and estimates for both). Unidimensional data were recovered well by both programs (LOGIST and MAXLOG) which is not surprising as the models are the same when there is only one dimension. The authors reported most of the differences between the two sets of estimates would be eliminated by rescaling.

For the two-dimensional simulated data, the multidimensional parameter

estimates from MAXLOG were highly correlated with the true parameters ($r > 0.98$ for item parameters and $r > 0.91$ for ability parameters). The LOGIST unidimensional estimates were not as strongly related to the true parameters with the exception of the difficulty parameter which was well recovered in the unidimensional analysis ($r = 0.98$ for difficulty; $r = 0.47$ for discrimination; $r = 0.68$ for abilities).

Results for the three-dimensional simulated data were similar. MAXLOG recovered good estimates of item parameters ($r > 0.91$ for difficulty and discrimination) and ability parameters ($r > 0.90$ for the three ability dimensions) while LOGIST accurately recovered only the item difficulty parameter ($r = 0.99$).

An analysis of variance was performed on the mean absolute deviation statistics with dimensionality of data (1, 2, or 3 dimensions) and the model (2PL or M2PL) as independent variables. Dimensionality, model, and their interactions were all found to be significant ($p \leq 0.01$). As the dimensionality of the data increased, the advantage gained by using a multidimensional model increased. Regardless of the dimensionality, the MAXLOG multidimensional estimation was better than the unidimensional estimation.

In the analysis of the real data, similar results were reported in that the multidimensional model fit the multidimensional data significantly better than the unidimensional model ($p \leq 0.01$). For the one-subtest real data, the fit of the 2PL model to the data was significantly better than the fit of the M2PL model ($p \leq 0.01$). McKinley and Reckase attributed this to the estimation procedure for the 2PL model perhaps being more robust than the M2PL model to violations of assumptions of the model found in the real data.

In summary, McKinley and Reckase reported that a multidimensional model more adequately fits simulated and real multidimensional response data than a unidimensional model and that the parameters of a multidimensional model can be

accurately estimated. These results are based on one replication only of the simulated data and on the factor analytic approach to determining dimensionality for the real data. It is also noteworthy that the multidimensional tests were composed of "pure" items. That is, the multidimensional tests could have been partitioned into unidimensional subtests. The results may have been different if multidimensional items had been used. However, the results do provide some evidence that MIRT analysis is preferable for multidimensional data.

McKinley and Reckase (1983d) analyzed real test data, structure unknown to the authors, assuming unidimensional and multidimensional IRT models. The unidimensional models used were the 2PL and the 3PL models (analyzed with LOGIST) and a M2PL model was used for the MIRT analysis (MAXLOG). The data were 2794 response strings to 50 items taken from several subtests of a standardized test (Iowa Test of Educational Development). The design for the analysis of the data was to begin with a unidimensional model, evaluate the fit of the model to the data, and then to increase the dimensionality of the model and perform the same analyses. The dimensionality of the model was increased until deviations from fit (as measured using the distribution of the residual covariance matrix) were acceptably small. Once the authors determined acceptable fit, the item parameter estimates were analyzed to determine the structure of the ability components required by the test. The authors expected to find two or more relatively distinct ability dimensions and to be able to classify items on the test into content categories based on which dimension was required for a correct response. However, the results were not as expected. Rather than distinct ability dimensions, the analysis revealed two highly correlated dimensions. Whereas they had expected M2PL to fit the data better than the unidimensional models, this was also not the case. Although the authors did not know the true dimensionality of the data, they suggested that it was predominantly unidimensional. The authors

suggested that the data appeared unidimensional because of the extreme item difficulties, i.e., a difficulty dimension was the dominant dimension.

Reckase (1985b) updated the McKinley and Reckase (1982a) M2PL model to define a multidimensional item difficulty (MID) to be composed of two parts: D, the distance from the origin of the θ-space to the point of maximum discrimination; and α_k, the angle the distance vector makes with the θ_k-axis. Items with the same MID values are considered to be unidimensional in the sense that examinees require the same composite of underlying abilities in order to respond correctly. Reckase and Ackerman (1986) then considered the relationship between the concept of unidimensionality and the direction of an item in a multidimensional space. The authors hypothesized that if items with equivalent MID directions were combined to form a test, the resulting test would meet the IRT requirement of unidimensionality (i.e., that for a given item, all persons with the same estimate of ability have the same probability of a correct response).

Real data from the ACT Mathematics Usage Test (40 items) for 1000 examinees were analyzed using MAXLOG for the M2PL model (two dimensions were specified). The MID estimates were computed for each item and the items were sorted according to direction. Using this information, four item sets were identified and analyzed using LOGIST (3PL) to determine whether or not each of the sets could be considered unidimensional. The results of the study tended to support the conception of unidimensionality suggested by a common direction in the multidimensional space for a set of items and the use of multidimensional difficulty statistics in forming unidimensional item sets. The study should be repeated with a larger number of items in the subtests so that more definitive conclusions could be reached.

The most relevant of the multidimensional studies to this research was that of McKinley and Reckase (1984). The purpose of this study was to assess the

effects of correlated abilities on test characteristics and to explore the effects of correlated abilities using a MIRT model which does not explicitly account for this correlation.

Two simulated tests were constructed. The first test was composed of two relatively unidimensional subtests. The items for the second test were all two-dimensional items. Each test had 50 items. For each test response, data were generated based on the M2PL model of McKinley and Reckase (1982a) using four groups of examinees which differed in the degree of inter-dimension correlation. There were eight data sets in all: two types of test (two unidimensional subtests, two-dimensional items) by four degrees of correlation between dimensions (0.00, 0.35, 0.50, 0.70). For each data set, two true abilities were selected randomly from a bivariate normal distribution (both means equal to 0.0, both standard deviations equal to 0.50) for 2000 simulated subjects such that the ability dimensions had the appropriate correlation.

Four types of analyses were run on each of the data sets, the most relevant to this research being a M2PL analysis using MAXLOG and a correlational analysis of true and estimated parameters. The results of the MIRT analysis differed depending on whether the test was composed of two unidimensional subtests or of two-dimensional items.

In all cases, the correlations between the generated abilities were very close to the specified true population values. For all eight data sets the difficulty parameter, d_j, was well recovered. (The parameter d_j should not be confused with the multidimensional difficulty distance, D.) It appears to be the most stable of the parameters.

When the test was composed of two unidimensional subtests, McKinley and Reckase found the true parameters were better recovered when the abilities were less correlated. The correlation between the estimates of the two ability

dimensions was not recovered. In all four data sets the correlation between the two ability parameter estimates was very close to zero. This would indicate that MAXLOG tended to force orthogonality of the estimated ability dimensions. When the ability dimensions were uncorrelated, the discrimination parameter estimates were fairly accurate. The discrimination parameters were only moderately well recovered when the correlation between the ability dimensions was high.

When the items on the test were themselves multidimensional, MAXLOG recovered the parameters less well than for the subtest data. The MAXLOG ability estimates were negatively correlated and this negative correlation increased in magnitude as the true correlation decreased. There seems to be no problem generating multidimensional data to satisfy a specified correlation but there are problems in recovering this correlation with MAXLOG. The discrimination parameters for these four data sets were less well recovered when the ability dimensions were not correlated than when they were correlated 0.70, although the discrimination indices were not well recovered in any of the cases.

McKinley and Reckase concluded that the latent structure of response data involved two concepts: latent item structure which refers to the number and interrelationships of the dimensions required for performance on the items (i.e., unidimensional items or multidimensional items); and the latent ability structure of an examinee which refers to the number and interrelationships of the dimensions underlying the examinee's responses. The correlation between these underlying ability dimensions also appears to be important in determining the latent ability structure. The most important result is that true dimensions are not well estimated when the dimensions are correlated or when the items themselves are multidimensional.

Multidimensional analysis of multidimensional data would appear from the results of these studies to be preferable to unidimensional analysis of the same

data. There are some problems associated with multidimensional analysis. For example, the inclusion of a pseudo-chance level parameter in the model affects the stability of the other item parameter estimates. Also, the estimation procedures may tend to force orthogonality. As the correlation between dimensions increases, the true parameters appear to be less accurately recovered. When the items themselves are multidimensional, this result is even more pronounced.

Summary and Research Problem

In IRT the observed item response patterns of examinees are used to obtain estimates of the unobserved latent abilities which account for performance on the items. The theory is focussed at the item level rather than the test level making it particularly useful in measurement situations such as multi-level testing, tailored testing, or item bias detection.

The original IRT models were based on the assumptions of local independence and unidimensionality. The unidimensionality assumption has been found to be a limiting aspect of the theory. Few real life situations are unidimensional in the sense that only one underlying trait accounts for correct performance. Yet the models and the estimation procedures forced this restriction on the data. Efforts have been made to identify measures of the dimensionality of data and to partition multidimensional tests into unidimensional subtests. Neither has been particularly successful. MIRT models have been proposed. However, for some time, estimation procedures were not available to analyze the multidimensional data. Unidimensional analysis was applied to known multidimensional data in efforts to determine whether multidimensional estimation procedures were required. In general, the unidimensional models were not found to be robust to multidimensional data, whether real or simulated.

Estimation procedures for analysis of multidimensional data were

introduced within the last five years. Only two multidimensional analysis programs have been readily available: TESTFACT and MAXLOG. Both programs require extensive amounts of computer time and large numbers of examinees. TESTFACT will not accommodate large numbers of items and does not retain the monotonicity of the IRS. MAXLOG requires large numbers of items in order to make stable parameter estimates and does not recover the true dimensions when the underlying abilities are correlated. Preliminary results from a third analysis program, MIRTE, indicate the program is more efficient and more accurate in its parameter estimations. Whether it can accommodate correlated dimensions has still not been determined.

The results of the research involving multidimensional analysis of multidimensional data are encouraging. Even given the imperfections of the multidimensional estimation procedures, multidimensional analysis provided better estimates for multidimensional data than a unidimensional analysis of the same data. No replications were made in any of the studies reported involving multidimensional analysis. In an exploratory sense of understanding the models and the estimation procedures this is acceptable. In order to make conclusions with some confidence, replications of these studies are required.

The effects of using data sets with a narrower range and a lower mean on one ability have not been investigated adequately. McCauley and Mendoza (1985), in a study of identification of item bias, generated two-dimensional data for items which required a secondary ability on which their two groups of examinees differed in mean level. However, the data were generated to conform to a specific factor structure and the analysis was done with LOGIST assuming a unidimensional IRT model. A factor analysis dimension and an IRT dimension are not necessarily equivalent. It becomes questionable then whether data generated to fit a specified factor structure should reflect the same structure when analyzed

according to an IRT model particularly when the IRT model used is unidimensional and the data are expected to be multidimensional. In any case, the results of this study do not assist in prediction of the outcome if the data were to undergo a multidimensional analysis.

Only McKinley and Reckase (1984) tested the effects of correlated abilities in a multidimensional analysis. Generating correlated data to fit the M2PL model was successful. Recovering the correlation between the two underlying dimensions was not successful. When the test was composed of two unidimensional subtests, the individual dimensions were recovered moderately well by MAXLOG but the correlation between the dimensions was not recovered. The estimation procedures tended to force the underlying dimensions to be orthogonal. When the items themselves were multidimensional, neither the ability dimensions nor the correlation between them was well recovered. The item difficulty parameter was well recovered in all cases. The item discrimination parameters were not well recovered when the ability dimensions were highly correlated. There appeared to be a confounding of item structure with ability structure.

The effects of both a restricted secondary ability and correlated abilities on parameter estimation need to be evaluated in a comprehensive, systematic manner. It is also important to determine the stability of the parameter estimates in multidimensional analysis procedures.

In summary, the purpose of this study is to determine the adequacy of two-dimensional ability and item parameter estimates using a MIRT analysis under the conditions of a differentiated ability on the second dimension and different degrees of correlation between the two latent abilities.

Chapter 3

METHODOLOGY

This chapter is divided into three sections: model and analysis program selection; research design and description of data sets; and procedures for generation and analysis of data.

Selection of a MIRT Model and Method of Parameter Estimation

The results of this study are dependent on the model selected. Selection of a MIRT model for a particular situation is not an easy task. Several issues need to be considered, both psychological and practical.

Model selection should be based initially on whether the model can adequately describe the interaction between a person and an item. The choice of a compensatory or a noncompensatory model should be made as much as possible on the extent to which the model represents the testing situation.

There are practical issues to be considered as well, such as the number of item parameters used, the estimation procedure, and the availability of an analysis program. In addition, efficiency of analysis and cost must remain factors in the decision. Most of the research involving the number of item parameters used (1, 2 or 3) has been done using unidimensional models but many of the results extend to the logic of a multidimensional space such as poor fit of a one-parameter model and the problems in estimating a pseudo-chance level parameter.

McKinley and Reckase (1983a) found that their M2PL model with item parameters for difficulty and discrimination explained multidimensional data better than a unidimensional model although correlations between the underlying abilities weren't recovered by their analysis program (McKinley & Reckase, 1984).

Muraki and Englehard (1985) used the multidimensional two-, and three-parameter normal ogive models for generation of data but analyzed the data using a linear factor analysis procedure. The inclusion of the guessing parameter resulted in less stable estimates for the other parameters.

If a multidimensional analysis is desired, estimation procedures are not yet well developed for multidimensional analysis with noncompensatory models. Estimation procedures are better developed for the compensatory models. Maximum likelihood estimation (MLE) is the most widely used procedure for estimation of the parameters in IRT. Maximum likelihood estimators are the values of the parameter estimates that make an observed data set appear most likely given the particular model. As the number of items or persons increases, the maximum likelihood estimators converge (in probability) to the parameters. The Newton-Raphson procedure (suggested by Bock & Lieberman (1970) and Bock (1972)) is an iterative method for determining the minimum of a nonlinear equation and it is used in many maximum likelihood procedures.

A third practical problem to be considered is the availability of a computer program to analyze the multidimensional data. Without an analysis program, multidimensional parameter estimation is theoretical. Few of these exist and none has undergone exhaustive testing. McKinley and Reckase (1983c) developed MAXLOG, a program in FORTRAN which performs joint maximum likelihood estimation on multidimensional data. In this program, abilities are considered to be uncorrelated. The TESTFACT program (Wilson, Wood, & Gibbons, 1984) performs full-information factor analysis. Bayesian estimation rather than MLE is used in TESTFACT. The discrimination values aren't constrained by the program to be above zero. They range from -1 to +1 similar to factor loadings. Intuitively, a negative discrimination value makes little sense. A third program, MIRTE

(Carlson, 1987), has recently been developed for analysis of data which fit the M2PL model of McKinley and Reckase (1982a).

Model selection for use with multidimensional data is hampered at this point in time by the complexity of the models available, the restriction of some models to specific situations, and the lack of computer programs to analyze the data.

Yen (1980) pointed out that use of an inappropriate model can introduce different types of errors in different situations. As a primary objective in this study is to determine the consequences of restricted range of a secondary required ability on parameter estimation, the inclusion of a pseudo-chance level parameter may weaken the internal validity of the study. Since the aim is to determine the accuracy of the ability and item parameter estimates, it would be appropriate to avoid model effects when possible. The inclusion of a c-parameter has been found to affect standard error of ability estimates and estimates of discrimination and difficulty. Therefore, a two-parameter model seemed most appropriate.

The M2PL model of McKinley and Reckase (1982a) was used for this study for the following reasons: (a) no pseudo-chance level item parameter is included; (b) estimation procedures are well developed for this model and an analysis program (MIRTE) exists to perform these estimations; and (c) it has been demonstrated in previous research that the model does reflect multidimensional data particularly well in the case of a two-dimensional ability space which is the scenario for this study.

The M2PL model proposed by McKinley and Reckase (1982a) was updated by Reckase (1985b, 1986). Henceforth, in this thesis, M2PL will be used to refer specifically to this multidimensional two-parameter logistic model rather than a family of multidimensional two-parameter logistic models.

The mathematical formula (given previously in Formula 3) for the M2PL model is repeated here:

$$P_{ij} = P(x_{ij} = 1 \mid \underline{a}_i, d_i, \underline{\theta}_j) = \frac{\exp(\underline{a}_i'\underline{\theta}_j + d_i)}{1 + \exp(\underline{a}_i'\underline{\theta}_j + d_i)}, \qquad (14)$$

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

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 on item i; $\underline{a}_i$ is a vector of m discrimination parameters; d_i is a parameter representing the difficulty of item i; $\underline{\theta}_j$ is a vector of m ability parameters for individual j; N is the number of examinees; n is the number of items; and m is the number of dimensions.

This model is compensatory in that it allows high proficiency on one dimension to compensate for low proficiency on other dimensions in arriving at a correct response to a test item. For a two-dimensional model the mathematical function becomes:

$$P_{ij} = \frac{e^{(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i)}}{1 + e^{(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i)}}, \qquad (15)$$

If an item is pure on dimension one, then the probability of a correct response doesn't depend on ability in dimension two as a_{2i} would be zero. However, for items requiring some amount of the second ability in order to respond correctly, the probability of a correct response does increase as θ_2 increases. This aspect of the model suits the proposed educational situation described earlier. For the ESL

students who may not have the necessary verbal ability (measured by θ_2) the probability of a correct response is smaller, even though the item may be designed and used to measure primarily mathematics ability (measured by θ_1). The compensatory model fits the situation in that it more realistically models the testing scenario where students will use whatever abilities they have in an attempt to respond correctly. As well the model fits the data in that there are varying degrees of the need for the second (verbal) ability.

Reckase (1985b) updated the M2PL model by defining a multidimensional item difficulty parameter, MID, composed of a distance D_i and an angle, α_{ik}, such that

$$D_i = -d_i / \left[\sum_{k=1}^{m} (a_{ik})^2 \right]^{0.5} \qquad (16)$$

This parameter represents the distance between the origin of the m-dimensional ability space and the point in the space where the item information is a maximum. The line joining this point to the origin is at an angle of α_{ik} to the k^{th} ability dimension where

$$\cos \alpha_{ik} = a_{ik} / \left[\sum_{k=1}^{m} (a_{ik})^2 \right]^{0.5} \qquad (17)$$

The size of the angle, α_{ik}, indicates to what extent ability on the kth dimension is required to respond correctly to the item. An item with $\alpha_1 < \alpha_2$ would require more of ability from the first dimension in order to respond correctly. The smaller the size of α_1, the closer the MID vector is to the θ_1 axis. The extent of the recovery of the angles is indicative then of how well the latent space is being recovered.

Reckase (1986) also defined a multidimensional discrimination parameter for

item i to be

$$MDISC_i = \left[\sum_{k=1}^{m} (a_{ik})^2 \right]^{0.5} \qquad (18)$$

This parameter is related to the item characteristic curve on the multidimensional item response surface above the line through the origin of the ability space and to the point of maximum information. It is proportional to the slope of that curve at the point of steepest slope and is therefore analogous to the unidimensional discrimination parameter (Carlson, 1987). Reckase (personal communication, December, 1986) suggested defining the latent space by selecting values for MDISC and MID (i.e., D and α) and determining the other parameters from these.

The analysis program selected to perform the estimation procedure is MIRTE (Carlson, 1987). This program provides estimates of item parameters and examinee abilities for the M2PL model. The method of estimation used is a variation of the joint maximum likelihood procedure using a modified Newton-Raphson iteration technique and the algorithm used is similar to that used in LOGIST. MIRTE was selected over MAXLOG because results of pilot testing with MIRTE have provided more accurate estimates more efficiently. MIRTE has been used in one recent study (Ackerman, 1987b) to estimate item parameters. The MIRTE version 2.00 used in this study was found to estimate parameters when dimensions were correlated better than MAXLOG (J. E. Carlson, personal communication, December, 1987).

The response to each item, conditional on the ability space, is assumed to be independent of that to every other item. Thus the likelihood of each n-element response vector $\underline{x}_j$ can be expressed as a product of probabilities:

$$\prod_{i=1}^{n} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}}, \text{ where } Q_{ij} = 1 - P_{ij} \tag{19}$$

Since the assumption is made that examinees respond to the items independently of each other, then the likelihood of a set of N response vectors is a product of probabilities, L:

$$L = \prod_{j=1}^{N} \prod_{i=1}^{n} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \tag{20}$$

Maximum likelihood estimates of the parameters of the model are those values that maximize the likelihood function, L, or equivalently, minimize F, where

$$F = - \ln L \tag{21}$$

In order to obtain estimates which minimize F, the partial derivative of F with respect to each parameter ($\underline{a}_i$, d_i, $\underline{\theta}_j$) is set to zero and solved. To ensure the existence of a minimum, the matrix of second partial derivatives of F with respect to each pair of parameters must be positive definite. Since the equations to be solved are nonlinear and there is no direct solution, the iterative Newton-Raphson procedure is used. (See Carlson (1987) for an explanation of the Newton-Raphson procedure in this program.)

The data for any examinees having perfect or zero number-right scores are eliminated before estimation procedures begin. These cases are identified in the output.

Initial estimates are required for both item and ability parameters before the iterations can proceed. Initial estimates for the discrimination parameters, $\underline{a}_i$, are supplied by the user and can be any positive number less than the user definable maximum value for the discrimination parameters. The initial difficulty

parameters are computed by the program and are a logarithmic function of the odds of being correct for each item. The initial difficulty estimates are transformed to a mean of zero and a standard deviation of two before iterations begin.

Initial ability estimates are also computed for each examinee as the sum of all the initial discrimination parameter estimates for those items which the examinee answered correctly. These values are scaled to a mean of zero and a standard deviation of one.

It is necessary to constrain the estimates at each stage in the MIRTE estimation program to prevent their drifting to unreasonably large or small values. Maximum and minimum values of ability estimates are specified by the user. These limits are also the maximum and minimum estimates for the difficulty estimates. The discrimination parameter estimates are constrained to a lower bound of 0.01 (to prevent all the discrimination parameter estimates becoming zero for an item, which would result in a totally nondiscriminating item). The upper bound for the discrimination parameter estimates can be specified by the user (although default values are provided for all user specified bounds).

In an iteration, if an estimate exceeds the upper or lower limit, the estimate is set to the limit value. Limits are set on all estimates before they are rescaled. Rescaling of all parameter estimates takes place after each step within the MIRTE estimation to avoid problems of indeterminacy. In the output provided by MIRTE, each estimate that failed to converge or was set to a limit is indicated.

As well as estimation of ability, item discrimination and difficulty, MIRTE provides estimates of standard errors for each of the parameter estimates. Estimates of the multidimensional item difficulty and discrimination are also provided.

Research Design and Data Description

This is a Monte Carlo study in that simulated abilities and item responses were generated to fit the M2PL model (assuming two ability dimensions) under specified conditions for two-dimensional items. The item responses were then analyzed using MIRTE (Carlson, 1987) and estimates were made of the underlying examinee abilities and item parameters.

The design was such that each of the possible combinations of three degrees of correlation and two ability vector distributions is tested. A notable difference of this study over previous studies which used simulated data is that the six different data sets and analyses were replicated 100 times. Previous research on simulated data has results from only one set of data or very few replications. This makes the interpretations and conclusions tenuous at best. The main advantage of the Monte Carlo procedure (i.e., the replications) in research using simulated data is the increased confidence with which the results can be held; the main disadvantages are the high cost of computer analysis, the possibility of not reflecting reality, and that the results are not generalizable beyond the specific conditions simulated. A sufficient number of replications is required if the level of precision is to be adequate. Given the amount of computer time required and the advice of J. Carlson, T. Ackerman, and M. Reckase (personal communication, December, 1986), 100 replications were judged sufficient to provide an adequate yet rigorous test of the program and design.

In this study two types of data were manipulated, the ability vectors for the examinees and the item parameters. The ability vectors are described first.

A two-dimensional ability vector, (θ_1, θ_2), was randomly generated for each examinee using the IMSL (1979) subroutine GGNSM. The different ability distributions for each of the six data sets are listed in Table 4. For the first three

data sets, both ability dimensions were normally distributed with mean of zero and standard deviation of one. The only difference among these three data sets (labeled A1, A2, A3) was the degree of correlation specified between θ_1 and θ_2. For Data Set A1, the two ability dimensions were uncorrelated; for Data Set A2, the correlation between θ_1 and θ_2 was 0.25; and for Data Set A3, θ_1 and θ_2 were correlated 0.50. The zero correlation was chosen to provide a situation in which the latent abilities were unrelated. The 0.25 correlation reflects the situation in which the two latent abilities have very little shared variance. An upper bound of 0.50 level of correlation was set as results of previous researchers have indicated a unidimensional analysis can be performed on multidimensional data correlated above 0.50 (Doody, 1985; Drasgow & Parsons, 1983; Hattie, 1984).

The three data sets labeled B1, B2, and B3 differed from the first data sets only in the distribution of θ_2. The second ability dimension, θ_2, was generated to be normally distributed with a mean of -1 and a standard deviation of 0.67. This restricted the range on the second ability dimension to approximately [-3, +1].

Table 4. True Ability Distributions for the Data Sets

Data Set	Distribution of θ_1	Distribution of θ_2	Approximate Range of θ_1	Approximate Range of θ_2	$\rho(\theta_1,\theta_2)$
A1	N(0,1)	N(0,1)	[-3,+3]	[-3,+3]	0.00
A2	N(0,1)	N(0,1)	[-3,+3]	[-3,+3]	0.25
A3	N(0,1)	N(0,1)	[-3,+3]	[-3,+3]	0.50
B1	N(0,1)	N(-1,0.45)	[-3,+3]	[-3,+1]	0.00
B2	N(0,1)	N(-1,0.45)	[-3,+3]	[-3,+1]	0.25
B3	N(0,1)	N(-1,0.45)	[-3,+3]	[-3,+1]	0.50

The item parameters (listed in Table 5) for the simulated test were set in advance and were chosen so that the test would be representative of a test requiring primarily ability from dimension one but also varying amounts of the

Table 5. True Item Parameters for the 104 Items

α_{11}	D_i	d_i	Item	MDISC	a_{11}	a_{12}	Item	MDISC	a_{11}	a_{12}
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°	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°	-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	0.518	66	1.70	1.642	0.44
15°	2.5	-5	15	2.00	1.932	0.518	67	1.70	1.642	0.44
15°	2.0	-4	16	2.00	1.932	0.518	68	1.70	1.642	0.44
15°	1.5	-3	17	2.00	1.932	0.518	69	1.70	1.642	0.44
15°	1.0	-2	18	2.00	1.932	0.518	70	1.70	1.642	0.44
15°	0.5	-1	19	2.00	1.932	0.518	71	1.70	1.642	0.44
15°	0.0	0	20	2.00	1.932	0.518	72	1.70	1.642	0.44
15°	-0.5	1	21	2.00	1.932	0.518	73	1.70	1.642	0.44
15°	-1.0	2	22	2.00	1.932	0.518	74	1.70	1.642	0.44
15°	-1.5	3	23	2.00	1.932	0.518	75	1.70	1.642	0.44
15°	-2.0	4	24	2.00	1.932	0.518	76	1.70	1.642	0.44
15°	-2.5	5	25	2.00	1.932	0.518	77	1.70	1.642	0.44
15°	-3.0	6	26	2.00	1.932	0.518	78	1.70	1.642	0.44
30°	3.0	-6	27	2.00	1.732	1.00	79	1.70	1.472	0.85
30°	2.5	-5	28	2.00	1.732	1.00	80	1.70	1.472	0.85
30°	2.0	-4	29	2.00	1.732	1.00	81	1.70	1.472	0.85
30°	1.5	-3	30	2.00	1.732	1.00	82	1.70	1.472	0.85
30°	1.0	-2	31	2.00	1.732	1.00	83	1.70	1.472	0.85
30°	0.5	-1	32	2.00	1.732	1.00	84	1.70	1.472	0.85
30°	0.0	0	33	2.00	1.732	1.00	85	1.70	1.472	0.85
30°	-0.5	1	34	2.00	1.732	1.00	86	1.70	1.472	0.85
30°	-1.0	2	35	2.00	1.732	1.00	87	1.70	1.472	0.85
30°	-1.5	3	36	2.00	1.732	1.00	88	1.70	1.472	0.85
30°	-2.0	4	37	2.00	1.732	1.00	89	1.70	1.472	0.85
30°	-2.5	5	38	2.00	1.732	1.00	90	1.70	1.472	0.85
30°	-3.0	6	39	2.00	1.732	1.00	91	1.70	1.472	0.85
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°	0.5	-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°	-0.5	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
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

second ability dimension in order to respond correctly. This reflects a situation in which a test designed primarily to measure mathematical ability also requires varying amounts of verbal ability in order to respond correctly to the items. The M2PL model required two discrimination parameters (a_1, a_2) and a parameter representing difficulty (d). As the multidimensional discrimination and the multidimensional difficulty were estimated by MIRTE, the true values for these parameters were also included in Table 5. (True values for MDISC, D, and α were selected initially as suggested by Reckase.) There were thirteen levels of item difficulty, D, covering the full ability range. The test consisted of 104 items, 26 requiring only the first ability for a correct response, 52 items requiring predominantly the first ability, and 26 items requiring equal amounts of both abilities.

The general design of the study involved two stages. In the first stage, three sets of simulated, two-dimensional ability vectors (A1, A2, A3; N = 2000) were generated. In the second stage, the three sets of simulated, two-dimensional ability vectors (B1, B2, B3; N = 2000) were generated to reflect a narrower range and lower mean on the secondary dimension.

The results gathered from analysis within the A data sets were used to determine the effect of different degrees of correlation between θ_1 and θ_2 on both the ability and item parameter estimates. By comparing corresponding data sets (e.g., A1 with B1, A2 with B2, A3 with B3) the effects of narrower range of the second dimension on parameter estimates were determined. Results gathered from analysis within the B data sets were used to determine the combined effects of narrower range and degree of correlation by comparing with the results from the A data sets.

Procedures for Generation and Analysis of Data

A flow chart of the data generation and analysis procedure for any data set is given in Figure 6. Given the item parameters and the ability vectors, the M2PL model was used to simulate item responses (0 if incorrect, 1 if correct) for each of the 2000 examinees to each of the 104 items for each data set. The FORTRAN program used in this simulation was M2PLGEN (Ackerman, 1985). The response vectors were generated by comparing $P(\underline{\theta})$, given the true $\underline{\theta}$ vector for an examinee, with a uniform random number (R) between 0 and 1 generated by the IMSL subroutine GGUBS. (This IMSL generator has been tested and recommended for use in simulation studies (Slaughter & Delucchi, 1986).) If $P(\underline{\theta}) > R$, then the item response is a 1; if $P(\underline{\theta}) < R$, then the item response is a zero. This was repeated 104 times to obtain a response for each item for each of the 2000 examinees. These item response vectors were then analyzed using MIRTE and estimates of both ability dimensions and the item parameters were made.

There were six data sets each replicated 100 times. The item parameters were the same for each data set as were the initial discrimination parameter estimates for the MIRTE analysis. The data sets differed in the distribution of the ability parameters (2 variations) and degree of correlation (3 different values) between the ability dimensions.

The six data sets were generated and analyzed independently. This procedure was as follows.

A random seed was provided, the item parameters were specified and the two ability vectors were generated. (In this case the random seed was a number from a lottery ticket.) The item responses were generated according to the M2PL model representing two-dimensional, two-parameter logistic data.

The response vectors were analyzed with MIRTE and estimates were

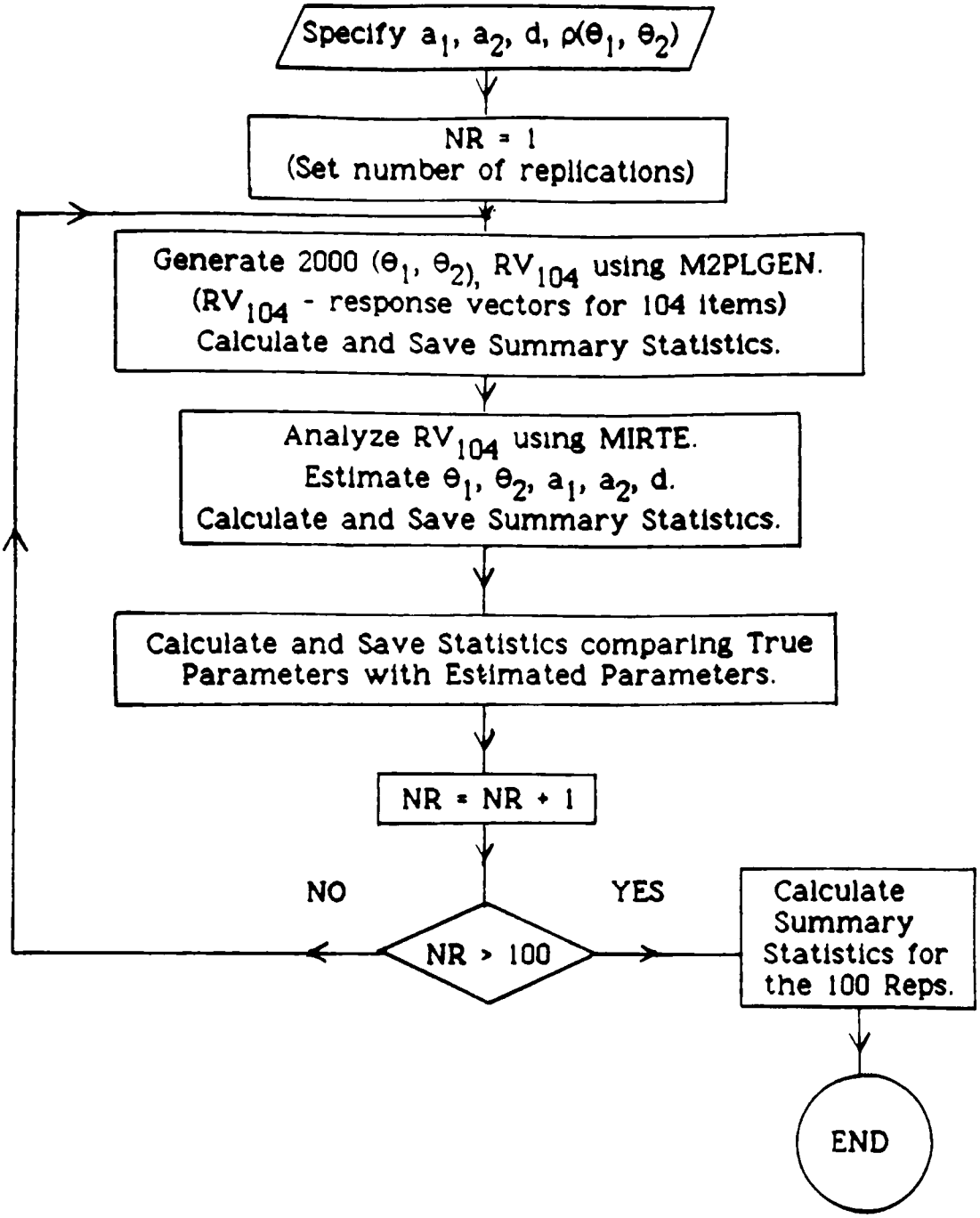


Figure 6. Flow chart of the procedure used to generate and analyze each data set.

obtained for θ_1, θ_2, d_i, a_{i1}, a_{i2}, $MDISC_i$, D_i, α_{i1}, and α_{i2}.

Descriptive statistics (mean, standard deviation, maximum, minimum) were calculated on the original parameters as well as their estimates. The average absolute deviation (AAD) of the estimate from the true value was calculated for the estimates of both abilities and item parameters, a_{i1}, a_{i2}, d_i. A correlation analysis was done on the variables. The variables in the item correlation matrix were NC (number correct), d, $\hat{d}$, a_1, $\hat{a}_1$, a_2, $\hat{a}_2$, D, $\hat{D}$, MDISC, $\widehat{MDISC}$, α_1, and $\hat{\alpha}_1$. The variables in the ability correlation matrix were θ_1, θ_2, $\hat{\theta}_1$, $\hat{\theta}_2$, and RS (raw score).

The random seed was incremented by two and the procedure was repeated, i.e., another 2000 ability vectors and response vectors for the 104 items were generated and analyzed. This process was replicated 100 times for each of the six data sets. In the data analysis, the MIRTE program doesn't always identify dimensions one and two correctly. In order to avoid confusing the dimensions during the 100 replications, a check was made during each analysis on the first thirteen item discrimination parameter estimates. These items were pure on θ_1. If the sum of the first thirteen a_1 estimates was less than the sum of the first thirteen a_2 estimates, the estimations for the dimensions were assumed to be interchanged and were then switched before continuing.

Descriptive statistics (mean, standard deviation, maximum, and minimum) were calculated on the 100 sets of summary statistics of each replication for each of the six data sets. The final statistics reported in Chapter 4 represent the means from the 100 replications of each data set. These statistics were used when comparing results between data sets.

Each job of 100 replications required 45,000-50,000 CPU seconds. The jobs were run in batch on an Amdahl 5880 processor with 64 megabytes of main memory using the VM/HPO operating system. Results appear in the next chapter.

Chapter 4

RESULTS AND DISCUSSION

This chapter is divided into five parts. In the first part results are presented for the ability vectors. The second part contains results concerning the item parameters. A discussion of the research questions is followed by the strengths and limitations of the research. Suggestions for future research conclude the chapter.

Generation and Recovery of Ability Dimensions

Two areas must be addressed in this section. It is necessary to determine how well the specified ability dimensions were generated and how well they were then estimated from the response vectors. The statistics given in the following tables are the mean values of the corresponding statistics determined for each of the 100 replications in each data set.

In the 600 replications there were no response vectors generated in which there were any perfect scores. There were 100 replications which reported raw scores of zero, most of these occurring in the A data sets with an increasing number as the correlation between the ability dimensions increased. (In data set A1, 9 replications reported zero raw scores; 28 in data set A2; 46 in data set A3; 3 in data set B1; 2 in data set B2; and 12 in data set B3.) Investigation of the A runs did not provide any rational explanation for this. More raw scores of zero in the A runs is contrary to what would be expected. The mean raw score for each of the A data sets was approximately 52 (Table 6) indicating an average conventional difficulty value of $p = 0.5$. The inclusion of raw scores of zero did not seem to affect this outcome. For the B data sets the raw score mean dropped to

approximately 47. This was expected as the range of θ_2 decreased and there was less 'ability' in the sample.

In agreement with the findings of Greaud (1988) the raw score appeared to be unaffected by changes in degree of correlation between θ_1 and θ_2. It was affected by the differentiated ability on the second dimension even though the items measured predominantly θ_1. Increasing the correlation between the ability dimensions resulted in an increase in the variance of the raw scores for both the A and B data sets, although the increase was greater in the A data sets. This may partially be attributed to the compensatory model and to the increase in inter-item covariance as $\rho(\theta_1,\theta_2)$ increases.

Table 6. Means of the Mean and Standard Deviation of the Raw Score (over 100 replications)

	Data Set					
	A1	A2	A3	B1	B2	B3
Mean	51.97	51.94	51.95	46.63	46.65	46.69
Std.Dev.	14.22	15.33	16.33	13.64	14.43	15.12

Generation of (θ_1,θ_2) Vectors.

The ability data in all six data sets were generated to fit the specifications stated. (Table 7 lists the means of the correlations, means, standard deviations and approximate ranges of the generated and estimated theta variables for all of the data sets.) The correlation between θ_1 and θ_2 for data generated over the 100 replications was recovered within 0.001 of the specified correlation. The means for θ_1 and θ_2 were all within 0.004 of those specified. There was very small variance (< 0.005) for these means and standard deviations in all data sets. There were no replications in which the ability data were not satisfactorily generated.

Table 7. Means of the Correlations, Means, Standard Deviations and Approximate Ranges for the Ability Dimensions (over 100 replications.)

Data Set		$r(\theta_1,\theta_2)$	θ_1	$s(\theta_1)$	Range of θ_1	θ_2	$s(\theta_2)$	Range of θ_2
A1	Generated	-0.001	-0.003	1.003	[-3.37, 3.52]	-0.001	0.997	[-3.40, 3.44]
	Estimated	0.062	0.000	1.001	[-3.81, 3.97]	0.000	1.001	[-3.76, 3.88]
A2	Generated	0.251	-0.004	1.002	[-3.46, 3.47]	-0.002	1.001	[-3.42, 3.43]
	Estimated	0.179	0.000	1.001	[-3.74, 3.79]	0.000	1.002	[-4.22, 4.27]
A3	Generated	0.500	-0.002	1.000	[-3.39, 3.44]	0.002	1.000	[-3.46, 3.43]
	Estimated	0.282	0.000	1.001	[-3.67, 4.12]	0.000	1.003	[-4.39, 4.51]
B1	Generated	-0.001	-0.003	1.003	[-3.37, 3.52]	-1.001	0.669	[-3.28, 1.31]
	Estimated	0.147	0.000	1.001	[-3.80, 3.92]	0.000	1.000	[-4.11, 3.49]
B2	Generated	0.251	-0.004	1.002	[-3.46, 3.47]	-1.001	0.671	[-3.30, 1.30]
	Estimated	0.201	0.000	1.001	[-3.84. 3.80]	0.000	1.002	[-4.67, 3.46]
B3	Generated	0.499	-0.002	1.000	[-3.39, 3.44]	-0.999	0.671	[-3.32, 1.30]
	Estimated	0.218	0.000	1.002	[-3.83, 4.01]	0.000	1.003	[-4.74, 3.68]

s - standard deviation

Recovery of Ability Parameters.

In each of the data sets over the 100 replications, $\hat{\theta}_1$ and $\hat{\theta}_2$ had means of 0.00 and standard deviations of 1.00. The standard deviation of the mean was less than 0.002 for all data sets. The recovery of these statistics is not meaningful as a measure of accuracy because the MIRTE program rescales the theta estimates to mean 0, standard deviation 1 after each iteration in order to prevent drifting of the estimates. Therefore, with the exception of the θ_2 estimates for the B data sets, the other theta estimates appeared to reflect the specified conditions even better than the generated data did.

The ranges of the estimated thetas were larger than those of the generated thetas. Changes in correlations between the thetas did not have a consistent effect on these ranges although there was some tendency for the range to increase as correlation increased. For the B data sets the range of the estimated θ_2 was much larger than that of the generated θ_2. This is expected as the standard deviation has increased from 0.67 to 1.00 because of the rescaling by MIRTE.

Table 8. Mean Values of Correlation Coefficients for Thetas (over 100 replications)

Data Set	$\rho(\theta_1,\theta_2)$	$r(\theta_1,\theta_2)$	$r(\hat{\theta}_1,\hat{\theta}_2)$	$r(\theta_1,\hat{\theta}_1)$	$r(\theta_2,\hat{\theta}_2)$	$r(\theta_1,\hat{\theta}_2)$	$r(\theta_2,\hat{\theta}_1)$
A1	0.000	-0.001	0.062	0.842	0.764	0.505	-0.295
A2	0.250	0.251	0.179	0.842	0.824	0.603	-0.050
A3	0.500	0.500	0.282	0.831	0.865	0.699	0.209
B1	0.000	-0.001	0.147	0.773	0.517	0.662	-0.170
B2	0.250	0.251	0.201	0.765	0.623	0.713	0.052
B3	0.500	0.499	0.218	0.744	0.721	0.755	0.247

The correlation between the ability dimensions was not well recovered (see Table 8 above). As $\rho(\theta_1,\theta_2)$ increased, MIRTE tended to produce ability estimates which were less correlated than the generated abilities. This result agrees with

that reported by Carlson (1987). The difference between $\rho(\theta_1,\theta_2)$ and $r(\hat{\theta}_1,\hat{\theta}_2)$ tended to increase as $\rho(\theta_1,\theta_2)$ increased. Only in the B2 data set was the value of $r(\hat{\theta}_1,\hat{\theta}_2)$ closer to the true correlation than that obtained for the corresponding A data set.

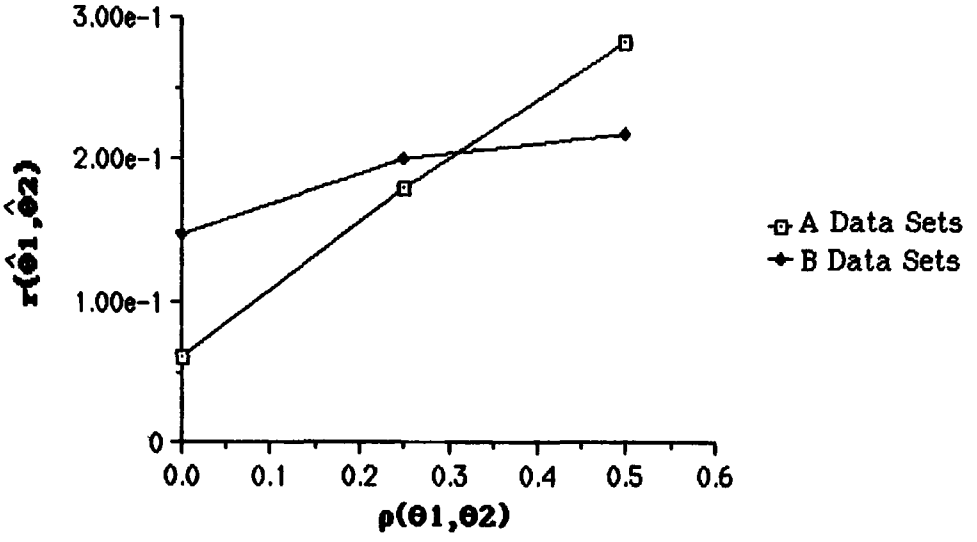


Figure 7. The relationship between $\rho(\theta_1,\theta_2)$ and $r(\hat{\theta}_1,\hat{\theta}_2)$ for the six data sets.

Figure 7 (above) shows the relationship between $\rho(\theta_1,\theta_2)$ and $r(\hat{\theta}_1,\hat{\theta}_2)$ for the A and B data sets. The estimation program, MIRTE, underestimated the true correlation between the dimensions except in the cases in which the true correlation was zero (Data sets A1 and B1). McKinley and Reckase (1984), in a similar but more limited piece of multidimensional research, reported that their MAXLOG program tended to force orthogonality on the ability dimensions in that, regardless of the specified correlation between the dimensions, the correlation was recovered at close to zero. While MIRTE appears to recover correlated dimensions

better than MAXLOG, the underestimates could be an indication that there is some tendency toward orthogonality. That the test had fewer items on which to estimate the second dimension would also affect recovery of the θ_2 parameter and consequently $r(\hat{\theta}_1,\hat{\theta}_2)$. There appeared to be interaction effects caused by the degree of correlation and differentiated ability. This is illustrated in Figure 7.

Figures 8 and 9 illustrate the recovery of the parameters θ_1 and θ_2 respectively for the six data sets. The drop in $r(\hat{\theta}_1,\hat{\theta}_2)$ in B3 is probably related to the larger drop in $r(\theta_1,\hat{\theta}_1)$ in B3 as compared to A3 (see Figure 8).

The relationship between the ability θ_1 and its estimate appeared to be well recovered in the A data sets ($r > 0.83$ for all three data sets). In the B data sets $r(\theta_1,\hat{\theta}_1)$ was lower ($r > 0.74$). The recovery of θ_1 deteriorated slightly as $\rho(\theta_1,\theta_2)$ increased while the recovery of θ_2 appeared to improve. This is evident from the patterns of the graphs in Figures 8 and 9.

The correlation between the ability θ_2 and its estimate was not recovered as well as that for θ_1 except in the data set A3 [i.e., $r(\theta_2,\hat{\theta}_2) < r(\theta_1,\hat{\theta}_1)$ for all data sets except A3]. As $\rho(\theta_1,\theta_2)$ increased, $r(\theta_2,\hat{\theta}_2)$ increased indicating that the relationship between θ_2 and $\hat{\theta}_2$ was being better recovered. It was expected that the correlation between θ_2 and $\hat{\theta}_2$ would be reduced in the B data sets because of the smaller range of θ_2 and this occurred for each of the three levels of $\rho(\theta_1,\theta_2)$ investigated. The reduction in the correlation $r(\theta_2,\hat{\theta}_2)$ for the B data sets over the corresponding A data sets was greater than the reduction in $r(\theta_1,\hat{\theta}_1)$. That is, the differentiated ability on the second dimension restricted the recovery of that dimension to a greater extent than the first dimension and this is congruent with expectations.

However, there seems to be an interaction between the degree of correlation between ability dimensions and the relationship between an ability parameter and

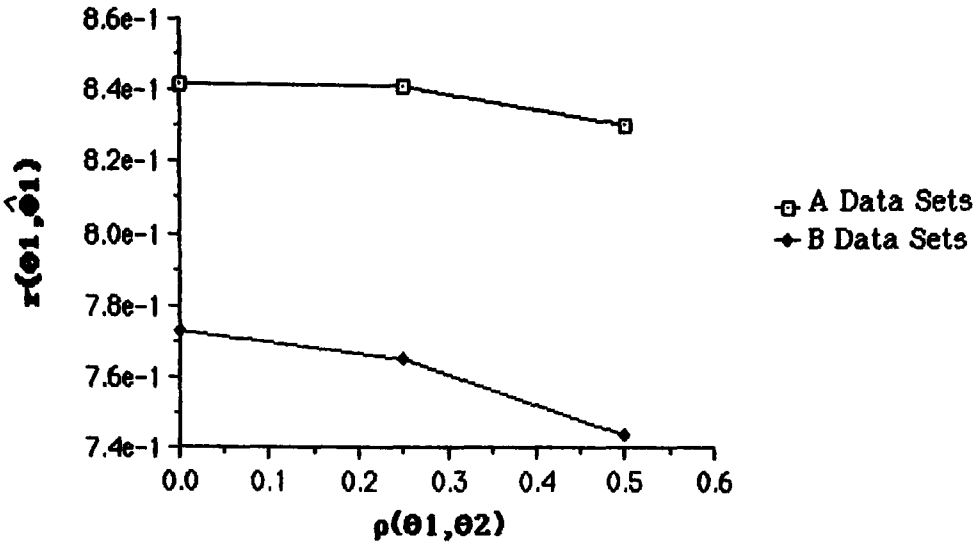


Figure 8. The recovery of θ_1 for the six data sets.

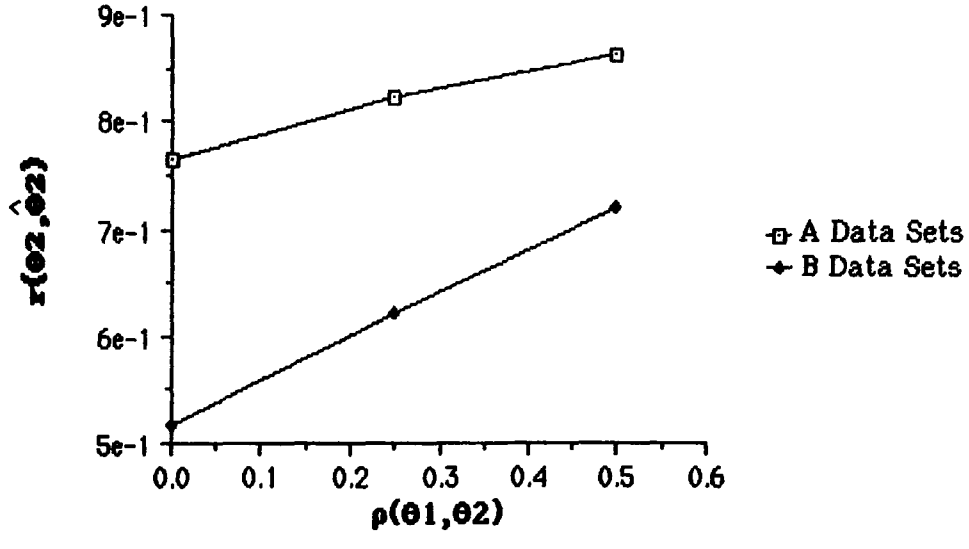


Figure 9. The recovery of θ_2 for the six data sets.

its estimate. As $\rho(\theta_1,\theta_2)$ increased, the correlation between the ability parameter and its estimate was reduced for θ_1 but improved for θ_2 which might suggest the two abilities are collapsing to a unidimensional space.

It is evident from Figure 10 that the estimate of θ_2 appeared to be more related to θ_1 than did the estimate of θ_1 to θ_2 [$r(\theta_1, \hat{\theta}_2) > r(\theta_2, \hat{\theta}_1)$ in all data sets]. As the correlation between the ability dimensions increased, the correlation between $\hat{\theta}_2$ and θ_1 also increased. When this increase in $\rho(\theta_1,\theta_2)$ was coupled with a lower mean and narrower range on the second ability dimension, the estimate of θ_2 seemed to be depending even more on the θ_1 ability. In the B data sets, $r(\theta_1, \hat{\theta}_2) > r(\theta_2, \hat{\theta}_2)$. The correlation between the estimate of θ_1 and true θ_1 was higher than $r(\theta_2, \hat{\theta}_1)$ for all data sets. The correlation $r(\theta_2, \hat{\theta}_1)$ was small (< 0.3) and fluctuated in size and sign. In this case as well there appeared to be some constant relationship in the estimation of these variables as $\rho(\theta_1,\theta_2)$ increased.

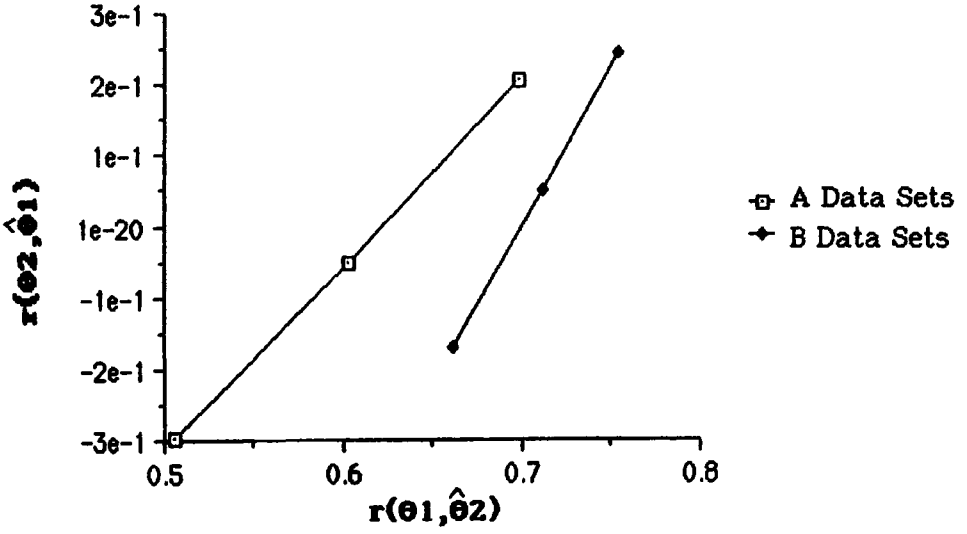


Figure 10. The relationship between $r(\theta_1,\hat{\theta}_2)$ and $r(\theta_2,\hat{\theta}_1)$ for the six data sets.

It would appear from the graphs in Figures 8, 9, and 10 that similar trends are occuring in the A and B data sets as $\rho(\theta_1,\theta_2)$ increases with respect to the ability estimates.

The correlations of the estimate with the true parameter are not direct measures of the recovery of the ability parameter. Two other measures were made which relate more directly to the adequacy of the recovery of the parameters. These were the average absolute deviation and the standard error (as calculated by MIRTE). The mean average absolute deviations of θ_i from $\hat{\theta}_i$, ($AAD(\hat{\theta}_i)$), are given in Table 9. The $AAD(\hat{\theta}_1)$ ranged from 0.446 to 0.459 for the A data sets. For the B data sets $AAD(\hat{\theta}_1)$ was larger (ranging from 0.463 to 0.566).

Table 9. Means of the Average Absolute Deviations for the Theta Estimates (over 100 replications)

Data Set	$\rho(\theta_1,\theta_2)$	$AAD(\hat{\theta}_1)$	$AAD(\hat{\theta}_2)$
A1	0.00	0.447	0.544
A2	0.25	0.446	0.470
A3	0.50	0.459	0.412
B1	0.00	0.463	0.856
B2	0.25	0.544	1.079
B3	0.50	0.566	1.047

As $\rho(\theta_1,\theta_2)$ increased there was a tendency for the $AAD(\hat{\theta}_1)$ to increase, moreso in the B data sets. The degree of correlation between the dimensions had a greater effect on the statistic $AAD(\hat{\theta}_2)$ than on $AAD(\hat{\theta}_1)$. In the A data sets, $AAD(\hat{\theta}_2)$ decreased as $\rho(\theta_1,\theta_2)$ increased. In the B data sets, $AAD(\hat{\theta}_2)$ increased as $\rho(\theta_1,\theta_2)$ increased. However, the distribution of θ_2 for the B data sets was such that the specified mean was -1. The estimates of θ_2 were scaled after each iteration to mean 0 which would result in the $AAD(\hat{\theta}_2)$ being larger. These values

of AAD($\hat{\theta}_2$) are not very useful because of the rescaling. That θ_2 appeared to be better recovered in the A data sets as $\rho(\theta_1,\theta_2)$ increased may be partly related to the compensatory nature of the M2PL model. There was very little variance over replications in the AADs (approximately 0.001 for $\hat{\theta}_1$ and 0.002 for $\hat{\theta}_2$) so that the thetas appear to have been recovered consistently within the six data sets. However, there is some cause for concern regarding the size of the AADs particularly in the A data sets where both dimensions were distributed over the full range.

The mean standard error of the thetas (as calculated by MIRTE) was approximately 0.259 for the A data sets and 0.287 for the B data sets. These values are almost half the size of the AADs except in the case of AAD($\hat{\theta}_2$) for the B data sets. The variance in these mean standard errors was very small (< 0.001) although this variance increased as the correlation between the dimensions increased and was larger for the B data sets. [The standard errors are not reported for each dimension as there was a problem separating them during the replications and se($\hat{\theta}_1$) was not distinguished from se($\hat{\theta}_2$).]

In summary, the theta parameters were recovered relatively well when there was no correlation between the ability dimensions and both abilities were distributed over the full range (data set A1). The ability θ_1 was recovered better than θ_2. As $\rho(\theta_1,\theta_2)$ increased, there appeared to be some improvement in the recovery of θ_2 (as shown by increased $r(\theta_2,\hat{\theta}_2)$ and smaller AAD($\hat{\theta}_2$) for the A data sets) while θ_1 was slightly less well recovered (as shown by reduced $r(\theta_1,\hat{\theta}_1)$ and larger AAD($\hat{\theta}_1$)). When the second ability was restricted (as in the B data sets), neither ability dimension was recovered as well as in the corresponding A data set although θ_2 seemed to be affected moreso than θ_1. Seemingly it became more difficult for the analysis program to distinguish between the two dimensions and

there was a tendency for the latent space to collapse. Higher average absolute deviations and larger standard errors for thetas in the B data sets indicate that it was more difficult to estimate the thetas than in the A data sets. There appeared to be some compound effects on the recovery of the ability parameters as well. As $\rho(\theta_1,\theta_2)$ increased, coupled with the restriction on the θ_2 dimension, θ_1 was less well recovered as evidenced by the higher average absolute deviations and the larger drop in $r(\theta_1,\hat{\theta}_1)$ from B1 to B3 as compared to the drop from A1 to A3 (0.029 compared to 0.011). An interaction was evident in the recovery of $\rho(\theta_1,\theta_2)$ (see Figure 7) where the increased correlation and differentiated ability on θ_2 caused a reduction in $r(\hat{\theta}_1,\hat{\theta}_2)$ and also in the recovery of the relationship between the ability parameter and its estimate.

Recovery of the Item Parameters

The recovery of the item difficulty parameters is presented followed by a presentation of the recovery of the item discrimination parameters. Recall that the same item parameters were used for all six data sets. Because of the nature of the MIRTE estimation program, theta estimates are used to estimate item parameters and item parameter estimates are used to estimate thetas. The final parameter estimates are, therefore, interdependent.

Item Difficulty Parameters.

The descriptive statistics for the difficulty parameter, d, and for the multidimensional difficulty parameter, D, are given in Table 10.

For the A data sets (both abilities covering the full range) the difficulty parameter, d, was relatively well recovered. Here the mean and standard deviation of the estimates of d were very similar to the true mean and standard deviation. As the correlation between the ability dimensions increased, the mean

of d was less well recovered appearing to be slightly overestimated. The standard deviation of d was also overestimated. The standard error of $\hat{d}$ (as calculated by MIRTE) decreased as $\rho(\theta_1,\theta_2)$ increased (all were approximately 0.11) while the AAD($\hat{d}$) increased slightly (all were approximately 0.23). It is logical that the items appear to be easier as the dimensions become more highly correlated. The mean of $\hat{d}$ was not very different from the true mean. Some of this difference might be attributed to error in a simulation study. Although 100 replications is more than previous multidimensional IRT research has reported, it could be argued that more replications (perhaps 500 to 1000) are required.

Table 10. Means of the Mean, Standard Deviation, Standard Error, Average Absolute Deviation and Correlation for Item Difficulty; Means of the Mean, Standard Deviation and Correlation for Multidimensional Difficulty (over 100 replications)

Data Set	$\hat{d}$	$s(\hat{d})$	$se(\hat{d})$	$AAD(\hat{d})$	$r(d,\hat{d})$	$\hat{D}$	$s(\hat{D})$	$r(D,\hat{D})$
True	0.000	3.771	-----	------	-----	0.000	2.058	------
A1	0.009	3.929	0.112	0.224	0.997	0.006	2.079	0.995
A2	0.010	3.936	0.109	0.228	0.997	0.005	2.028	0.994
A3	0.030	3.936	0.106	0.232	0.997	0.012	1.999	0.991
B1	-0.726	3.995	0.137	0.811	0.984	0.460	2.535	0.958
B2	-0.734	4.001	0.132	0.827	0.982	0.434	2.459	0.956
B3	-0.716	4.044	0.123	0.834	0.982	0.397	2.247	0.969

s- standard deviation; se - standard error (calculated by MIRTE)

The multidimensional difficulty, D, was slightly overestimated in all three A data sets. The mean of D was not as well recovered when $\rho(\theta_1,\theta_2) = 0.50$ (similar to that of d). The standard deviation of $\hat{D}$ decreased as $\rho(\theta_1,\theta_2)$ increased and became underestimated for non-zero correlations between the ability dimensions. In contrast to d however, with increase in correlation, D indicated

that items were slightly more difficult. This may be partly error or caused by poorer estimates of the a_is as multidimensional difficulty is a function of the difficulty parameter, d, and the multidimensional discrimination statistic, MDISC, and therefore, the estimates of multidimensional difficulty would be affected by discrimination parameter estimates.

The correlation between d and $\hat{d}$ was very high for all the A data sets as were the correlations between D and $\hat{D}$ (all $r > 0.99$). The size of these correlation coefficients along with the small standard errors and $AAD(\hat{d})$ are indications that the difficulty parameters, d and D, were well recovered in the A data sets.

In the B data sets, the recovery of d and D was much poorer. The lower mean and restriction of the range of θ_2 seemed to greatly affect the mean of $\hat{d}$. The rescaling of θ_2 to mean of zero, standard deviation of 1, made the "sample" in the B data sets appear more able than in fact the original sample was. This made the estimates of d (and D as well) appear as though the items are more difficult. The standard deviations increased as $\rho(\theta_1,\theta_2)$ increased and were slightly larger than the true standard deviation or those found in the corresponding A data sets. The standard errors of $\hat{d}$ were also larger than the standard errors found in the corresponding A data sets and, as with the A data sets, the standard errors decreased as $\rho(\theta_1,\theta_2)$ increased. The $AAD(\hat{d})$ increased as $\rho(\theta_1,\theta_2)$ increased and were almost four times the size of the corresponding $AAD(\hat{d})$ in the A data sets.

For the multidimensional difficulty vector parameter, D, the mean was greatly overestimated although not as much as $\rho(\theta_1,\theta_2)$ increased. The overestimation appears to be a result of the items appearing more difficult as the θ_2 estimates were rescaled. The standard deviation of D was overestimated somewhat but not nearly to the extent of the overestimation of the mean. The standard deviation of $\hat{D}$ also decreased as $\rho(\theta_1,\theta_2)$ increased. The parameter D is a

function of d and MDISC and is therefore affected by these estimates. In the B data sets d and D had opposite signs and the items appeared to become less difficult as $\rho(\theta_1,\theta_2)$ increased. In the A data sets $\hat{d}$ and $\hat{D}$ were both positive and very close to the true mean of zero. The fact that the mean estimates were both slightly positive is likely random error.

The correlation coefficients may provide some evidence of the recovery of the parameters. Correlations between d and $\hat{d}$ were very high for the B data sets (all $r > 0.98$) as were the correlations between D and $\hat{D}$ (all $r > 0.95$), although all correlations were less than the corresponding correlation in the A data sets. As $\rho(\theta_1,\theta_2)$ increased, the $r(d,\hat{d})$ decreased slightly (as $r(D,\hat{D})$ had done in the A data sets). This was not expected but the change was very small and may not be meaningful. The value of $r(D,\hat{D})$ increased as $\rho(\theta_1,\theta_2)$ increased for the B data sets. This was expected as difficulty was better recovered for higher correlations (better estimates of D and s(D)).

Table 11 gives correlations between the number correct (NC) for an item and each of the two item difficulty parameters (d and D) and their estimates. The NC was highly correlated with d for all data sets with the correlation increasing as $\rho(\theta_1,\theta_2)$ increased. The recovery of this correlation in $r(NC,\hat{d})$ was good, particularly in the A data sets. The results are similar for correlations between NC and D or its estimate. The variable NC did not correlate highly with any other item variables in the study (all other correlations were < 0.11). The correlation between D and d was -1.000 and this was recovered as approximately -0.98 for the A data sets and at -0.95 for the B data sets. As $\rho(\theta_1,\theta_2)$ increased, this correlation was not affected very much in the A or B data sets.

The recovery of the difficulty parameters was affected more by the differentiated ability on the second dimension than by the changes in $\rho(\theta_1,\theta_2)$.

When both abilities covered the full range, both d and D were relatively well recovered. However, with the lower mean and the restriction of range on θ_2, both difficulty parameters were less well estimated, apparently an artifact of rescaling. There is a suggestion of some interaction effects of correlation of abilities and differentiated θ_2 ability. As $\rho(\theta_1,\theta_2)$ increases, the mean of $\hat{D}$ increases slightly in the A data sets but decreases in the B data sets. The mean of $\hat{d}$ increases slightly in the A data sets but the pattern is inconsistent in the B data sets. In the A data sets, changes in $\rho(\theta_1,\theta_2)$ effect only a small change in the means of $\hat{D}$ and $\hat{d}$. In the B data sets, the change in the mean of $\hat{D}$ as $\rho(\theta_1,\theta_2)$ increases is much larger. Here the items are becoming less difficult as the dimensions become more correlated (although D is overestimated). This is what one would expect to occur.

Table 11. Mean Correlations for NC, d, $\hat{d}$, D, and $\hat{D}$ (over 100 replications)

Data Set	r(NC, d)	r(NC, $\hat{d}$)	r(NC, D)	r(NC, $\hat{D}$)	r($\hat{d}$, $\hat{D}$)
A1	0.982	0.981	-0.980	-0.975	-0.985
A2	0.984	0.982	-0.982	-0.976	-0.984
A3	0.985	0.983	-0.983	-0.976	-0.980
B1	0.972	0.967	-0.970	-0.940	-0.951
B2	0.973	0.967	-0.971	-0.941	-0.947
B3	0.974	0.968	-0.973	-0.956	-0.957

Item Discrimination Parameters.

There are four variables to consider in this section, the two discrimination parameters, a_1 and a_2, the multidimensional discrimination parameter, MDISC, which is a function of a_1 and a_2, and the angle α_1 (α_2 is the complement of α_1). Descriptive statistics for these variables are given in Table 12.

Table 12. Means of the Mean, Standard Deviation, Standard Error, and Average Absolute Deviation for a_1 and a_2, and for the Mean and Standard Deviation for MDISC and α_1 (over 100 replications)

Data Set	$\hat{a}_1$	$s(\hat{a}_1)$	$se(\hat{a}_1)$	$AAD(\hat{a}_1)$	$\hat{a}_2$	$s(\hat{a}_2)$	$se(\hat{a}_2)$	$AAD(\hat{a}_2)$	$\widehat{MDISC}$	$s(\widehat{MDISC})$	$\hat{\alpha}_1$
True	1 637	0.251	-----	------	0.678	0.496	-----	------	1 850	0 151	22 50
A1	1.195	0.569	0.099	0.500	1.379	0 512	0.096	0 707	1 957	0 288	49 07
A2	1.201	0 528	0.095	0 486	1 448	0 582	0.094	0 775	2 013	0.319	49 40
A3	1.202	0 502	0.093	0.490	1 510	0 628	0.093	0 836	2 057	0 381	49 98
B1	1.076	0.551	0.119	0.623	1 228	0 449	0.112	0 653	1 736	0 398	49 09
B2	1.094	0.557	0.138	0 620	1 298	0.501	0 108	0 708	1 803	0.449	49 76
B3	1.139	0.599	0 134	0 624	1 408	0 534	0.103	0 791	1.922	0.495	50 94

s - standard deviation, se - standard error (calculated by MIRTE)

In all six data sets, a_1 was underestimated and a_2 was overestimated. In the 104 items there were only eight different values for a_1 and seven for a_2. In order that the simulated test primarily measure dimension one, the values of a_1 were equal to or larger than those of a_2. The discrimination parameters did not cover the space particularly well. As $\rho(\theta_1,\theta_2)$ increased, the means of $\hat{a}_1$ and $\hat{a}_2$ increased. The mean of $\hat{a}_2$ was larger than the mean of $\hat{a}_1$ in all six data sets. The standard deviation of $\hat{a}_1$ was more than twice that of a_1. As $\rho(\theta_1,\theta_2)$ increased, in the A data sets, this standard deviation decreased; in the B data sets, $s(\hat{a}_1)$ increased as $\rho(\theta_1,\theta_2)$ increased. The standard deviation of $\hat{a}_2$ increased as $\rho(\theta_1,\theta_2)$ increased for both the A and B data sets and was smaller in the B data sets than in the corresponding A data sets. The standard error of $\hat{a}_1$ was larger in the B data sets than in the A data sets (as was the case for the ability and difficulty estimates). This statistic was smaller for non-zero $\rho(\theta_1,\theta_2)$ in the A data sets but became larger in the B data sets when $\rho(\theta_1,\theta_2) \neq 0$. The standard error of $\hat{a}_2$ was smaller than the corresponding standard error of $\hat{a}_1$ in each of the data sets. As $\rho(\theta_1,\theta_2)$ increased, the standard error of $\hat{a}_2$ decreased suggesting that a_2 was being better estimated. This represents an interaction effect of correlation of abilities and differentiated ability on θ_2 with $s(\hat{a}_1)$ and $se(\hat{a}_1)$. The size of $AAD(\hat{a}_1)$ was smaller than any of the corresponding values of $AAD(\hat{a}_2)$ and showed no predictable trend as $\rho(\theta_1,\theta_2)$ changed. The average absolute deviation was affected by the restriction of the ability on the θ_2 dimension. In the case of a_1, the restricted ability on θ_2 resulted in larger values of $AAD(\hat{a}_1)$ for any given correlation between the ability dimensions. For a_2, the reverse was true. Correlated abilities and differentiated ability on θ_2 interacted to affect $AAD(\hat{a}_2)$ in that the $AAD(\hat{a}_2)$ was smaller when the second ability was restricted for any of the given $\rho(\theta_1,\theta_2)$. This is in keeping with the smaller $s(\hat{a}_2)$. This suggests that a_2

was being better recovered when θ_2 was restricted.

The mean of the multidimensional discrimination parameter, MDISC, was overestimated in all data sets except B1 and B2. An overestimation of the mean should be expected because of the large overestimation of a_2 which is part of the value of MDISC. As $\rho(\theta_1,\theta_2)$ increased, the mean of $\widehat{MDISC}$ increased. The standard deviation of MDISC was also overestimated in all data sets, was higher for the B data sets, and the overestimation increased as $\rho(\theta_1,\theta_2)$ increased.

Correlation coefficients provide some additional evidence of the adequacy of parameter recovery (Table 13).

Table 13. Mean Correlation Coefficients for Item Discrimination Values (over 100 replications)

Data Set	$r(a_1,\hat{a}_1)$	$r(a_2,\hat{a}_2)$	$r(\hat{a}_1,\hat{a}_2)$	$r(a_1,\hat{a}_2)$	$r(a_2,\hat{a}_1)$	$r(MDISC,\widehat{MDISC})$	$r(\alpha_1,\hat{\alpha}_1)$
A1	0.834	0.893	-0.765	-0.572	-0.865	0.600	0.943
A2	0.818	0.899	-0.769	-0.587	-0.830	0.565	0.933
A3	0.760	0.895	-0.735	-0.587	-0.747	0.502	0.907
B1	0.530	0.523	-0.428	-0.309	-0.543	0.296	0.630
B2	0.460	0.511	-0.401	-0.306	-0.455	0.269	0.586
B3	0.431	0.514	-0.459	-0.285	-0.403	0.285	0.564

In all cases a_1 correlated more highly with $\hat{a}_1$ than with $\hat{a}_2$. Similarly, a_2 correlated more highly with $\hat{a}_2$ than with $\hat{a}_1$ except in the B1 data set. (In B1 the correlation of a_2 with $\hat{a}_1$ was only slightly larger in magnitude. All $r(a_1,\hat{a}_2)$ and $r(a_2,\hat{a}_1)$ were negative.) The above correlations were considered as evidence that the check used during the replications adequately fixed the dimensions. The correlation coefficients $r(a_2,\hat{a}_2)$ were larger than the corresponding $r(a_1,\hat{a}_1)$ in all of the data sets except B1 although this may be attributed to the larger standard

deviation of a_2 as compared to that of a_1. The greater variability in a_2 (almost four times that of a_1) would allow for higher correlations. The true a_1 mean was almost 2.5 times the size of the true a_2 mean. That the means of their estimates are comparable in size to each other and the mean of $\hat{a}_2$ is always larger than that of $\hat{a}_1$ would indicate that the discrimination parameters appear to be dispersed across the $\theta_1\theta_2$-space. Carlson (1987) reported that with the MIRTE program the estimation of the a-parameters appeared to be sensitive to the distribution of the discrimination parameters in the generated data. He found the estimates of the a-parameters did not correlate as highly with the generating parameters and this also seemed to lower somewhat the size of the correlations between the ability estimates and parameters and the difficulty estimate and parameter although the greatest impact was on the correlations between the discrimination parameters and estimates. Clearly, the discrimination parameters were not fully representative of the space.

Figures 11 and 12 graphically illustrate the recovery of the discrimination parameters, a_1 and a_2.

Correlation coefficients indicate that the correlations between the a_is were better recovered in the A data sets than in the B data sets. As $\rho(\theta_1,\theta_2)$ increased, the correlation coefficient between the a_1 and $\hat{a}_1$ decreased in both the A and B data sets. This trend did not hold for a_2 in either the A or the B data sets. Restricting the range of θ_2 ability greatly reduced the correlation between the discrimination parameter and its estimate for both a_1 and a_2. As with the difficulty parameters, the differentiated ability on the second dimension seemed to hinder the recovery of the discrimination parameters. The combination of differentiated ability and changes in degree of correlation between the ability dimensions did not reveal predictable trends in the recovery of the discrimination

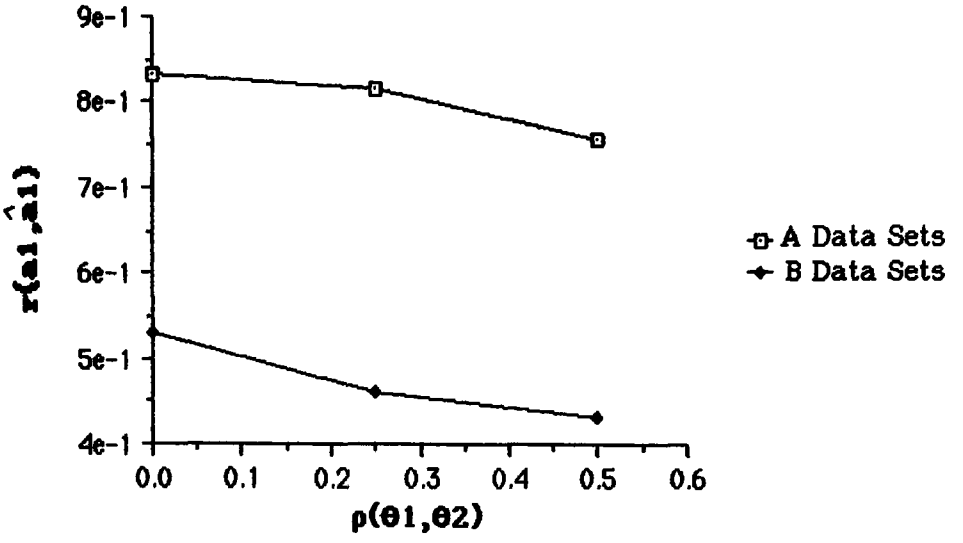


Figure 11. The recovery of a_1 for the six data sets.

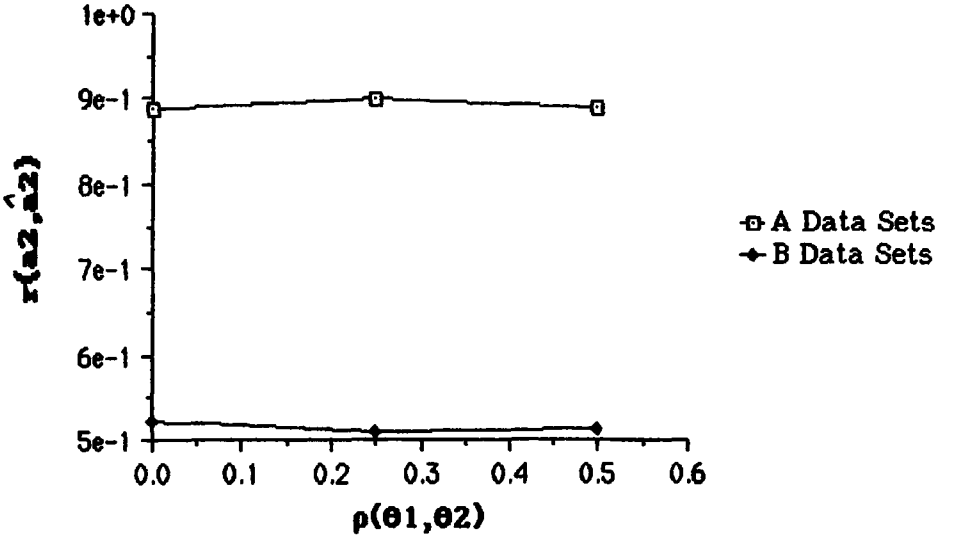


Figure 12. The recovery of a_2 for the six data sets.

parameters. The true correlation between a_1 and a_2 was -0.738. This was recovered relatively well in the A data sets (-0.735 to -0.769) but was greatly underestimated in the B data sets (-0.401 to -0.459).

The multidimensional discrimination parameter, MDISC, did not correlate as highly with its estimate as did the other discrimination parameters with their estimates. For the A data sets the correlation was highest (0.600) when the ability dimensions were uncorrelated. This correlation decreased as $\rho(\theta_1,\theta_2)$ increased. For the B data sets the correlations between the ability dimensions were consistently very low. This also suggests the differentiated ability on θ_2 may be hampering the recovery of the discrimination parameters.

There appeared to be a rotational indeterminacy in the recovery of the discrimination parameters and a tendency to spread the discrimination parameters over the entire space even though they originally did not cover the entire space. This was supported by the statistics on the angle estimates, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ (see Table 12 above). The angles are a function of the discrimination parameters and relate to the ability dimensions in that the smaller the angle, the more the item requires that corresponding dimension in order to respond correctly. Originally α_1 had a mean of 22.50°. This was recovered in all data sets at approximately 49°. Similarly the mean of α_2 (67.50°) was recovered at approximately 41°. The original standard deviation of 16.85° increased to approximately 20°. The importance of the angle α_1, like the discrimination parameter a_1, was underestimated while the importance of the corresponding parameters for the second dimension was overestimated. As with the a_i parameters and MDISC, there was a higher correlation of the α_k with its estimate ($|r| > 0.91$) in the A data sets than in the B data sets ($|r| > 0.56$) (see Table 13 above). As the correlation between the ability dimensions increased, this correlation coefficient decreased for both A

and B data sets. There seemed to be an attempt to cover the entire space in estimation of parameters related to discrimination. In the A and B data sets the angle α_2 and the discrimination parameter a_2 seem to be assuming more importance in the recovery of all parameters. The space is better covered by a_2 than a_1 although neither was well distributed over the space. The range of a_2 was larger than that of a_1. This may be partially causing the poorer recovery of a_1 and α_1 which in turn affects recovery of θ_1 and d.

The estimation program, MIRTE, also produces a summary of the distribution of the residual covariances between items. The residual of examinee j on item i, r_{ij}, is the difference between the examinee's response on item i and the probability of that examinee answering correctly item i as estimated by substituting estimates of the item and ability parameters into Formula (15) (see p. 57). Table 14 lists the summary (by percentages) of the frequency distribution of the mean residual covariances between all pairs of items. As the correlation between the ability dimensions increased, there were more items with smaller residual covariances. The differentiated ability on the second dimension resulted in larger residual covariances.

As in most previous research using the compensatory M2PL model, the difficulty parameters were recovered best, particularly in the A data sets. As $\rho(\theta_1,\theta_2)$ increased, the difficulty parameter, d, tended to be slightly overestimated in the A data sets. In the B data sets, d was greatly underestimated. This seems likely to be an artifact of the scaling procedures as the standard errors were small even in the B data sets. Difficulty would surely be recovered better if ESL type groups were part of a group with full language abilities. This might help to alleviate the rescaling problem.

Table 14. Summary (by percentages) of the Frequency Distribution of the Mean Residual Covariances (over 100 replications)

Data Set	< 0.011	0.011-0.012	0.013-0.014	0.015-0.016	0.017-0.018	0.019-0.020	> 0.020
A1	99.50	0.31	0.12	0.05	0.02	0.02	0.00
A2	99.56	0.28	0.11	0.04	0.01	0.01	0.00
A3	99.61	0.24	0.09	0.03	0.01	0.01	0.01
B1	98.92	0.53	0.26	0.13	0.07	0.04	0.04
B2	99.03	0.48	0.24	0.12	0.06	0.03	0.03
B3	99.30	0.36	0.15	0.08	0.08	0.03	0.03

The multidimensional difficulty parameter, D, is a function of d and MDISC and its recovery is, therefore, related to the recovery of the d and a_i variables. In the A data sets, recovery of D was adequate but it was largely overestimated in the B data sets. McCauley and Mendoza (1985) reported that a unidimensional analysis of two-dimensional data with a lower mean on the second dimension recovered larger difficulty estimates. This may correspond to the parameter D which represents a vector length from the origin of the latent space to the point of maximum information for an item. For a less able sample, an item may appear more difficult. The high correlation coefficients between the item parameter and its estimate suggest that relative order is preserved.

The degree of correlation between the ability dimensions affected recovery of the discrimination parameters. In all of the data sets an increase in $\rho(\theta_1,\theta_2)$ resulted in poorer recovery of a_1 as measured by $r(a_1,\hat{a}_1)$ [a_1 was underestimated in all data sets]. This corresponds to the poorer recovery of θ_1 as $\rho(\theta_1,\theta_2)$ increased. The a_2 parameter (overestimated in all data sets) tended to be slightly better recovered for $\rho(\theta_1,\theta_2) \neq 0$ in the A data sets although the trend was not as

definite. In the B data sets, a_2 tended to be less well recovered for non-zero $\rho(\theta_1,\theta_2)$ as measured by $r(a_2,\hat{a}_2)$.

The differentiated ability on θ_2 restricted the recovery of both a_1 and a_2. Both discrimination parameters were better estimated given full ability ranges on both dimensions. The a_2 parameter appeared to be more dominant in both the A and B data sets in spite of the reverse being true.

The combined effects of degree of correlation and differentiated ability on recovery of a_1 or a_2 were evident in their standard errors which moved in opposite directions to each other for the B data sets. The $se(\hat{a}_1)$ increased from B1 to B3 and $se(\hat{a}_2)$ decreased. In the A data sets $se(\hat{a}_1)$ and $se(\hat{a}_2)$ were very similar in size and both decreased as $\rho(\theta_1,\theta_2)$ increased.

Discussion of Research Questions

Three research objectives were presented in Chapter 1: to determine how parameter estimates are affected by a) different degrees of correlation between the two ability dimensions; b) responses of examinees with ability on the dominant dimension normally distributed over the full range but with a lower mean and narrower ability range on the second dimension; and c) the interaction of the differential ability on the second dimension and degree of correlation. These will be discussed in turn.

Effects of Increasing the Correlation between Ability Dimensions.

Increasing the correlation between the ability dimensions (from 0.00 to 0.50) had no effect on raw score means but the variance of the raw scores increased.

The correlation between the estimated ability dimensions was lower than the true correlation except at $\rho(\theta_1,\theta_2) = 0.00$ (when it was overestimated). As $\rho(\theta_1,\theta_2)$ increased, the correlation was less well recovered. However, $r(\hat{\theta}_1,\hat{\theta}_2)$ was

a linear function of $\rho(\theta_1,\theta_2)$ for both the A and B data sets indicating a consistency in the recovery of the correlation between dimensions.

Increasing $\rho(\theta_1,\theta_2)$ resulted in reduced $r(\theta_1,\hat{\theta}_1)$ while $r(\theta_2,\hat{\theta}_2)$ improved slightly. As $\rho(\theta_1,\theta_2)$ increased, the $AAD(\hat{\theta}_1)$ increased and the $AAD(\hat{\theta}_2)$ decreased. These results are consistent with the results reported for the discrimination parameters. Similarly the mean of the discrimination parameter, a_1, was underestimated while that of a_2 was overestimated. The second dimension seemed to be assuming a dominance over the first as $\bar{\hat{a}}_2 > \bar{\hat{a}}_1$ for all data sets. As $\rho(\theta_1,\theta_2)$ increased, the mean of a_1 was better estimated but that of a_2 was less well estimated. The standard deviation of a_1 and a_2 were overestimated but changes in $\rho(\theta_1,\theta_2)$ had opposite effects here. In the A data sets, $s(\hat{a}_1)$ improved (decreased) as $\rho(\theta_1,\theta_2)$ increased but $s(\hat{a}_2)$ became worse (increased). Standard errors of both discrimination parameters improved (decreased) as $\rho(\theta_1,\theta_2)$ increased. The correlation between the ability dimensions had no effect on $AAD(\hat{a}_1)$ but $AAD(\hat{a}_2)$ increased as $\rho(\theta_1,\theta_2)$ increased, a change opposite to that of the standard error. As $\rho(\theta_1,\theta_2)$ increased, both $r(a_1,\hat{a}_1)$ and $r(\alpha_1,\hat{\alpha}_1)$ decreased. There was no trend relating to $\rho(\theta_1,\theta_2)$ in changes in $r(a_2,\hat{a}_2)$. MDISC and its standard deviation were less well estimated as $\rho(\theta_1,\theta_2)$ increased reflecting poorer recovery of the a_i parameters.

In all of the situations regarding thetas and a_is described above, there seems to be a trend. The trend evident in looking at the results for θ_1, a_1, and α_1 is that the relationship between the first dimension parameters and their estimates is less well recovered as $\rho(\theta_1,\theta_2)$ increases and there is a tendency for the second dimension to become more dominant ($\bar{\hat{a}}_2 > \bar{\hat{a}}_1$). There seems to be a collapsing of the space so that a_1 and a_2, α_1 and α_2, and even θ_1 and θ_2 become more similar. Much of the discrepancy here can be related to the discrimination parameters.

The range of a_2 was larger than that of a_1. This may be contributing to the poorer recovery of a_1. The test space is not well covered by the discrimination parameters and this may also be contributing to the poor estimates of both a_1 and a_2. This would correspond to the finding of Carlson (1987) that distribution of the discrimination parameters can have an effect on the recovery of all parameters but most seriously affects the recovery of the discrimination parameters. Possibly some model effects are occurring here as a compensatory model was chosen.

As $\rho(\theta_1,\theta_2)$ increased, effects on the difficulty parameters were small. The standard deviation of $\hat{d}$ increased, the se($\hat{d}$) decreased, and the AAD($\hat{d}$) increased. There were no effects of correlation on r(d,$\hat{d}$) which had very high values. The difficulty vector, D, was affected differently. Both the standard deviation of $\hat{D}$ and r(D,$\hat{D}$) decreased as $\rho(\theta_1,\theta_2)$ increased although the items appeared to become slightly more difficult. This parameter is a function of d and MDISC and these results are probably caused by the poorer recovery of the discrimination parameters as the deterioration in the recovery of D is similar to that of MDISC (r(MDISC,$\widehat{\text{MDISC}}$) also decreased as $\rho(\theta_1,\theta_2)$ increased).

Finally, as $\rho(\theta_1,\theta_2)$ increased, more items had smaller residual covariances. This was as expected.

Effects of a Differentiated Ability on θ_2.

Lower mean raw scores were reported in the B data sets. The sample was less able and this result was expected.

A differentiated ability on θ_2 resulted in $\rho(\theta_1,\theta_2)$ being less well recovered except at $\rho(\theta_1,\theta_2) = 0.25$. This is discussed below as an interaction. The poorer recovery of θ_1 and θ_2 would affect $r(\hat{\theta}_1,\hat{\theta}_2)$.

The relationships between the ability parameters and estimates were not as well recovered in the B data sets as in the A data sets ($r(\theta_1,\hat{\theta}_1)$ and $r(\theta_2,\hat{\theta}_2)$ were

both reduced). The differentiated ability on θ_2 restricted the recovery of the relationship in the second dimension moreso than in the first. The estimates of parameters are interdependent because of the iterative nature of the estimation procedure. The smaller range of θ_2 would reduce $r(\theta_2,\hat{\theta}_2)$. Because the ability space was not covered as well in the B data sets, the recovery of all parameters was expected to deteriorate. Larger AADs and standard errors were found in the B data sets as compared to the corresponding A data sets.

The differentiated ability on θ_2 seemed to improve the recovery of a_2 (smaller $s(\hat{a}_2)$, smaller $AAD(\hat{a}_2)$, and a mean estimate slightly closer to the true mean of a_2). The discrimination parameter a_1 was less well recovered (larger $se(\hat{a}_1)$ and $AAD(\hat{a}_1)$) although the mean estimate of a_1 improved as did the standard deviation. The correlations $r(a_1,\hat{a}_1)$, $r(a_2,\hat{a}_2)$, $r(\hat{a}_1,\hat{a}_2)$, r(MDISC, $\widehat{\text{MDISC}}$), and $r(\alpha_1,\hat{\alpha}_1)$ were all lower in the B data sets than in the corresponding A data sets. As in the A data sets, there was a tendency for the second dimension to assume a dominance over the first. The sample did not have a full range of ability resulting in their being not as able in the B data sets. Therefore, estimates of item parameters were expected to be poorer than in the A data sets. Because of the iterative procedures in the estimation program, this should also result in poorer ability estimates. The differentiated ability on θ_2, coupled with the discrimination parameters not covering the space, adversely affected recovery of thetas and discrimination parameters.

The difficulty parameters were not as well recovered under the B data set conditions. The means of $\hat{d}$ and $\hat{D}$ varied greatly from the true means. There were larger values for $s(\hat{d})$, $se(\hat{d})$, $AAD(\hat{d})$, $s(\hat{D})$, and reduced values for $r(d,\hat{d})$ and $r(D,\hat{D})$ than found in the A data sets. It was expected that difficulty would be less well estimated in the B data sets. The items appear to be more difficult to a less able

sample (the mean of $\hat{D}$ greatly increased). The rescaling of θ_2 by MIRTE had an effect on difficulty estimates. The response vectors were generated for a differentiated θ_2 ability. The parameter estimates were made by rescaling θ_2. This made the items appear more difficult. Although rescaling seems to cause a problem with the difficulty estimates, the difficulty parameters remained the most stable throughout the analyses.

As expected, larger residual covariances were found in the B data sets.

Interaction Effects of Correlated Abilities and Differentiated Ability on θ_2.

There were four possible interaction effects, the first in the recovery of $\rho(\theta_1,\theta_2)$ (see Figure 7 above). There was an interaction between the level of correlation of abilities and differentiated ability on θ_2 on the estimated correlation of abilities. For the B data sets, there was little effect of correlated abilities. For the A data sets, much steeper slopes resulted when $r(\hat{\theta}_1,\hat{\theta}_2)$ was plotted against $\rho(\theta_1,\theta_2)$. There was a poorer recovery of $\rho(\theta_1,\theta_2)$ in the B data sets with the exception of B2. The difference in $r(\hat{\theta}_1,\hat{\theta}_2)$ was small between data set A2 and B2 and perhaps is not as meaningful. Indeed this may not be a true interaction even though the lines cross as the B data sets consistently appear to recover $\rho(\theta_1,\theta_2)$ less well. The slightly better recovery of the $\rho(\theta_1,\theta_2)$ of 0.25 may in fact be an artifact of a regression line showing no relationship and consistently estimating correlation close to 0.25 regardless of the true correlation. One would expect $\rho(\theta_1,\theta_2)$ to be better recovered in the full distribution of θ_2 at any level of correlation.

A second interaction occurred between correlation of abilities and correlation of each ability estimate with its parameter. As $\rho(\theta_1,\theta_2)$ increased, $r(\theta_1,\hat{\theta}_1)$ decreased while $r(\theta_2,\hat{\theta}_2)$ increased (see Figures 8 and 9 above). Increased

$\rho(\theta_1,\theta_2)$ had more adverse affects on $r(\theta_1,\hat{\theta}_1)$ than $r(\theta_2,\hat{\theta}_2)$. The $\hat{\theta}_2$ appeared to depend more on θ_1 ability (i.e., $r(\theta_1,\hat{\theta}_2)$ increased as $\rho(\theta_1,\theta_2)$ increased and was larger in the B data sets than in the A data sets). This was not the case with $\hat{\theta}_1$ which did not seem to depend on θ_2 in either A or B data sets. The $\theta_1\theta_2$-space would appear to be collapsing. The distribution of the discrimination parameters may be contributing to this result as much as differentiated ability and correlation between abilities.

A third interaction was found as $se(\hat{a}_i)$ and $s(\hat{a}_i)$ were affected by correlation of abilities and differentiated ability on θ_2. As $\rho(\theta_1,\theta_2)$ increased, the $s(\hat{a}_1)$ decreased in the A data sets but increased in the B data sets, whereas $s(\hat{a}_2)$ increased in both the A and B data sets. In the A data sets, the $se(\hat{a}_1)$ decreased as $\rho(\theta_1,\theta_2)$ increased but increased in the B data sets. The $se(\hat{a}_2)$ decreased as $\rho(\theta_1,\theta_2)$ increased in both A and B data sets. There is an interaction effect on $s(\hat{a}_i)$ and $se(\hat{a}_i)$. Increasing the degree of correlation between abilities and a differentiated θ_2 ability combine to give poorer recovery of a_1. While it would be expected that the recovery may deteriorate in B data sets, it was not expected that increasing $\rho(\theta_1,\theta_2)$ would cause further deterioration. As the abilities became more correlated, more information is being used to estimate the second dimension ($se(\hat{a}_2) < se(\hat{a}_1)$ and $se(\hat{a}_2)$ decreases as $\rho(\theta_1,\theta_2)$ increases). As well, the $AAD(\hat{a}_i)$ both increased as correlation increased. These results might be related to the recovery of the mean of $\hat{a}_2$ as being larger than the mean of $\hat{a}_1$ and to the possible collapsing of the space. Clearly, the B samples didn't cover the ability space adequately. The rescaling of the θ_2 may be contributing to this interaction.

A fourth interaction was found between the correlation of abilities and differentiated ability on θ_2 and the mean of $\hat{D}$. Surprisingly, in the A data sets, as $\rho(\theta_1,\theta_2)$ increased, $\bar{\hat{D}}$ changed very little. In the B data sets, as $\rho(\theta_1,\theta_2)$ increased,

$\bar{\hat{D}}$ decreased (the items appear to be getting easier). This was as expected. Since D is a function of d and MDISC, and a_2 (a part of MDISC) was better estimated in the B data sets, this may explain why D became smaller (indicating easier items) but d did not change. A differentiated ability on θ_2 affected the size of the difficulty means moreso than the degree of correlation.

There are three issues of concern with respect to this research: the problems caused by the rescaling of the θ_2 estimates; the recovery of the two-dimensional space; and the dimensionality of the items.

The rescaling of the θ_2 estimates in the B data sets seemed to affect estimates of difficulty as well as estimates of thetas and discriminations. The estimates of means of d and D were adversely affected in the B data sets. The estimates of the mean of a_2 improved in the B data sets. It cannot be determined from the results reported here the extent of the effects of rescaling but it appears that the rescaling problem affects all parameter estimates somewhat.

The recovery of the structure of the ability space is also a concern. There was a tendency for the space to collapse as the abilities became more correlated. This may relate to a rotational indeterminacy in the recovery of the abilities. In the initial research design, some items pure on the second dimension were included in order to anchor the abilities in an attempt to improve the recovery of all parameters. It was recommended that this not be done as the test would not simulate the desired condition (M.D. Reckase, personal communication, December, 1986). This might be reconsidered in a future design. The collapsing of the space as $\rho(\theta_1,\theta_2)$ increased not only affected the theta estimates but also the discrimination estimates. In the B data sets, the structure of the latent space was recovered less well than in the A data sets. In retrospect, combining corresponding A and B data sets prior to analysis of the raw score vectors would provide a sample

which more typically represents the situation in which ESL students would likely be placed and would have allowed for better coverage of the $\theta_1\theta_2$-space. This might improve the estimation of some parameters and it would also partially eliminate the rescaling problem.

The third issue is the dimensionality of the item space. Twenty-six of the items were unidimensional (pure on a_1). The remaining 78 were two-dimensional, 52 requiring primarily ability on the first dimension for a correct response, 26 requiring equal amounts of both abilities. 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 a_2 over a_1, or α_2 over α_1. The poor recovery of the discrimination parameters also affected recovery of the difficulty and ability parameters. The item space seemed to become somewhat unidimensional. The estimates of the a_is were more alike and the size of the α_i angles moved towards 45° with α_2 becoming dominant. Since the range of a_2 was greater than that of a_1, this could have affected the dominance of a_2 over a_1.

Interpretation of parameter estimates appears to depend on the model, $\rho(\theta_1,\theta_2)$, and the characteristics of the data set. There is every indication from the results of this research that there are indeed three components of multidimensionality (subject dimensionality, test dimensionality, and the interaction of the two) as suggested by McKinley and 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 B data sets were low on ability in the second dimension.

As for the analogy of the ESL students, would those students be penalized in placement based on the results of this test? Clearly their raw scores on the tests were lower. As the ability dimensions became more correlated, the raw score for these students improved only slightly. McKinley and Reckase (1984) reported that $\rho(\theta_1,\theta_2)$ was an important factor in the latent ability structure. In terms of the recovery of the primary ability dimension, θ_1, the ESL students portrayed in the B data sets would have poorer recovery of this dimension as indicated by $r(\theta_1,\theta_1)$ and $AAD(\theta_1)$. If the M2PL model were chosen to represent the response data and MIRTE were used to analyze the data, these students would probably be penalized if their θ_1 estimates were used to determine placement. However, because of the rescaling, the question of how the ability estimates of the ESL students are affected cannot really be determined. If the A and B data sets had been pooled together, it might be possible to determine how the estimates of θ_1 are affected.

Strengths and Limitations

It is important to recall that the results presented above were based on 100 replications of each of the six data set conditions. There was very little variance on any of the summary statistics calculated over the 100 runs. The results can be considered somewhat stable. Monte Carlo simulation provides a means of systematically examining parameter estimation under a variety of conditions and of identifying the effects of various kinds of differences in the ability distribution. Since there are infinitely many multidimensional data sets possible, this research project is necessarily limited to a few of these possibilities. The strength of Monte Carlo studies is that the "real" parameters are known and can be compared with the estimates. The use of simulated data ensures close model-data fit, an issue of contention with real data sets.

Any IRT model is a simplification of reality. The task was to choose a model best representing the given situation. A compensatory M2PL model was chosen in order to more realistically simulate the characteristics and attitudes students would bring to a testing situation particularly those who may lack some ability. For any given ability vector the compensatory model gives an equal or higher probability of a correct response than the noncompensatory model. Two ability dimensions were chosen because most published tests require reading and another ability being tested. Two dimensions also provided a reasonable starting point to determine the accuracy of parameter estimation. The same item parameters were used throughout as a control on extraneous effects.

The limitations of the model must be considered. The choice of the M2PL compensatory model was made as it seemed to fit the problem more closely and an estimation program was available to analyze the data generated. Although the inclusion of a guessing parameter has been found previously to cause less stable estimates of other parameters (Muraki & Englehard, 1985), it is not known if this is true with MIRTE. The inclusion of the c-parameter coupled with the differentiated ability on the second dimension and differences in degree of correlation would have required a larger sample of items and examinees if reasonable estimates were expected. This was not within the scope of the research. Green, Yen, and Burket (1988) have reported advantages of using the three-parameter model. As a version of MIRTE now exists which includes the c-parameter, it would be beneficial to determine how well it functions. There is some question as to whether separate difficulty parameters for each of the dimensions might have improved the results of the study. A model such as that of Bogan and Yen (1983) could be used in a replication of the study.

With respect to the estimation program, it is not known how different

sample sizes of examinees or items affect the estimation procedures. The rescaling of the θ_2 estimates has adversely affected results. The rescaling problem relates to the basic assumption of item parameter estimates being person free and ability parameter estimates being item free. This holds only as long as there is coverage of the item parameters by abilities of the group. The MIRTE program arbitrarily fixes ability at mean 0, standard deviation,1. Perhaps some prior knowledge of the sample should allow for changes in this although it is seldom known in real life. It would be useful to have combined corresponding A and B data sets, as would be the case in real life, to better determine how restricted θ_2 ability affects θ_1 estimates.

The perception that the two-dimensional space was collapsing may be related to the test parameters selected. Use of a test more representative of the two-dimensional space might have resulted in better estimates for all parameters.

Indeed, the design of the research may be as important as the performance of the program. Two thousand subject estimates are available to estimate parameters for 104 items and vice versa making the estimation of the abilities more difficult. This accounts for the larger AADs and standard errors for the ability estimates.

A contribution made by this research towards future research using MIRTE is the knowledge that the program performed consistently, efficiently, and somewhat effectively thus eliminating the stringent necessity for and expense of a large number of replications of simulated data.

Suggestions for Future Research

Several questions remain at the conclusion of this research which suggest future studies. These are summarized briefly.

Are the results affected by the estimation procedures and/or the model

chosen? Replication of the research using different models (perhaps the M3PL model of Bogan and Yen (1983) or a noncompensatory model) would indicate to what extent model choice affected results. Inclusion of a guessing parameter in the model would provide additional information. A more recent version of MIRTE allows for inclusion of the c-parameter.

It would be useful as well to estimate item parameters only while holding the given ability parameters fixed and vice versa to determine further the efficiency of the MIRTE program. These results could be compared with those obtained when item and ability parameters are simultaneously estimated. Presumably both item and ability parameters would be better estimated. However, one could study the effects of each by varying the other parameters, i.e., specifying different conditions for item parameters in order to determine the effects on the ability estimates and vice versa.

Corresponding A and B data sets could be combined in order to present the ESL-type group in a large sample of wider variability more typical of a real life situation. This should solve some of the rescaling and space problems.

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

It would be useful to determine how well the ability dimensions were recovered at different ability levels rather than just at the mean level of ability although the standard errors, average absolute deviations and correlations do give some indication of overall recovery. This could be ascertained by looking at the θ-vectors in different sections of the $\theta_1\theta_2$-space and comparing the original (θ_1,θ_2)

with its estimate. It would also be useful to know how influential the second ability dimension became as the items required more of this ability for a correct response.

Another area of interest is that of item difficulty. Further analysis of the examinee results on easy versus difficult items at different ability levels would provide useful information for test builders.

A test with a wider range of discrimination values could determine how discrimination values affect recovery of item and ability parameters. Analysis of discrimination parameter recovery in different areas of the ability space could also be useful. Providing more items requiring both dimensions and some items pure on both dimensions would provide some indication of how the discrimination values need to be chosen to improve estimates. The poor recovery of the difficulty and discrimination parameters is a cause for concern as difficulty and discrimination are major selection criteria for test construction.

This research study provides encouraging results for those working in multidimensional item response theory. An important finding is the capability of MIRTE to retain the structure of the data and the people. 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 judicious to further develop estimation programs so that rotational solutions could be produced which might alleviate the tendency to collapse a two-dimensional space as the correlation between the dimensions increases.

Chapter 5

SUMMARY

This research study was designed to determine how well multidimensional IRT ability and item parameters would be estimated under certain specified conditions. The conditions were different degrees of correlation between the two ability dimensions and a differentiated ability on a second dimension.

The results of the research indicated that as the ability dimensions became more correlated, there was a tendency for the two-dimensional ability space to collapse. MIRTE tended to underestimate the degree of correlation between the ability dimensions but did not force orthogonality on the dimensions. Of the item parameters, the difficulty parameter was recovered most successfully. As the ability dimensions became more highly correlated, the discrimination parameter estimate for the predominant dimension (a_1) was underestimated while discrimination on the second dimension (a_2) was overestimated. The discrimination parameters in general were not well recovered. Increasing the correlation between the ability dimensions tended to result in even poorer recovery of the discrimination parameters. For correlated dimensions there appeared to be a confounding of item structure and ability structure as found by McKinley and Reckase (1984). The discrimination parameters did not cover the latent space adequately. In the recovery there was a tendency to spread the discrimination parameters over the entire latent space. This also occurred with the ability estimates and would indicate some rotational indeterminacy in the recovery of the multidimensional correlated latent space.

Restrictions on the second ability dimension resulted in poorer estimation for parameters of both ability dimensions. The differentiated ability on θ_2

appeared to cause a large shift in the estimates of d, underestimating the mean but retaining the internal structure of the item difficulties. The restrictions on the second ability dimension made the recovery of the discrimination parameters much worse than in the A data sets. The rescaling of the θ_2 estimates clearly affected the parameter recovery for the B data sets, particularly item difficulty.

Four interaction effects of correlation of abilities and a differentiated ability on θ_2 were noted. The correlation of abilities and differentiated ability on θ_2 interacted with recovery of $\rho(\theta_1,\theta_2)$, with recovery of the $r(\theta_i,\hat{\theta}_i)$, with the discrimination parameters (in $s(\hat{a}_i)$, $se(\hat{a}_i)$, and $AAD(\hat{a}_i)$), and with the recovery of the mean of the estimate of D. The rescaling of θ_2 and the poor coverage of the ability space and the item space partially explained these effects.

There were several aspects of the research design which made it different from previous research, the most obvious being the large number of replications (100) of each of the six simulated data set conditions. The conclusions based on these replications are stronger than those of most simulation MIRT studies which have tended to have few or no replications.

Previously researchers have considered effects of different degrees of correlation between ability dimensions but none has combined this with a differentiated ability on one of the dimensions in order to simulate the educational situation of ESL students, a growing segment of our student population.

Finally, this research study represents a stringent test of a relatively new multidimensional analysis program, MIRTE. The program has been used in some pilot testing situations and a few pieces of research involving only one replication but none to the extent it has been tested here.

The results of this research are encouraging for the future study of MIRT. The continued development of practical, useful and informative multidimensional

models coupled with the refinement and improvement of multidimensional estimation procedures are required if the necessary levels of accuracy of assessment and test fairness are to be achieved. Multidimensional item response theory has the capacity to provide required information about why certain groups of people perform as they do as well as to improve test construction and examinee assessment techniques. Multidimensional models, estimation procedures, and different latent space structures must continue to be explored.

REFERENCES

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

Ackerman, T. A. (1987a, April). *A comparison study of the unidimensional IRT estimation of compensatory and noncompensatory multidimensional item response data.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Ackerman, T. A. (1987b, 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, DC.

Ackerman, T. A., & Spray, J. A. (1986, April). *A general model for item dependency.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50(1),* 3-16.

Ansley, T. N. (1984). An empirical investigation of the effects of applying a unidimensional latent trait model to two-dimensional data. (Doctoral dissertation, University of Iowa, 1984). *Dissertation Abstracts International, 45/07, 2074A.*

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(1),* 37-48.

Ansley, T. N., & Forsyth, R. A. (1987, April). *The effects of applying the unidimensional one-parameter logistic model to multidimensional data.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17(4),* 283-296.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37(1),* 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika, 46,* 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35,* 179-197.

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, Quebec.

Carlson, J. E. (1987). *Multidimensional Item Response Theory Estimation: A Computer Program* (ACT Research Report 87-19). Iowa City, IA: American College Testing Program.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlation between items or between tests. *Psychometrika, 10,* 1-19.

Cattell, R. B. (1964). Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology, 55,* 1- 22.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum.

Choppin, B. (1982). The Rasch model for item analysis. In B. Choppin (Ed.), *A critical comparison of psychometric models for measuring achievement* (Methodology project). Washington, DC: National Institute of Education.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5-32.

Divgi, D. R. (1980, April). *Dimensionality of binary items: use of a mixed model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Doody, E. (1985). *Examining the effects of multidimensional data on ability and item parameter estimates using the three-parameter logistic model.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 258 992)

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22(4),* 249-262.

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

Embretson (Whitely), S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186.

Embretson, S. E. (1985). *Component latent trait models: A remedy for local dependence?* Paper presented at the Psychometric Society meeting, Nashville.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Gamache, L. M. (1983). *Comparison of traditional and latent trait procedures in analysis and selection of rating scale items.* (ERIC Document Reproduction Service No. ED 230 578)

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.

Green, D. R., Yen, W. M., & Burket, G. R. (1988, April). *Experiences in the application of item response theory in test construction.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test dimensionality. *Educational and Psychological Measurement, 37(4),* 827-838.

Hambleton, R. K. (1980). Latent ability scales: interpretation and uses. *New Directions for Testing and Measurement, 6,* 73-97.

Hambleton, R. K. (1982). The three-parameter logistic model. In B. Choppin (Ed.), *A critical comparison of psychometric models for measuring achievement* (Methodology Project). Washington, DC: National Institute of Education. (ERIC Document Reproduction Service No. ED 224 823)

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14(2),* 75-96.

Harper, D. (1972). Local dependence latent structure models. *Psychometrika, 37,* 53-57.

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

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19(1),* 49-78.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9(2),* 139-164.

Hoover, H. D. (1983). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice, Winter 1984,* 8-14.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item Response Theory: Application to Psychological Measurement.* Homewood, Il: Dow-Jones-Irwin.

Huynh, C-L., & Ansley, T. N. (1988, April). *An empirical study of grouping effects with chi-square statistics in fitting the three-parameter logistic model in IRT settings.* 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.

Jannarone, R. J. (1986). Conjunctive item response theory kernels.

Psychometrika, 51, 357- 373.

Jones, D. H., Wainer, H., & Kaplan, B. (1984). *Estimating ability with three item response models when the models are wrong and their parameters are inaccurate* (Research Report 84-26). Princeton, NJ: Educational Testing Service.

Kim, J., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues.* Beverly Hills, CA: Sage Publication.

Kingsbury, G. G., & Weiss, D. J. (1979, April). *Relationships among achievement level estimates from three characteristic curve scoring methods* (Research Report 79-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Kingston, N. M. (1986). *Assessing the dimensionality of the GMAT verbal and quantitative measures using full-information factor analysis* (Research Report 86-13). Princeton, NJ: Educational Testing Service.

Kolen, M. J., & Harris, D. J. (1987, April) *A multivariate test theory model based on item response theory and generalizability theory.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61,* 273-287.

Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh, 62A,* 74-82.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al., *Measurement and Prediction.* Princeton: Princeton University Press.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18(2),* 109-118.

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

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18,* 57-75.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989-1020.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247-264.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.*

Reading, Massachusetts: Addison-Wesley.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179-193.

Lumsden, J. (1961). The construction of unidimensional tests. *Psychological Bulletin, 58(2),* 122-131.

McCauley, C. D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9(4),* 389-400.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100-117.

McDonald, R. P. (1982). Unidimensional and multidimensional models for item response theory. In *Proceedings of IRT and CAT Conference,* Wayzata, MN. (ERIC Document Reproduction Service No. ED 264 264)

McDonald, R.P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27,* 82-99.

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, Quebec. (ERIC Document Reproduction Service No. ED 228 326)

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9(1),* 49-57.

McKinley, R. L., & Reckase, M. D. (1982a, May). *An analysis of the characteristics of a family of IRT models.* Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec.

McKinley, R. L., & Reckase, M. D. (1982b, March). *Multidimensional latent trait models.* Paper presented at the National Council on Measurement in Education, New York.

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). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report ONR-83-2). Iowa City, Iowa: American College Testing Program.

McKinley, R. L., & Reckase, M. D. (1983c). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation, 15(3),* 389-390.

McKinley, R. L., & Reckase, M. D. (1983d, April). *The use of item response theory analysis on dichotomous data from multidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

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

Mislevy, R. J. (1983). Item response theory for grouped data. *Journal of Educational Statistics, 8(4),* 271-288.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49(3),* 359-381.

Mislevy, R. J. (1985). *Bayes modal estimation in item response models* (Research Report 85-33). Princeton, NJ: Educational Testing Service.

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

Mulaik, S. A. (1972, March). *A mathematical investigation of some multidimensional Rasch models for psychological tests.* Paper presented at the annual meeting of the Psychometric Society, Princeton, NJ.

Muraki, E. (1985). *Full-information factor analysis for polychotomous item response.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Muraki, E, & Englehard, G. (1985). Full-information item factor analysis: applications of EAP scores. *Applied Psychological Measurement, 9(4),* 417-430.

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

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* (Vol. 4). Berkeley: University of California Press.

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. (1981, August). *The formation of homogeneous item sets when guessing is a factor in item responses* (Research Report 81-5). Columbia: University of Missouri, Department of Educational Psychology.

Reckase, M. D. (1985a). *Models for multidimensional tests and hierarchical structured training materials* (Research Report ONR-85-1). Iowa City, Iowa: American College Testing Program.

Reckase, M. D. (1985b, April). *The difficulty of test items that measure more than one ability.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Reckase, M. D. (1985c, August). *Trait estimates from multidimensional items.* Paper presented at the annual meeting of the American Psychological Association, Los Angeles, CA.

Reckase, M. D. (1986, April). *The discriminating power of items that measure more than one dimension.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M. D., & Ackerman, T. A. (1986, April). *Building a test using items that require more than one skill to determine a correct answer.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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. (1982a, July 27-30). *Some latent trait theory in a multidimensional latent space.* Paper presented at the Invitational Conference on IRT and CAT, Wayzata, MN.

Reckase, M. D., & McKinley, R. L. (1982b, August 23-27). *The feasibility of a multidimensional latent trait model.* Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Reckase, M. D., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional IRT models.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Reiser, M. (1983). An item response model for the estimation of demographic effects. *Journal of Educational Statistics, 8,* 165-186.

Rogers, H. J., & Hambleton, R. K. (1987, April). *Evaluation of computer simulated baseline statistics for use in item bias studies.* Paper presented ant the annual meeting of the American Educational Research Association, Washington, DC.

Rogers, H. J., & Hattie, J. A. (1987). A monte carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11(1),* 47-57.

Samejima, F. (1973a). A comment on Birnbaum's three parameter logistic model in latent trait theory. *Psychometrika, 38,* 221-233.

Samejima, F. (1973b). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203-219.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39,* 111-121.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical testing theory. *Psychometrika, 42(2),* 193-198.

Slaughter, R. E., & Delucchi, K. L. (1986, April). *The quality of random number generation on the IBM Personal Computer: Implications for monte carlo simulation methods.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Slinde, J. A., & Linn, R. L. (1979). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement, 16,* 159-165.

Stegelmann, W. (1983). Expanding the Rasch model to a general model having more than one dimension. *Psychometrika, 48(2),* 259-267.

Stout, W. (1984). *A statistical procedure for assessing test dimensionality.* Measurement Series 84-2. Illinois University, Urbana, Il. (ERIC Document Reproduction Service No. ED 249 281)

Sympson, J. B. (1978). 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.

Takane, Y., & de Leeuw, J. (1986). *The relationship between item response theory and factor analysis of discretized variables* (Natural Sciences and Engineering Research Council of Canada, Grant A6394). Montreal, Quebec: McGill University.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7,* 215-232.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of Item Response Theory.* Vancouver, British Columbia: Educational Research Institute of British Columbia.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1986, April). *The effects of two-dimensional data on unidimensional IRT parameter estimates.* Paper presented at the annual meting of the American Educational Research Association, San Francisco, CA.

Weiss, D. J., & Suhadolnik, D. (1982, July). Robustness of adaptive testing to multidimensionality. *Proceedings of the IRT and CAT Conference,* Wayzata, MN. (ERIC Document Reproduction Service No. ED 264 270)

Wherry, R. J., & Gaylord, R. H. (1944). Factor pattern of test items and tests as a function of the correlation coefficient: Content, difficulty, and constant error factors. *Psychometrika, 9,* 237-244.

Whitely, S. E. (1980). Multi-component latent trait models for ability tests. *Psychometrika, 45,* 479-494.

Whitely. S. E. (1981). Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement, 18,* 67-84.

Wilson, D., Wood, R. L., & Gibbons, R. (1984). *TESTFACT: Test Scoring and item factor analysis.* [Computer program]. Moorsville, IN: Scientific Software, Inc.

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

Wright, B. D., & Panchapakesan, N. A. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17(4),* 297-311.

Yen, W. M. (1985). Increasing item complexity: a possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50(4),* 399-410.

Zwick, R. (1988, April). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24(4),* 293-308.