UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MULTIDIMENSIONAL ITEM RESPONSE THEORY: A SAS MDIRT MACRO AND EMPIRICAL STUDY OF PIAT MATH TEST

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

SUNG-HYUCK LEE

Norman, Oklahoma

2007

UMI Number: 3255213

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI 3255213
Copyright 2009 by ProQuest LLC.
All rights reserved. This edition of the work is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

MULTIDIMENSIONAL ITEM RESPONSE THEORY : A SAS MDIRT MACRO AND EMPIRICAL STUDY OF PIAT MATH TEST

A DISSERTATION APPROVED FOR THE

DEPARTMENT OF PSYCHOLOGY

BY

Joseph L. Rodgers

Robert Terry

Larry E. Toothaker

Jorge L. Mendoza

Craig J. Russell

© Copyright by SUNG-HYUCK LEE 2007

All Rights Reserved

ACKNOWLEDGMENTS

I want to give a special thank to my beloved wife, Su-Jung Park who passed away from stomach cancer last summer. She showed and taught me what it was to be a Christian throughout her fight with the disease. She was never afraid of her death because she believed that the Heaven where she would take her everlasting rest with God was the ultimate end of our life. The memory of her sacrificial and positive attitude to the life will last in my mind as an unforgettable blessing from God.

I want to express my sincere gratitude to Dr. Joseph Rodgers who was my major advisor. This dissertation could not have been written without his guidance and encouragement. He not only served as my supervisor but also was a very supportive and reliable friend throughout my personal crisis. He and the other faculty members, Dr. Robert Terry, my co-advisor, Dr. Larry Toothaker, Dr. Jorge Mendoza, and Dr. Russell Craig patiently guided me through the dissertation process, I thank them all.

I can't miss giving my thanks to brothers and sisters at Norman Korean Baptist Church. Without their spiritual and financial support I might have been lost in my personal heartbreaking circumstance. I owe my wholehearted appreciation to my parents and every member of my family who sacrificed their life for me. Finally, I thank God for providing me with such wonderful committee members, sincere friends, faithful brothers and sisters and a supportive family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

ABSTRACT

Even though unidimensional item response theory (IRT) provides a better framework for practical test settings than classical test theory (CTT), theoretical and empirical evidence shows that most response data violate the assumption of unidimensionality. There are several computer programs dedicated to estimating parameters based on the multidimensional perspective (MIRT). However, their accessibility is still costly, and they are not easy to use. In this paper, we present a SAS macro called MDIRT-FIT, to increase accessibility to the benefits obtained from this recent measurement theory development. The program is developed to estimate parameters based on a compensatory multidimensional item response theory (MIRT) model for dichotomous data. The full information item factor analysis model with an EM algorithm suggested in Bock & Aitken (1988) is implemented in the SAS programs. The estimation program written in SAS/IML® provides both parameters of MIRT and parameters of the factor analysis model with their associated standard errors and overall model fit statistics. The maximum number of latent traits that can be estimated with this program is limited to five latent dimensions because of both computational burden and practical sufficiency. The accuracy and stability of the SAS macro is examined by utilizing simulated data of examinees' responses. The PIAT math test, a subset of the Peabody Individual Achievement Test, was examined to validate the comparability of the SAS macro to TESTFACT which is widely used for estimating parameters of MIRT models.

Multidimensional Item Response Theory: A SAS MDIRT Macro and Empirical Study of PIAT Math Test

Introduction

Emergence of Modern Test Theory

Owing to its role within society, measurement has been a hot area for a long time. For example, society has encouraged measurement experts to invent precise instruments to measure intelligence, and has used intelligence as a selection criterion. Besides being used for selection, measurement has often been used as a tool for assignment of individuals to various positions of a society. On the other hand, members of society have a strong desire to actualize themselves. Being selected is important because it means successful competition and a chance to fulfill oneself. Therefore, measurement by testing has been a subtle and complicated arena where the roles of society and the individual may often collide.

Realizing the critical role of measurement to society and its members, measurement experts started developing fair measurement procedures that satisfy both society and the individuals. Since then, good measurement tools have been sought so that society could select qualified applicants, and its members could be tested by a fair test. In order to make a good test what was needed was a good test theory quantifying the

property of an object being observed, assigning a number to it, and providing for interpretation of the meaning of the number assigned. What was additionally required of a desirable test theory was that it should be not only theoretically plausible but could also be empirically fit to observed data and practically applied to real world settings. Several decades of effort searching for the desirable characteristics of a good test theory were captured in the concepts of validity and reliability of a test, which are believed prerequisite for a good test theory.

However, constructing a valid and reliable test is a very complicated process, especially when the attribute of an object being measured is not observable (e.g. intelligence). In measurement, the abstract and philosophical construct is called various names such as latent ability, attribute, factor, or dimension (Hambleton & Swaminathan, 1985). The underlying latent variable is not directly observed but is measured by way of manifest variables which are assumed to have a significant correlation with the targeted latent trait. What has been needed was a reasonable latent trait theory that stated the functional relationship between observed variables and the underlying latent trait.

It was classical test theory (CTT) that drew the attention of psychological and educational researchers in their attempt to measure the underlying latent trait. CTT provides a simple model which states that the observed score on a test is the sum of the true score and measurement error. It assumes that the expected mean of the observed score is equal to the true score underlying the test performance of an examinee if she or he is administered the same test infinitely many of times. CTT has been helpful in developing important concepts for item and test analysis, and test construction. Those

concepts include the reliability coefficient, the correlation between replicated measurements, the Spearman-Brown formula for test lengthening, Kuder-Richardson's coefficients for internal inconsistency of a test comprising binary items, Cronbach's alpha for internal consistency, and the corrections for attenuation and validity of a test.

However, CTT has exposed many undesirable features in practical testing situations because it heavily relies on the concept of parallel forms, which are nearly impossible to achieve in realty. Researchers must be content with either lower-bound estimates of reliability or reliability estimates with unknown biases (Hambleton and van der Linden, 1982).

Another test theory which is more appropriate for measuring an underlying latent trait and overcoming the limitations of CTT is item response theory (IRT). The concept of modern IRT traces back to Thurstone (1925), when he tried to arrange the items of the Binet & Simon test used for estimating children's metal age on an age-graded scale (Bock, 1997). Thurstone also plotted each item with the percentage of correct responses to an item on the vertical axis and chronological age of a respondent on the horizontal axis. His work helped practitioners at that time to estimate children's mental age because they could rank test items from easy to hard on the continuum of mental intelligence and narrow them down by administering items with high discriminating power for a specific mental age. It is noticeable that the S-shaped plot he drew is very similar to the modern item characteristic curve, which is at the core of IRT.

IRT is a modern test theory that provides a more intuitive and more informative measurement model than classical test theory. IRT is more intuitive than its predecessor

in that, unlike classical test theory, IRT has accounted for the variability of items in their difficulty and discrimination power as a source of variability in human performance. In other words, IRT makes it possible to measure proficiency level on the latent trait after controlling for nuisance variables (e.g. item difficulty, item discrimination, pseudo guessing parameter). IRT is more informative than CTT in that it allows for administrators to identify the underlying proficiency level of test takers more precisely with smaller measurement error.

IRT provides many advantages that its old competitor does not in practical testing situations. First, the estimated item parameters are invariant with regard to who is sampled from the population, and the estimated proficiency level remains constant regardless of which items are administered if the item characteristic curve fits the given data well. In CTT, the estimated item parameters depend on the characteristics of a group of respondents. The calibrated item difficulty tends to be higher for test takers with a high proficiency level than test takers with low proficiency level. The estimated item discrimination tends to be higher for a heterogeneous group than a homogeneous group on the latent trait. In other words, the usefulness of estimated item parameters is limited to the group from which they are obtained and the meaning of the estimated proficiency levels is limited to the item set from which they are obtained (Hambleton, 1991).

Second, IRT is more suitable than CTT when the proficiency levels of examinees are compared. CTT assumes that the standard error of a test is constant across the entire continuum of the targeted latent trait. However, this assumption is not usually correct, because the standard error is minimized when the overall test difficulty is matched with

the underlying latent trait level of an examinee, and it is larger for those with a high or low proficiency level. A standardized test based on CTT gives the most valid measurement of examinees whose proficiency level is close to the test difficulty. Therefore, comparing two test scores from different sets of items is not easily handled because they are not on the same scale. Through IRT, two test scores from two different item sets can be easily equated using linking items, because they are compared on the same scale (e.g. they are linearly related).

Third, IRT can provide a more efficient test in terms of test length or measurement precision. Typically, with a standard test based on CTT an examinee is required to respond to all items, from the easiest item to the hardest item, to estimate the proficiency level underlying his or her response to the items. The main reason that a standard test does not give a precise estimate for those who have a high or low proficiency level in terms of standard error is that the difficulty of items is not matched with their proficiency level. It is well known that when hard items are administered to respondents with low proficiency level or easy items are given to respondents with high proficiency level, measurement error increases. On the other hand, IRT allows the test administrator to select an item approximately matched with the proficiency level of an examinee, because it is possible to predict how an examinee performs on a particular test item using the item characteristic curve.

Fourth, CTT depends on the important parameters of the latent trait (true score) on the assumption that strictly parallel tests are available (Lord, 1980). It is nearly impossible to meet this assumption in reality because it requires the same mean, same

variance between two tests, and same covariance with other criterion tests. However, IRT provides the concept of test information, which replaces the reliability index in CTT.

Fifth, IRT makes computerized adaptive test (CAT) come to the life. In CAT, test items are automatically selected by the computer from an item bank containing calibrated items. In a traditional paper and pencil tests based on CTT, examinees are required to attempt exactly the same items from the easiest and the hardest regardless of their proficiency level, which increases the measurement error of examinees whose proficiency level is higher or lower than average. However, CAT based on IRT can successively present items matched to the current estimated proficiency level of an examinee, depending on the previous responses of the examinee, until some desired measurement precision is reached. Those procedures result in more efficient testing than that based on CTT, in terms of test length or measurement precision.

Overview of Unidimensional IRT Models

Item Characteristic Curve (ICC)

The item characteristic curve (ICC) is at the core of IRT in that the validity of the process of estimating parameters entirely depends on its appropriateness to model the testing behavior of examinees. It is a probabilistic model that describes the interaction between the underlying proficiency level of an examinee and the item parameters (Ackerman, 1992). ICC is a mathematical function that predicts the probability that a respondent at a certain proficiency level makes a successful response to an item. When the model fits response data, it allows test administrators to predict how a particular

examinee performs on a particular calibrated item. Therefore, in CAT, the proficiency level of an examinee can be estimated more precisely and more efficiently by presenting items matched with the examinee's latent trait level.

The Two-Parameter Normal Ogive Model

The normal ogive model (cumulative normal distribution) was used to fit the data from early psychophysical experiments. It was chosen because of not only its similarity to the observed data from early psychophysical experiments but also its well-established mathematical characteristics. The mathematical form of the normal ogive model is expressed below.

$$P_i(\theta_j) = P(x_i = 1 \mid \alpha_i, \beta_i, \theta_j) = \int_{-\infty}^{z_i=\alpha_i(\theta_j-\beta_i)} \frac{1}{\sqrt{2\pi}} e^{-(z_i^2/2)} dz \text{ , where} \quad (1)$$

$P_i(\theta_j)$ is the probability that examinee *j* correctly responds to item *i*

x_i is an actual response to item *i* (1 = correct response, 0 = incorrect response)

α_i is discrimination parameter of item *i*

β_i is difficulty parameter of item *i*

θ_j is trait parameter of examinee *j*

z_i is limit of the integral, $\alpha_i(\theta_j - \beta_i)$, in standard deviation units

Even though the normal ogive model was first used for the ICC, the logistic ogive model is preferred because of its similarity to the cumulative normal distribution

and its computational efficiency. Haley (1952) showed that the difference between the normal ogive model and the logistic ogive model is smaller than 0.01 across the entire continuum of the latent trait in terms of predicted probability of correct response (when the 1.702 adjustment constant is used). In addition, it is an attractive alternative to the normal ogive model because it does not involve numerical integration that causes additional computational burden. In this dissertation, two computer programs (SAS IRT-FIT and SAS POLYIRT-FIT) based on the logistic ogive model are developed to analyze dichotomously-recorded items, which are most commonly used in general test settings.

The Two-Parameter Logistic Model (2PLM)

The two-parameter logistic model (2PLM) has two item parameters that determine the shape of the item characteristic curve. Item difficulty is a location parameter that is determined by the θ value at which an examinee has a probability of 50% for a correct response. A difficult item is located to the right (large θ) and an easy item is located to the left (small θ), as shown in the Figure 1 (compare item 1 with item 2), when the other parameters are held constant. The discrimination parameter is proportional to the slope at the inflection point. When the other parameters are fixed, an item with high discriminating power has a steeper slope than an item with low discriminating power, as shown in Figure 1 (see item 2 and item 3). The probability of a correct response given the item parameters and the ability parameter is defined as below.

$$P_i(\theta_j) = P(x_i = 1 \mid \alpha_i, \beta_{i,} \theta_j) = \frac{1}{1 + e^{-D\alpha_i(\theta_j - \beta_i)}}, \text{ where} \quad (2)$$

$P_i(\theta_j)$ is the probability that examinee *j* with θ getting the item *i* correct

e is 2.718, the base of the natural logarithm

D is 1.702, which is a scaling factor

x_i is a response to item *i* (1 for correct response, 0 for incorrect response)

α_i is discrimination parameter of item *i*

β_i is difficulty parameter of item *i*

θ_j is ability parameter of examinee *j*

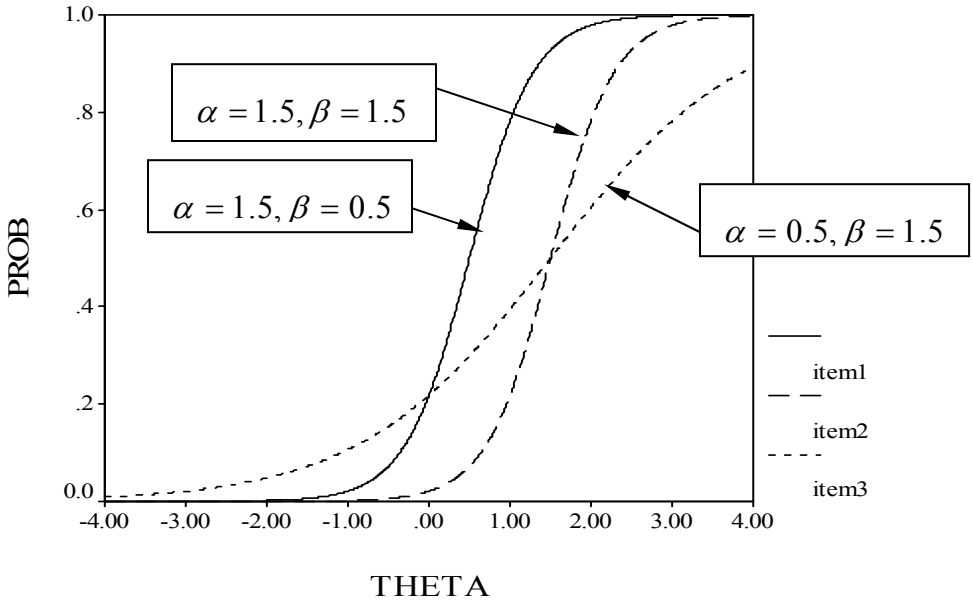

Figure 1. ICC of two-parameter logistic model

The Three-Parameter Logistic Model (3PLM)

In the three-parameter logistic model (3PLM), a pseudo guessing parameter is added to model the phenomenon that an examinee with a low ability gets an item correct by chance. The pseudo guessing parameter is a lower asymptote of the item characteristic curve, as shown in Figure 2. It is notable that in the 3PLM, the difficulty parameter is not defined on the trait continuum where an examinee has a 50% chance to get the item correct. It is determined where a respondent has [c + (1-c)/2] chance to get the item correct (e.g. difficulty of item 2 = 0.2+(1-0.2)/2 = 0.6) because of the effect of the guessing parameter on the ICC. The 3PLM becomes identical to the 2PLM when the pseudo guessing parameter is zero. The mathematical form of the three-parameter logistic model is expressed as below.

$$P_i(\theta_j) = P(x_i = 1 \mid \alpha_i, \beta_i, c_i, \theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-D\alpha_i(\theta_j - \beta_i)}}, \text{ where} \qquad (3)$$

c_i is guessing parameter of item *i*, and

all the other components are defined as for the two-parameter logistic model.

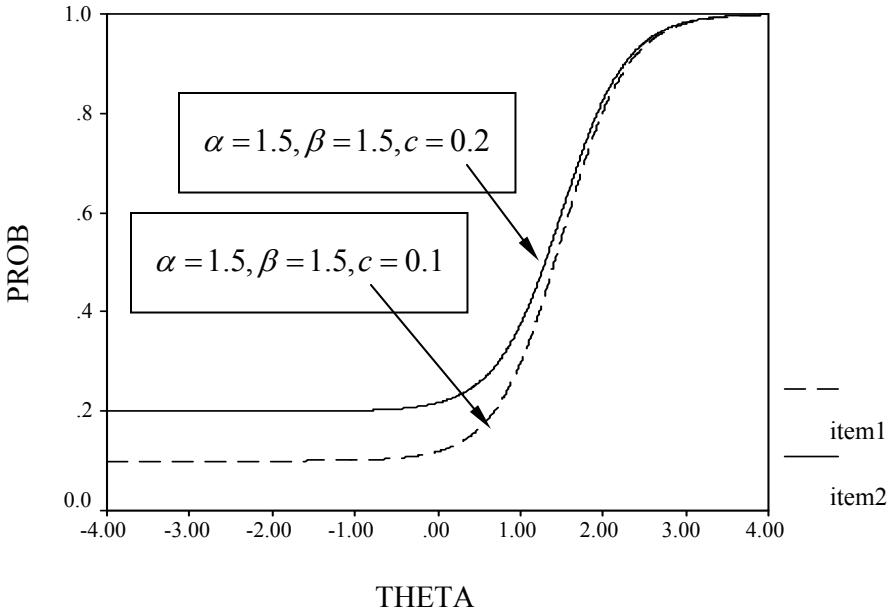

Figure 2. ICC of three parameter logistic model

The Graded Response Model for Polytomous Data

Many psychological testing instruments measuring attitude or personality include multiple ordered-categories items (e.g Likert-type item) for which the previous models are not appropriate. Samejima (1969) suggested an IRT model that can handle items that have multiple categories and are recorded in an ordered fashion. It is an extension of the two parameter logistic model in that the same model is used for boundary characteristic curves (Baker, 1992) that are needed to estimate boundary threshold parameters. Boundary threshold parameters provide the major advantage of Samejima's polytomous model over the previous IRT model for dichotomous data. The latent trait level can be more precisely estimated because the successive boundary threshold parameters can extract more information about the underlying trait level of an examinee than a single

threshold parameter. The strategy Samejima used to generate boundary characteristic curves is to treat the response of examinees as dichotomously recorded. For instance, 1 is assigned to the first category as a correct response and 0 is assigned to the remaining m-1 categories as incorrect responses. For the second boundary characteristic curve, 1 is assigned to categories 1 and 2 as a correct response while 0 is assigned to the rest of the categories as incorrect responses, and so on. The model for boundary characteristic curves is expressed as

$$P_k^*(\theta_j) = \frac{1}{1+e^{-D\alpha_i(\theta_j-\beta_{ik})}} \quad \text{, where} \tag{4}$$

k = 1, 2 …m-1 (m is the number of categories),

$P_k^*(\theta_j)$ is the probability that examinee j responds to category k or a lower category

$P_0^* = 1$, $P_m^* = 0$,

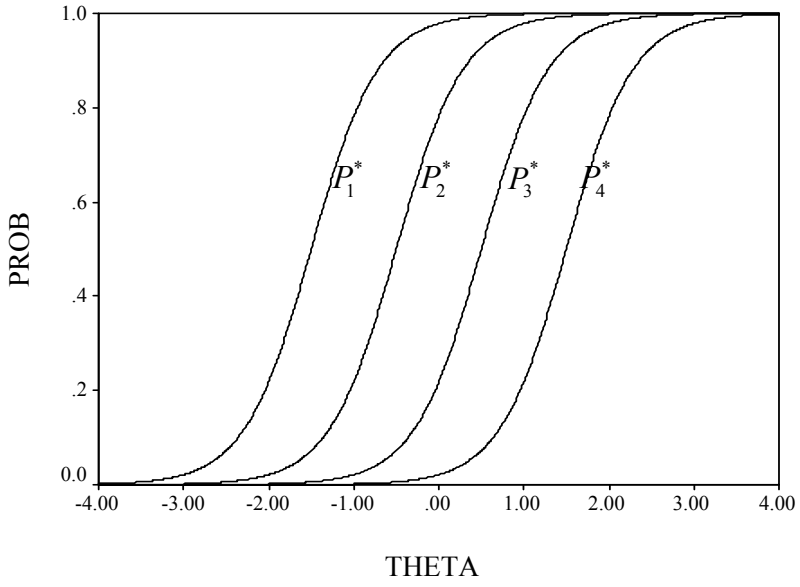

Figure 3. Boundary Characteristic Curves of Graded Response Model

In Figure 3, four boundary threshold parameters (-1.5, -0.5, 0.5, 1.5) and a common discrimination parameter of 1.5 are used for a graded response item with five ordered categories. Boundary characteristic curves are used to produce item response category characteristic curves (IRCCC) that define the probability that an examinee responds to a particular category of an item (Baker, 1992). The relationship between the boundary characteristic curve and IRCCC is expressed as

$$P_k(\theta_j) = P^*_{k+1} - P^*_k = \frac{1}{1+e^{-D\alpha_i(\theta_j-\beta_{iK+1})}} - \frac{1}{1+e^{-D\alpha_i(\theta_j-\beta_{ik})}}, \text{ where} \quad (5)$$

$P_k(\theta_j)$ is the probability that examinee j responds to category k (k = 1, 2, . . . m) and $\sum_{k=1}^{m} P_k = 1$.

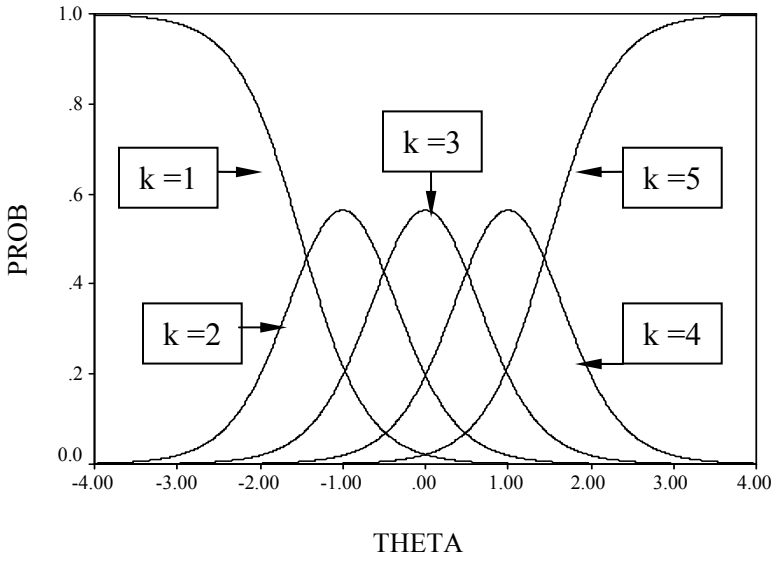

Figure 4. IRCCC of Graded Response Model

The narrower and the more peaked the IRCCC's are, the better they are at discriminating among latent trait levels (Embretson and Reise, 2000). Boundary threshold parameters are determined at the intersection where two adjacent IRCCC's meet.

Rasch vs. Two or Three Parameter Models

The one-parameter logistic model (known as Rasch model) is not included in this dissertation not because it is unimportant, but because the procedure for estimating parameters is relatively easy. However, it has interesting features that the 2PM and 3PM do not. First, it estimates only one item parameter (difficulty), assuming all items have an equal discriminating power. Second, a different estimation method, called conditional maximum likelihood (CLM), is utilized. The sufficient statistic (the sum of correct responses) for the underlying latent trait is available in the estimation process. Third, it is based on a different philosophy about measurement. It puts more emphasis on obtaining desirable measurement properties (called specific objectivity) than capturing all aspects of the observed response data in developing item response models.

This has been a controversial issue between these two different philosophical lines throughout the history of IRT models. Theorists following Rasch have asserted that IRT models providing specific measurement property should be preferred. In fact, it is only the Rasch model that provides a kind of ratio scale for estimated ability and item parameters. To be more specific, it is true in the Rasch model that two estimated proficiency levels can be compared in terms of the probability of correct response to an item without reference to item parameters, and vice versa. On the other hand, theorists

advocating the use of 2PLM and 3PLM have insisted that models should not determine the response data and the empirical data themselves should determine the properties of the model. Thissen (2001, p. 90) mentioned that "the properties of the model for measurement are a consequence of the observed item-response data, as summarized by the parametric model: Items are assumed to measure as they do, not as they should."

Advent of Multidimensional Item Response Theory (MIRT)

Advantages Multidimensional IRT over Unidimensional IRT

Even though unidimensional IRT gives better solutions to the test practitioner than CTT, a more sophisticated test theory like multidimensional item response theory (MIRT) has been required to accommodate a complex reality. The major problem with unidimensional IRT is that it cannot handle many empirical data that are potentially multidimensional. If the unidimensional IRT model does not fit a set of response data, we might suspect that the test is measuring more than two underlying latent traits. Many researchers have shown that psychological factors like cognitive skill, motivation, and test-anxiety, as well as the targeted latent trait, have an effect on the test performance of an examinee. In reality, it is almost impossible to make a test that purely measures a single latent trait only, which is especially true when a test measures a latent construct related to human cognition.

Using multidimensional items to measure a single trait may weaken the construct validity of a test, because the construct validity of the test depends on the degree to which a theoretical construct and a specific measuring tool or a procedure agree. When the

assumption of unidimensionality is violated, the main source that decreases construct validity comes from the fact that there is no unique one-to-one matching between multidimensional latent spaces and the targeted unidimensional space (Ackerman, 1992). This means that it is not even guaranteed that examinees at the same unidimensional proficiency level are measured on the same composite of multiple abilities when they are administered multidimensional items.

What makes it worse is that multidimensional trait dimensions are not independent of item difficulty (Reckase, 1985; Reckase, Carlson, Ackerman, & Spray, 1986; Ackerman, 1989; Ackerman, 1991). They showed that easy items tend to measure the targeted trait and difficult items tend to measure the nuisance trait. This implies that examinees taking a CAT might be measured on completely different composites of multiple latent traits, depending on their unidimensional proficiency level (Ackerman, 1991). The construct validity of CAT based on unidimensional IRT can be secured only when the assumption of unidimensionality is satisfied. The net result is that the more a test deviates from unidimensionality, the more the construct validity of the test is likely to decrease.

Another limitation of applying a unidimensional IRT model to a multidimensional test is well documented by many researchers (Ackerman, 1992, 1996; Reckase, 1985; Reckase & McKinley, 1991). Ackerman explained why examinees cannot be ranked in order as is done on a single dimension when the assumption of unidimensionalty is violated. From the unidimensional IRT perspective, the justification of assigning a meaningful score to the latent construct of an examinee entirely depends

on the extent to which a test measures the latent trait. Furthermore the invariant property of the estimated item parameters and estimated latent trait level does hold only if the assumption of unidimensionality is satisfied. Therefore, when a test is measuring more than two underlying latent traits, ordering examinees on one dimension does not make sense because they are identified as coordinates on multiple latent trait spaces. From a multidimensional IRT perspective, the concept of invariant estimated item parameters and proficiency level need to be modified because of the multidimensionality of test items. Estimated item parameters are invariant and proficiency level estimates of examinees remain constant only if test items are measuring the same composite of multiple latent traits.

In addition, the concept of differential item function (DIF) based on unidimensional IRT can be more clearly comprehended from a multidimensional perspective. It is well known that the main source of differential item function (DIF) is that test items are measuring unintended latent traits. Oshima and Miller (1992) examined how often multidimensional items whose means on nuisance trait for two subgroups are different can be identified as biased items using item bias indices based on a unidimensional IRT. Multidimensional items in this study are detected as biased most of the times, albeit the detection rate of DIF depends on the number of multidimensional items, discrimination power of those items, and the degree to which those items measure the nuisance trait. The result agrees with what Ackerman (1992) suggested for potentially biased items in which the conditional means or variances of subgroups on a nuisance trait are different.

Finally, the computerized adaptive test (CAT) based on MIRT gives advantages over CAT based on traditional IRT. Segall (1996) showed that administering multidimensional items not to measure a single trait but to measure multiple traits at the same time could obtain substantial gains over administering separate unidimenisional CAT's. First, he demonstrated that the multidimensional item selection strategy incorporating a Bayesian framework could considerably increase the efficiency of a measurement. When the dimensions are correlated, using the prior knowledge of one dimension for the rest of the dimensions in item selection can result in the reduction in test lengths or greater precision. In addition, multidimensional CAT provides more adequate and efficient coverage of content in CAT, which is almost impossible in unidimensional CAT because of the confounding between item difficulty and latent dimensions. Therefore, when the assumption of unidimensionality is violated, MIRT is an attractive alternative to unidimensional IRT because the multidimensional perspective may provide a better way to get around the problems that unidimensional IRT cannot easily handle.

Overview of MIRT Models

Compensatory Multidimensional 3PL Model

There are two types of MIRT models, depending on whether compensation of high proficiency on one trait for low proficiency on other traits is available or not. Reckase (1983) suggested a multidimensional extension of the three-parameter logistic model. In the model, the probability of a correct response to an item is linearly related to

the sum of k weighed proficiency levels. Also, it is evident that the weights represent the impact on the dimensions and they are compensating for each other. The mathematical function is expressed as below.

$$P(x_{ij}=1 \mid a_i, d_i, c_i, \theta_j) = c_i + (1-c_i)\frac{1}{1+\exp[-D(\sum_{k=1}^{g} a_{ik}\theta_{jk}^{'} + d_i)]}, \text{ where} \qquad (6)$$

D is scaling constant which is 1.702

a_i is a row vector of discrimination parameters of item *i* on k dimensions (k = 1, 2, … g)

d_i is a difficulty parameter of item *i*

c_i is a guessing parameter of item *i*

θ_j is a row vector of ability parameters of examinee *j* on k dimensions (k = 1, 2, … g)

Noncompensatory Multidimensional 3PL Model

The noncompensatory MIRT model was suggested by Sympson (1978). Unlike compensatory MIRT, the probability of successful performance on an item depends on the product of the success on each of the underlying dimensions. Therefore, failure on any dimension ends up with failure to solve the item. In other words, higher proficiency on one dimension does not compensate for lower proficiency on the other dimensions.

$$P_i(\theta_j) = \phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-z_i}^{\infty} e^{(-1/2)Z^2} dz$$

$$P(x_{ij} = 1 \mid a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i)\prod_{k=1}^{g} \frac{1}{1 + \exp[-Da_{ik}(\theta_{jk} - b_{ik})]}, \text{ where} \qquad (7)$$

D is scaling constant which is 1.702

a_{ik} is discrimination parameter of item *i* on k[th] dimension (k = 1, 2, … g)

b_{ik} is difficulty parameter of item *i* on k[th] dimension (k = 1, 2, … g)

c_i is a guessing parameter of item *i*

θ_{jk} is ability parameter of examinee *j* on k[th] dimension (k = 1, 2, … g)

Graphical Representation of Multidimensional Items

Displaying a multidimensional item in the space of multiple latent traits is restricted to the two-dimensional case because of the limitation of visual representation. McKinley & Reckase (1983) proposed a multidimensional two parameter logistic model (M2PLM) which is a direct extension of unidimensional 2PLM. The mathematical form of the model is expressed as below and examples of six multidimensional items are presented using the model.

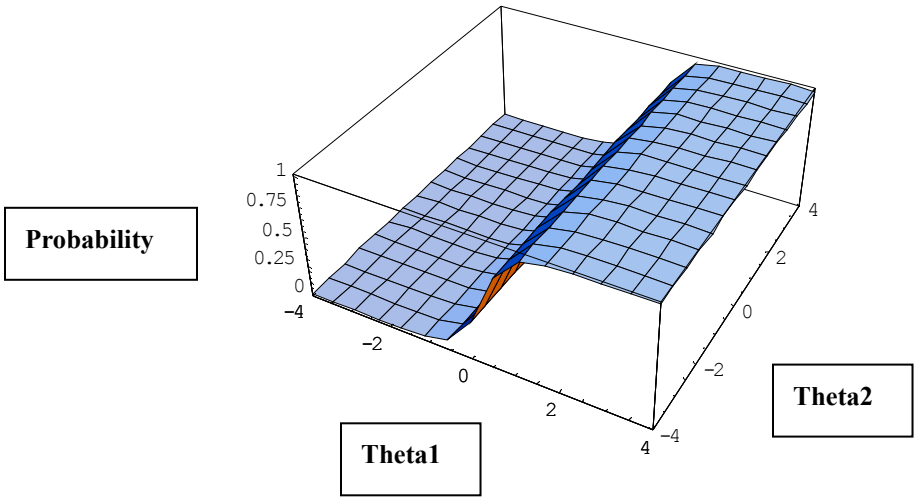

Figure 5a. IRS of a Multidimensional item with $\alpha_1 = 2.0, \alpha_2 = 0, d = 0.5$

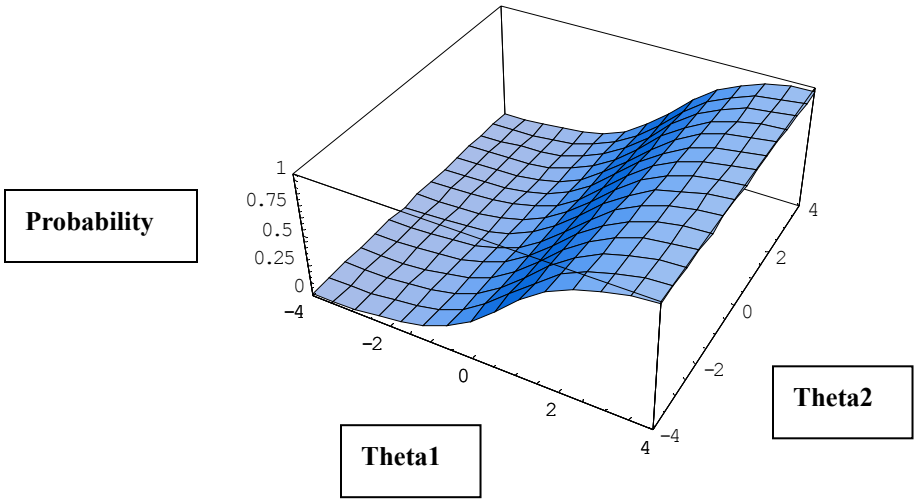

Figure 5b. IRS of a Multidimensional item with $\alpha_1 = 0.7, \alpha_2 = 0, d = 0.5$

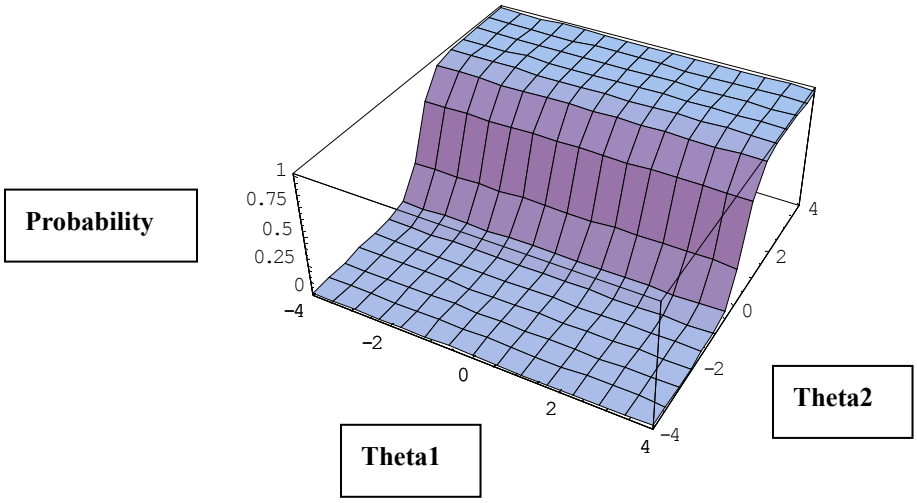

Figure 5c. IRS of a Multidimensional item with $\alpha_1 = 0, \alpha_2 = 2.0, d = 0.5$

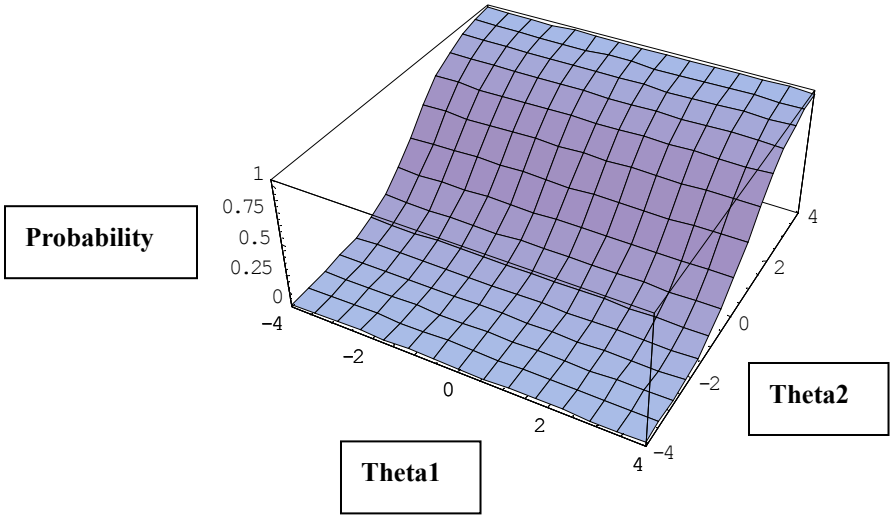

Figure 5d. IRS of a Multidimensional item with $\alpha_1 = 0, \alpha_2 = 0.7, d = 0.5$

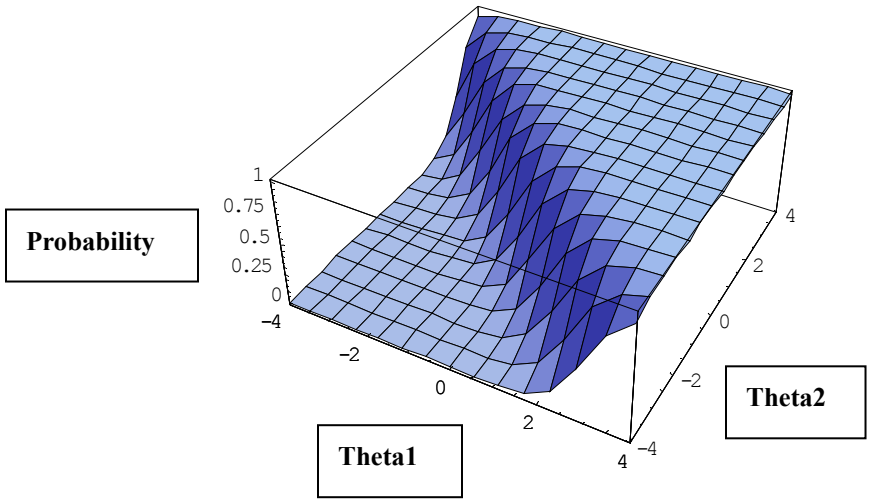

Figure 5e. IRS of a Multidimensional item with $\alpha_1 = 1.5, \alpha_2 = 1.5, d = -1.5$

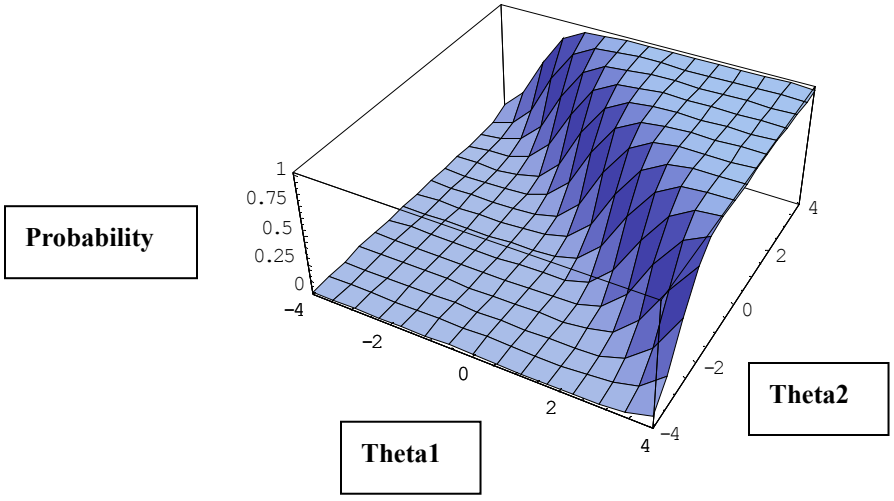

Figure 5f. IRS of a Multidimensional item with $\alpha_1 = 1.5, \alpha_2 = 1.5, d = 1.5$

Reckase (1885) named the multidimensional trace surface of a multidimensional item as the item response surface (IRS) rather than ICC as in unidimensional IRT. The IRS has different characteristics from undimensional ICC. First, the probability of correct response monotonically increases as each level of latent traits increases and low proficiency level on one dimension can be compensated for high proficiency level on any other dimension. Second, the shape of The IRS depends on not only a single difficulty parameter but multiple discrimination parameters of an item. For example, two items presented in Figure 5a and Figure 5b show the effect of discrimination power on IRS. They measure only the first dimension but the item in Figure 5a has more discriminating power than the item in Figure 5b giving a steeper IRS. Similarly, the items displayed in Figure 5c and Figure 5d primarily measure the second trait only. However, the item in Figure 5c discriminates the second trait better than the item in Figure 5d. Figure 5e and Figure 5f are showing two multidimensional items that are equally discriminating two latent traits well. The difference between them is that the item in Figure 5e requires higher proficiency level on both traits than the item in Figure 5f.

However, the interpretation of the IRS is not as easy as it looks at the first glance. It is notable that the IRS of an MIRT model gives numerous slopes depending on different combinations of latent traits. Ackerman (1996) showed graphically how the information of an item varies depending on the different combinations of multiple latent traits. The item difficulty and item discrimination of a multidimensional item depend on a particular composite of latent traits. Reckase (1985, 1991) suggested how to determine multidimensional item difficulty (MDIFF) and multidimensional discrimination (MDISC)

of a multidimensional item hoping that characterizing multidimensional items with a single number is helpful in ranking and selecting them. Those formulae are shown below.

$$MDISC_i = \sqrt{\sum_{k=1}^{m} \alpha_{ik}} \qquad (8)$$

$$MDIFF_i = \frac{-d_i}{MDISC_i} \qquad (9)$$

$a_{ik} = \arccos \frac{\alpha_{ik}}{MDISC}$, where

$MDISC_i$ multidimensional item discrimination of item i

$MDIFF_i$ multidimensional item difficulty of item i

α_{ik} slope of item i on k^{th} dimension

d_i multidimensional difficulty of item i

a_{ik} angle between a line made by MDIFF and k^{th} axis for i^{th} item

MDIFF is the distance from the origin to the point of the steepest slope in θ space in the direction that gives the best discriminating power. MDISC is a sensitivity of a multidimensional item to the difference in a particular composite of traits in the direction that gives the best discriminatory power. Both MDIFF and MDISC are convenient to summarize multidimensional items and may be used as the counterparts of unidimensional IRT. However, the item information function (MIF), an additional important concept in MIRT, depends on the direction that is being measured in θ space, which is determined by the formula below. In summary, the most informative point of an

multidimensional item can be uniquely determined with the information of MDIFF, MDISC and the angle that is being measured in θ space.

$$I_{ia}(\theta) = \frac{\left[\nabla_a P_i(\theta)\right]^2}{P_i(\theta)[1-P_i(\theta)]} = D^2(a_i`u_i)^2 P_i(\theta)Q_i(\theta)$$, where

a a vector of angles determined by

$I_{ia}(\theta)$ the information of item i in the direction of a in M2PLM

u_i a vector of directional cosine

(Reckase & McKinley, 1991; Reckase, 1997)

Computer Programs for IRT Models

Many computer programs conducting IRT parameter estimation have been developed to meet the increasing computational demand. Popularly used computer programs based on unidimensional IRT are LOGIST (Wingersky, 1983), BILOG (Mislevy & Bock, 1983), PARSCALE (Muraki & Bock, 1996) and MULTILOG (Thissen, 1991). They can be categorized depending on the response type of the data and the estimation method implemented. LOGIST and BILOG are both used to analyze dichotomously responded data, but they are different in that the joint maximum likelihood estimation method is implemented in LOGIST whereas the marginal maximum likelihood estimation method is implemented in BILOG. PARSCALE is used to estimate parameters when the response categories are graded as in Likert-type scale. MULTILOG is used when the responses of examinees are nominally scored.

Since the advantages of a multidimensional approach to measuring underlying latent traits over a unidimensional approach were recognized in the testing arena, the demand for computer software that can implement multidimensional test theory has increased. Commonly used computer programs equipped with a multidimensional perspective are MIRTE (Carlson, 1987), TESTFACT (Wilson, Wood, & Gibbons, 1984) and NOHARM (Fraser, 1988). Those programs are flexible in that they can be used to estimate parameters based on both unidimensional IRT and multidimensional IRT but they have been developed from distinctive principles. In MIRTE, which is an updated version of MAXLOG (McKinley & Reckase, 1983), the joint maximum likelihood estimation method suggested by Birnbaum (1968) is expanded to the multidimensional perspective. Full information item factor analysis using the EM algorithm is implemented in TESTFACT. In NOHARM the normal ogive ICC is approximated by a third degree Hermite-Tchebycheff polynomial with least squares criterion.

Development of SAS IRT Programs

Research Topics

The primary purpose of the current dissertation is to develop reliable and precise SAS computer programs that can be used to calibrate item parameters and estimate the proficiency level of examinees. Two computer programs that conduct parameter estimation based on unidimensional IRT are developed to accommodate various response types. First, a computer program implementing the marginal maximum likelihood estimation method suggested by Bock (1981) is developed to deal with dichotomous data.

Second, another computer program based on the polytomous IRT model by Samijima (1969) is developed to analyze categorical and ordered response data.

The unidimensional estimation program is extended to the multidimensional perspective, which is needed to reveal a more complex and more realistic structure of latent traits. The last computer program, based on full information factor analysis (Bock, Gibbons & Muraki, 1988), is developed to estimate parameters of multidimensional items. Even though the multidimensional normal ogive model was suggested to describe the interaction between multiple latent traits of an examinee and a multidimensional item, the multidimensional logistic ogive model was implemented in the SAS MDIRT macro because of the computational efficiency and numerical equivalency ($|\phi(z) - \psi(z)| < 0.01$ for all z).

Following, the computer programs developed in this study will be simulated with computer-generated data to examine the stability, precision and comparability to their competitors (e.g. BILOG, PARSCALE, and TESTFACT). In addition, newly developed SAS macros are demonstrated and validated with an empirical application. The math items from the Peabody Individual Achievement Test (PIAT-Math) will be used to validate and demonstrate the utility of the SAS IRT programs.

Computer Language used in the Study

Albeit item response theory has come to fruition as computer technology has advanced, there are still some obstacles to make it easily accessible. In the market, there are several computer programs dedicated to IRT, but accessibility to them is still costly

and requires quite some time in getting used to the program language. The computer programs developed in this research are written in SAS, which is one of the most popular and familiar statistical packages and data management to statistics-related people. This will hopefully increase availability and accessibility to IRT estimation software. The SAS interactive matrix language (IML) is used because it is a powerful and flexible programming tool, which makes it easy to use for matrix mathematics.

The Estimation Method in SAS Unidimensional IRT Programs

In this dissertation, the marginal maximum likelihood estimation (MMLE) method for the two parameter logistic model and the Bayesian marginal maximum likelihood estimation (BMMLE) method for the 3PLM are implemented. They not only overcome some limitations associated with other estimation methods, but also facilitate the calibrating process by applying the EM algorithm (Bock, 1981). It is well known that estimated structural parameters are not necessarily consistent when structural parameters are jointly estimated with incidental parameters (Neyman & Scott, 1948). In IRT, item parameters are considered as structural while parameters of the underlying latent trait are considered as incidental, because the number of ability parameters estimated increases as the number of respondents increases. For the Rasch model in which all items are assumed to have the same discrimination power, the conditional maximum likelihood estimation method can be an alternative to MMLE because it provides a sufficient statistic (e.g. the sum of correct responses) for latent trait level at each level of the latent trait. However, this does not hold for the two or three parameter model.

In MMLE, assuming the population distribution of a latent trait is known, item parameters are estimated independently of incidental parameters by integrating over the distribution of the latent trait. Thus, MMLE is freed from the inconsistency problem, which is not true for other estimation techniques (Baker, 1992). However, MMLE does not provide an appropriate estimate for an item to which all examinees respond correctly or an item to which all respondents answer incorrectly. In addition, MMLE is not capable of calibrating the proficiency level of an examinee when she or he makes all correct responses or all incorrect responses to the test items, unless some constraint is imposed. In contrast, BMMLE assumes a distribution of parameters in the population, which prevents item or proficiency parameters from drifting toward positive or negative infinity.

Three Steps of Estimation common to all SAS Unidimensional IRT Programs

The process of estimating parameters consists of three stages. In the first stage, initial values, which are used as starting values in the next stage, are computed. Obtaining initial values close to true parameters is important because it reduces the number of iterations required in estimating parameters. The importance of good starting values becomes more evident when the number of estimated parameters for an item increases. Baker (1988) showed that whether the Newton-Raphson iteration process converges depends on initial estimates being very close to the actual value when the guessing parameter is added.

In the second stage, the item parameters are estimated by implementing the estimation method of marginal maximum likelihood with the EM algorithm. The EM

algorithm (Damster, Laird, and Rubin, 1977) is divided into two processes, which are the expectation process and the maximization process. The expected posterior probability of correct responses to an item conditional on each trait level is computed in the expectation process and the item parameters are estimated at the value that maximizes the likelihood function based on the expected posterior probability in the maximization process. Those two steps are iterated until the convergence of estimated parameters is achieved by using the Newton-Raphson iteration method. It is notable that the process of estimating the item parameters does not depend on the process of estimating the trait parameters of examinees because of the EM algorithm. The EM algorithm will be described in detail later.

In the last stage, the examinee parameters are estimated by using the item parameters obtained in the second stage. Unlike the item parameters, examinee parameters are estimated by implementing the maximum likelihood estimation method without the E-M step.

Estimation Method Implemented in SAS Multidimensional IRT Program

In this MDIRT-FIT macro, the full information item factor analysis model (Bock, Gibbons & Muraki, 1988) with marginal maximum likelihood estimation with the EM algorithm is implemented to estimate parameters. Even though Takane (1987) demonstrated the equivalence of marginal likelihood of the two-parameter normal ogive model and factor analysis on dichotomous data, factor analysis on dichotomized variables has produced unsatisfactory results. It is difficult in some cases to compute desirable

tetrachoric correlation coefficients and the computational burden increases as the number of items increases. The full information item factor analysis model gives a better solution to those problems by directly modeling item responses of examinees instead of modeling the pairwise tetrachoric correlation coefficients.

STAGE 1: Computing Initial Value

Computing Tetrachoric Correlation. In the 1940's, the phi correlation coefficient, conventionally used as a linear description of two binary variables, it was shown to be an inappropriate measure to represent the true relationship between two dichotomous variables with underlying normal distributions. Researchers found several problems with the phi correlation coefficient when it was used for continuous variables. In particular, it tends to decrease when the difficulty levels of two variables are different. In other words, the phi correlation coefficient no longer has an upper bound of unity when the difficulty of two variables is different, which results in misrepresenting the true relationship between two continuous variables (McDonald, 1985, pp.198). Due to this drawback, it has been suggested that the tetrachoric correlation coefficient which is a Pearson product-moment correlation for binary variables with an underlying normal distribution be substituted for the phi correlation.

		Item 1		
		Right	Wrong	
Item 2	Right	P_{11} (a)	P_{12} (b)	$P_{1\bullet}$
	Wrong	P_{21} (c)	P_{22} (d)	$P_{2\bullet}$
		$P_{\bullet 2}$	$P_{\bullet 1}$	

Table 1. 2×2 Contingency Table. $*$( ) frequency for each cell

A good estimation method for the tetrachoric correlation was suggested by Brown (1977). In this article, the integral of the bivariate normal distribution is approximated by the tetrachoric series expansion when the estimated $|r|$ is smaller than 0.95. Otherwise, 32 Gaussian quadrature points are used to evaluate the tail of the bivariate normal distribution. The estimated tetrachoric correlation is accurate to at least three decimal places unless one of the cell probabilities is less than 0.0001. The standard deviation of the estimated tetrachoric correlation is provided as an option in this SAS program.

The Newton-Raphson method is implemented to estimate the population tetrachoric correlation. Yule's Y (equation 10), one of the indices representing the magnitude of association between two dichotomous variables, is used as the initial estimate of r. The Tetrachoric series (equation 11) and its derivative with respect to r (equation 12) are computed to find the next approximation at each iterative step. Each new estimate of r is obtained from the previous one by the Newton-Raphson method (equation 13). The following equations support this development:

$$\text{Yule's Y} = \frac{\left[\sqrt{ad} - \sqrt{bc}\right]^2}{ad - bc} \tag{10}$$

$$L(z_1, z_2, r) = \int_{-\infty}^{z_2}\int_{-\infty}^{z_1} \phi(x_1, x_{2,} r)dx_1 dx_2 = P_{1\bullet}P_{\bullet 1} + \sum_{n=1}^{\infty} \frac{r^n}{n!}\phi(z_1, z_2, 0)v_{n-1}w_{n-1}, \text{ where} \tag{11}$$

$v_0 = 1$ $\qquad$ $w_0 = 1$

$v_1 = z_1$ $\qquad$ $w_1 = z_2$

. $\qquad$.

. $\qquad$.

$v_n = z_1 v_{n-1} - (n-1)v_{n-2}$ $\qquad$ $w_n = z_2 w_{n-1} - (n-1)w_{n-2}$

$$\frac{\partial L}{\partial r} = \sum_{n=1}^{\infty} \frac{r^n}{(n-1)!}\phi(z_1, z_2, 0)v_{n-1}w_{n-1} \tag{12}$$

$$r_{t+1} = r_t + \frac{L(z_1, z_2, r)}{\frac{\partial L}{\partial r}} \tag{13}$$

(Brown, 1977)

However, the tetrachoric correlation has its own shortcomings when used for factor analysis on dichotomous data. It is not assured that the observed tetrachoric correlation matrix is positive definite, if not, then, it is not a complete correlation matrix but an imaginary correlation matrix. In addition, it was found that factor analysis on the matrix of tetrachoric correlations tends to extract more factors than are actually present in the data. Therefore, a smoothing procedure must be implemented before the tetrachoric correlation matrix is used for factor analysis when it is not positive definite.

Smoothing Tetrachoric Correlation. If the matrix of the observed tetrachoric correlation (see equation 15) is not positive definite, (which means one of the eigenvalues is negative), a nonnegative definite correlation (positive semi-definite) matrix can be obtained by correcting the negative eigenvalues without changing the sum of the eigenvalues. For an example, the reader should see the manual of TESTFACT (1984, pp. 790).

$$R^S = [Diag(KD^C K')]^{-1/2} KD^C K'[Diag(KD^C K']^{-1/2}, \text{ where} \quad (14)$$

R^S is a smoothed tetrachoric correlation matrix

D^C is a corrected diagonal matrix with corrected eigenvalues

K is a diagonal matrix with eigenvalues of R

K' is a transposed matrix of K

(Knol & Berger, 1991)

Factor Analysis for Initial Estimates of Item Parameters. To obtain initial values close to true item parameters, a factor analysis on smoothed tetrachoric correlation matrix needs to be conducted. To extract factor loadings from the smoothed correlation matrix, the minimum residual (MINRES) method is implemented. MINRES factor analysis (Harman, 1976) determines a new factor pattern matrix so that the objective function (equation 16) is minimized (using the least-squares criterion) with only off-diagonal elements of the correlation matrix accounted for. Then, the obtained factor

loadings are rotated to simple structure (varimax rotation), using an option in SAS MDIRT program. The equations to support this process are the following:

$$R = \begin{bmatrix} r_{11} & r_{12} & . & r_{1n} \\ r_{21} & r_{22} & . & r_{2n} \\ . & . & . & . \\ r_{n1} & _{n2} & & r_{nn} \end{bmatrix} \tag{15}$$

$$F(A) = \sum_{j=1}^{n} \sum_{k=1}^{n} (r_{jk} - \hat{r}_{jk}), \quad (j \neq k), \text{ where} \tag{16}$$

R an observed correlation matrix

$F(A)$ is objective function that is minimized

r_{jk} is an observed correlation between variable j and k

$\hat{r}_{jk}$ is an estimated correlation between variable j and k

MINRES proceeds by estimating initial communalities for each item using squared multiple correlation (17), replacing them with the diagonal elements of the smoothed correlation matrix, extracting initial factor loadings by eigen decomposition (equation 18) given that the number of factors is determined in advance, estimating parameters (factor loadings) that minimize the objective function (19), and storing the updated factor loadings (20) repeating previous steps (19, 20) until a desired criterion is reached.

Squared Multiple Correlation (SMC) : $1-\frac{1}{r_{jj}}$, where (17)

r_{jj} is the diagonal element in R^{-1} corresponding to variable j

Eigen Decompositon : $R = \Lambda`\Lambda$, $\Lambda = CD^{1/2}$, where (18)

Λ is the factor pattern matrix

C is the matrix comprising of eigen vectors

D is the diagonal matrix of eigenvalues

Incremental change of factor loading : $\Lambda(\Delta) = R_j^0 \Lambda(\Lambda`_j \Lambda_j)^{-1}$, where (19)

R_j^0 is the row vector of residual correlation of variable j with all other variables (zero for the self-residual)

Λ_j is the factor matrix with the elements in row j replaced by zeros

$\Lambda(\Delta)$ is the incremental change of factor loadings at each iterative step

$\Lambda_{t+1} = \Lambda_t + \Lambda(\Delta)$, where (20)

Λ_{t+1} is the factor loading at the $(t+1)^{th}$ iterative step

Interchangeability Between FA Parameters and MIRT Parameters. The full information item factor analysis model (Bock, Gibbons & Muraki, 1988) applied Thurston's multiple factor model to describe the underlying response process. The response process (called an item variable by Baker, 1992, pp.8) is a hypothetical continuous random variable over a population of subjects representing a subject's

propensity to respond correctly to an item. The factor analysis model is:

$$y_i = f_1\theta_1 + f_2\theta_2 .... + f_m\theta_m + \varepsilon \qquad (21)$$

with $y_i \sim \mathrm{N}(0,1)$, $\theta_1 \sim \mathrm{N}(0,1)$, …… $\theta_m \sim \mathrm{N}(0,1)$,

assumed for the response of subject *j* to an item *i*

$x_{ij} = 1$ if $y_i = f_{1j}\theta_{1j} + f_{2j}\theta_{2j} .... + f_{mj}\theta_{mj} + \varepsilon > \gamma_i$

$x_{ij} = 0$ if $y_i = f_{1j}\theta_{1j} + f_{2j}\theta_{2j} .... + f_{mj}\theta_{mj} + \varepsilon < \gamma_i$, where

γ_i is item threshold, then

$$P(x_{ij} = 1 \mid \theta_1, \theta_2, ... \theta_m) = \frac{1}{\sqrt{2\pi}\sigma} \int_{Z_i(\theta_j)}^{\infty} e^{-\frac{1}{2}z^2} dz = \phi(Z), \text{ where}$$

$$Z_i(\theta_j) = \frac{f_{1j}\theta_{1j} + f_{2j}\theta_{2j} ... + f_{mj}\theta_{mj} - \gamma_i}{\sigma_i} \quad \text{and}$$

$$\sigma_i = \sqrt{1 - (f_{1j}^2 + f_{2j}^2 ... f_{mj}^2)}$$

The corresponding multidimensional logistic model is

$$P(x_{ij} = 1 \mid \theta_1, \theta_2, ... \theta_m) = \frac{1}{1 + e^{-DZ_i(\theta_j)}} = \psi(Z), \text{ where}$$

$$Z_i(\theta_j) = \alpha_{1j}\theta_{1j} + \alpha_{2j}\theta_{2j} .... + \alpha_{mj}\theta_{mj} + d_i$$

Therefore, the parameters of factor analysis can be easily interpreted as the parameters of MIRT and vice versa.

$$\sigma_i = \sqrt{1 - \sum_{k=1}^{m} f_{ik}^2} \quad , (22) \qquad \alpha_{ik} = \frac{f_{ik}}{\sigma_i} \quad , (23) \qquad d_i = -\frac{\gamma_i}{\sigma_i} \quad , (24)$$

Where,

f_{ik} is the factor loading of item *i* on the k[th] dimension

α_{ik} is discrimination parameter of item *i* on the k[th] dimension

$z_i(\theta_j)$ is the normal deviate of item *i* conditional on θ_j

σ_i is the standard error of unique factor in FA

d_i is the intercept of item *i*

Gaussian-Hermite Quadrature Method. Numerical integration of the normal distribution is a challenging process in terms of computer running time. The most popular solution to the problem is to evaluate the continuous normal distribution by a discrete distribution on a small number of points as shown in Figure 6. For example, we can approximate the bivariate normal distribution by summing the volume of hexahedrons under the surface of bivariate normal distribution. The center of a hexahedron and its volume are called a joint quadrature point and weight, respectively. Quadrature points are the roots of Hermite polynomials over the interval $[-\infty, \infty]$ with its weighting function of e^{-x^2}. Computed quadratures are multiplied by $\sqrt{2}$ and weights are divided by $\sqrt{\pi}$ to obtain better approximation.

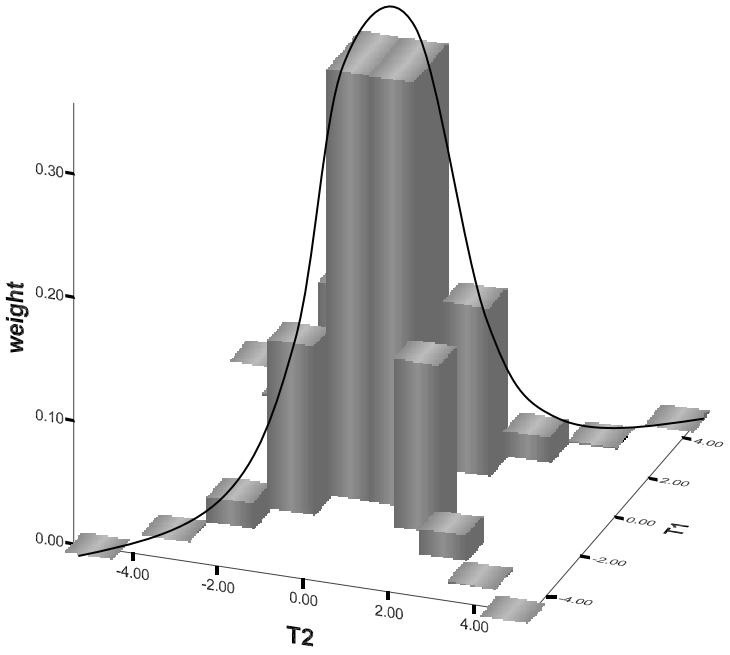

Figure 6. approximation of bivariate PDF by hexahedrons

STAGE 2: Estimating Item Parameters with MML

Likelihood Method with EM Algorithm. It was pointed out by Neyman and Scott (1948) that the inconsistent estimates are caused by the dependence of the estimation process of item parameters on the estimation process of ability parameters. Bock & Liberman (1970) suggested the marginal maximum likelihood estimation (MMLE) to circumvent the inconsistency issue on the estimates of item parameters. They

estimated item parameters in the marginal distribution to preclude the effect of ability (nuisance variable) parameters from the estimation process of item parameters by integrating over the ability distribution.

$$P(\underline{U}_j) = \int_{-\infty}^{\infty} P(\underline{U}_j \mid \theta_j, \alpha_i, d_i) g(\theta \mid \tau) d\theta \text{, where} \qquad (25)$$

$P(\underline{U}_j)$ is the marginal probability that an examinee with ability θ_j has a response vector $\underline{U}$ with respect to item parameters and the population ability density function.

α_i is a vector of discrimination parameters of item *i*

d_i is item difficulty parameter of item *i*

θ_j is a vector of ability parameters of examinee *j*

$g(\theta \mid \tau)$ is a probability density function of ability

Computing the marginal probability $P(\underline{U}_j)$ demands a great deal of computational work because it includes numerical integration of the multivariate normal distribution which is assumed for the underlying latent traits. The Gaussian quadrature method is used to compute the marginal probability of a response pattern of an examinee for a set of test items. A numerical approximation to the integration over the ability distribution is conducted by summing the product of the likelihood at all joint quadarature points and their associated weights.

$$P(\underline{U}_j)=\sum_{k1=1}^{p}\sum_{k2=1}^{q}\cdots\sum_{km=1}^{t}\prod_{i=1}^{n} P_i(X_{k1k2\ldots km})^{U_{ij}} Q(X_{k1k2\ldots km})^{1-U_{ij}} A(X_{k1})A(X_{k2})\ldots A(X_{km}) \qquad (26)$$

$$P(\underline{U}_j)= \sum_{k1=1}^{p}\sum_{k2=1}^{q}\cdots\sum_{km=1}^{t}\prod_{i=1}^{n} L(X_{k1k2\ldots km})A(X_{k1})A(X_{k2})\ldots A(X_{km}), \qquad (27)$$

where

$P(\underline{U}_j)$ is the marginal probability that an examinee with θ_j has a response vector $\underline{U}$ with respect to the item parameters and the population ability density function.

$P(X_{k1k2\ldots km})^{U_{ij}}$ is the probability of a correct response at a joint quadrature point (k1= 1, 2… p), (k2 = 1, 2… q) …., and (km = 1, 2 … t)

$Q(X_{k1k2\ldots km})^{1-U_{ij}}$ is the probability of an incorrect response at a joint quadrature point (k1= 1, 2… p), (k2 = 1, 2… q) …., and (km = 1, 2 … t)

$L(X_{k1,k2\ldots km})$ is the estimated likelihood at a joint quadrature point (k1= 1, 2… p), (k2 = 1, 2… q) …., and (km = 1, 2 … t)

$A(X_{km})$ is the weight at k^{th} quadrature point on m^{th} dimension

The EM algorithm is a useful estimation technique to find the maximum likelihood estimates of parameters in a probabilistic model that depends on an unobservable latent variable. Bock & Aitkin (1981) proposed the EM algorithm be implemented because the parameter estimation process (MMLE) of IRT models, a set of latent trait models, fit the paradigm. In addition, employing the EM algorithm with MMLE adds more velocity to the estimation process.

Expectation Process. The expected number of examinees ($\bar{f}_{ik1k2...km}$) and the expected number of correct responses ($\bar{r}_{ik1k2...km}$) at all pairs of quadrature points ($X_{ik1k2...km}$) corresponding to *m* latent traits are computed in the expectation step of the EM algorithm. The provisional statistics are called "artificial data" because they are computed based on the assumed population distribution of latent trait (Baker, 1992). In other words, the expected likelihood including the unobservable latent trait is computed as though it were observed.

$$\bar{f}_{ik1,k2..km} = \sum_{j}^{N} \left[\frac{L(X_{k1}, X_{k2}, ...X_{km}) A(X_{k1}) A(X_{k2}) ... A(X_{km})}{\sum_{k1}^{p} \sum_{k2}^{q} ... \sum_{km}^{t} L(X_{k1}, X_{k2}, ...X_{km}) A(X_{k1}) A(X_{k2}) ... A(X_{km})} \right] \tag{28}$$

$$\bar{r}_{ik1,k2..km} = \sum_{j}^{N} \left[\frac{u_{ij} L(X_{k1}, X_{k2}, ...X_{km}) A(X_{k1}) A(X_{k2}) ... A(X_{km})}{\sum_{k1}^{p} \sum_{k2}^{q} ... \sum_{km}^{t} L(X_{k1}, X_{k2}, ...X_{km}) A(X_{k1}) A(X_{k2}) ... A(X_{km})} \right] \tag{29}$$

,

where

$\bar{f}_{ik1,k2...km}$ is the expected number of examinees at the joint quadrature point (k1 = 1, 2… p), (k2 = 1, 2… q) …., and (km = 1, 2 … t)

$\bar{r}_{ik1,k2...km}$ is the expected number of correct responses at the joint quadrature point (k1 = 1, 2… p), (k2 = 1, 2… q) …., and (km = 1, 2 … t)

$L(X_{k1,k2...km})$ is the estimated likelihood at each of joint quadrature points (k1= 1, 2…

p), (k2 = 1, 2… q) …., and (km = 1, 2 … t)

$A(X_{km})$ is the weight at the k[th] quadrature point on the m[th] dimension

u_{ij} is a response of examinee *j* to item *i* (1 for a correct response, 0 otherwise)

Maximization Process. The estimates of item parameters are computed using the marginal distribution of the item response vector ***U*** with respect to the item parameters. Item parameter estimates are determined where the marginal likelihood function is maximized. To determine provisional estimates of item parameter, the first, the second, and the cross derivatives of the log likelihood function with respect to each item parameter are expressed in terms of those artificial statistics. The result of the differentiation of the log likelihood function with respect to each of the item parameters is shown below in general form. To find more detail about the differential process of the log likelihood function, see Baker (1992, Ch. 6).

$$L = \sum_{j=1}^{N} P(\underline{U}) \qquad (30)$$

$$\log L = \log \sum_{j=1}^{N} P(\underline{U}) \qquad (31)$$

The first derivative is computed from the following:

$$\frac{\partial \ln L}{\partial \alpha_{im}} = \sum_{k1=1}^{p} \sum_{k2}^{q} .... \sum_{km=1}^{t} D(X_{km})[(\bar{r}_{ik1k2...km} - \bar{f}_{ik1k2...km} P_i(X_{k1k2...km})] \qquad (32)$$

$$\frac{\partial \ln L}{\partial d_i} = \sum_{k1=1}^{p}\sum_{k2}^{q}....\sum_{km=1}^{t} D[(\bar{r}_{ik1k2...km} - \bar{f}_{ik1k2...km} P_i(X_{k1k2...km})] \quad (33)$$

The first derivative is computed from the following:

$$\frac{\partial^2 \ln L}{\partial^2 \alpha_{im}} = -D^2 \sum_{k1=1}^{p}\sum_{k2=1}^{q}...\sum_{km=1}^{t} (X_{km}^2)\bar{f}_{ik1k2...km} P_i(X_{k1k2...km}) Q_i(X_{k1kx...km}) \quad (34)$$

$$\frac{\partial^2 \ln L}{\partial^2 d_i} = -D^2 \sum_{k1=1}^{p}\sum_{k2=1}^{q}...\sum_{km=1}^{t} \bar{f}_{ik1k2...km} P_i(X_{k1k2...km}) Q_i(X_{k1kx...km}) \quad (35)$$

The cross derivative is computed from the following:

$$\frac{\partial^2 \ln L}{\partial \alpha_{im}\alpha_{in}} = -D^2 \sum_{k1=1}^{p}\sum_{k2=1}^{q}...\sum_{km=1}^{t} (X_{km})(X_{kn})\bar{f}_{ik1k2...km} P_i(X_{k1k2...km}) Q_i(X_{k1kx...km}), \quad m \neq n \quad (36)$$

$$\frac{\partial^2 \ln L}{\partial \alpha_{im} d_i} = -D^2 \sum_{k1=1}^{p}\sum_{k2=1}^{q}...\sum_{km=1}^{t} (X_{km})\bar{f}_{ik1k2...km} P_i(X_{k1k2...km}) Q_i(X_{k1k2...km}) \quad (37)$$

Newton-Raphson Iteration Method. The expectation process and the maximization process are looped inside the Newton-Raphson method. The iteration process continues until the convergence criterion (e.g. 0.001) is satisfied. The approximation at the t^{th} cycle is updated by computing the change from the previous approximation and adding it to the approximation at the $(t-1)^{th}$ cycle. For example, the inversed information matrix consisting of the second derivatives on its diagonal and cross derivatives and the vector consisting of the first derivatives with respect to parameters of item *i* are multiplied to compute the amount of change from the previous iteration.

$$\begin{bmatrix} \hat{\alpha}_{i1} \\ \hat{\alpha}_{i2} \\ . \\ . \\ \hat{\alpha}_{im} \\ \hat{d}_i \end{bmatrix} = \begin{bmatrix} \hat{\alpha}_{i1} \\ \hat{\alpha}_{i2} \\ . \\ . \\ \hat{\alpha}_{im} \\ \hat{d}_i \end{bmatrix} + \begin{bmatrix} \frac{\partial^2 \ln^L}{\partial^2 \alpha_{i1}} & \frac{\partial^2 \ln^L}{\partial \alpha_{i1} \partial \alpha_{i2}} & . & . & \frac{\partial^2 \ln^L}{\partial \alpha_{i1} \partial \alpha_{im}} & \frac{\partial^2 \ln^L}{\partial \alpha_{i1} \partial d_i} \\ \frac{\partial^2 \ln^L}{\partial \alpha_{i2} \partial \alpha_{i1}} & \frac{\partial^2 \ln^L}{\partial^2 \alpha_{i2}} & . & . & \frac{\partial^2 \ln^L}{\partial \alpha_{i2} \partial \alpha_{im}} & \frac{\partial^2 \ln^L}{\partial \alpha_{i2} \partial d_i} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \frac{\partial^2 \ln^L}{\partial \alpha_{im} \partial \alpha_{i1}} & \frac{\partial^2 \ln^L}{\partial \alpha_{im} \partial \alpha_{i2}} & . & . & \frac{\partial^2 \ln^L}{\partial^2 \alpha_{im}} & \frac{\partial^2 \ln^L}{\partial^2 \alpha_{im} d_i} \\ \frac{\partial^2 \ln^L}{\partial d_i \partial \alpha_{i1}} & \frac{\partial^2 \ln^L}{\partial d_i \partial \alpha_{i2}} & . & . & \frac{\partial^2 \ln^L}{\partial d_i \partial \alpha_{im}} & \frac{\partial^2 \ln^L}{\partial^2 d_i} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \ln^L}{\partial \alpha_{i1}} \\ \frac{\partial \ln^L}{\partial \alpha_{i2}} \\ . \\ . \\ \frac{\partial \ln^L}{\partial \alpha_{im}} \\ \frac{\partial \ln^L}{\partial d_i} \end{bmatrix} \quad (38)$$

STAGE 3: Estimating Ability Parameters

Estimation Methods. At the third stage, the Bayes expected a posteriori (EAP) estimation procedure is implemented to estimate ability parameters assuming the obtained item parameters at the second stage are true item parameters. The process of computing the EAP estimates of a subject's ability is non-iterative and can be easily obtained. For instance, the EAP ability estimate of the j^{th} examinee on the first dimension (equation 39) and the posterior standard deviation (PSD) of the estimate can be directly approximated by equation 40. The maximum likelihood estimation (ML) procedure is available as well. The first (equation 41), second (equation 42), and cross derivatives (equation 43) of the log likelihood function with respect to multiple ability parameters are needed because Fisher's scoring method (equation 44) is implemented. The estimated parameters are updated until they reach a predetermined criterion value. The information matrix provides the variances of the estimated ability parameters and the inverse

information matrix provides measurement errors of those ability parameters.

$$E(\theta_{j1} \mid \underline{U}, \xi) = \overline{\theta}_{j1} = \frac{\sum_{k1}^{p} X_{k1} \left[\sum_{k2}^{q} \cdots \sum_{km}^{t} L(X_{k1,k2\ldots km}) A(X_{k2}) \ldots A(X_{km-1}) \right] A(X_{k1})}{\sum_{k1}^{p} \sum_{k2}^{q} \cdots \sum_{km}^{t} L(X_{k1,k2\ldots km}) A(X_{k1}) A(X_{k2}) \ldots A(X_{km})} \tag{39}$$

$$\text{PSD of } \overline{\theta}_{j1} = \sqrt{\frac{\sum_{k1}^{p} (X_{k1} - \overline{\theta}_{j1})^2 \left[\sum_{k2}^{q} \cdots \sum_{km}^{t} L(X_{k1,k2\ldots km}) A(X_{k2}) \ldots A(X_{km-1}) \right] A(X_{k1})}{\sum_{k1}^{p} \sum_{k2}^{q} \cdots \sum_{km}^{t} L(X_{k1,k2\ldots km}) A(X_{k1}) A(X_{k2}) \ldots A(X_{km})}} \tag{40}$$

$$\frac{\partial L}{\partial \theta_{jm}} = D \sum_{i=1}^{n} \alpha_m W_{ij} \left(\frac{u_{ij} - P_{ij}}{P_{ij} Q_{ij}} \right) \tag{41}$$

$$\frac{\partial^2 L}{\partial^2 \theta_m} = -D^2 \sum_{i=1}^{n} \alpha_m^2 W_{ij} \tag{42}$$

$$\frac{\partial^2 L}{\partial \theta_m \partial \theta_n} = -D^2 \sum_{i=1}^{n} \alpha_m \alpha_n W_{ij} \tag{43}$$

$$\theta^{(k+1)} = \theta^{(k)} - I_n^{-1}(\theta) \frac{\partial \ln L(\theta^{(k)})}{\partial \theta} \tag{44}$$

$$\begin{bmatrix} \hat{\theta}_{j1} \\ \hat{\theta}_{j2} \\ \cdot \\ \hat{\theta}_{jm} \end{bmatrix} = \begin{bmatrix} \hat{\theta}_{j1} \\ \hat{\theta}_{j2} \\ \cdot \\ \hat{\theta}_{jm} \end{bmatrix} + \begin{bmatrix} \frac{\partial^2 L}{\partial^2 \theta_{j1}} & \frac{\partial^2 L}{\partial \theta_{j1} \partial \theta_{j2}} & \cdot & \frac{\partial^2 L}{\partial \theta_{j1} \partial \theta_{jm}} \\ \frac{\partial^2 L}{\partial \theta_{j2} \partial \theta_{j1}} & \frac{\partial^2 L}{\partial^2 \theta_{j2}} & \cdot & \frac{\partial^2 L}{\partial \theta_{j2} \partial \theta_{jm}} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial^2 L}{\partial \theta_{jm} \partial \theta_{j1}} & \frac{\partial^2 L}{\partial \theta_{jm} \partial \theta_{j2}} & & \frac{\partial^2 L}{\partial^2 \theta_{jm}} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial L}{\partial \theta_{j1}} \\ \frac{\partial L}{\partial \theta_{j2}} \\ \cdot \\ \frac{\partial L}{\partial \theta_{jm}} \end{bmatrix} \tag{45}$$

Simulation of Parameter Recovery

Method

A Monte Carlo method was used to examine how close parameters recovered by the SAS MDIRT program are to the true parameters. Ability parameters of the j^{th} examinee up to five dimensions ($\theta_{j1}, \theta_{j2}, ...\theta_{j5}$) are generated from an NID (0,1) distribution and these values were used as examinees' true known ability levels. Discrimination and intercept parameters were generated from a uniform distribution. Discrimination parameters of the i^{th} item on up to five dimensions ($\alpha_{i1}, \alpha_{i2} ....\alpha_{i5}$) were generated from a uniform distribution [0, 2]. The intercept parameter of the i^{th} item is generated from a uniform distribution [-3, 3]. To simulate the response vector of an examinee, a randomly generated number from a uniform distribution [0, 1] is compared to the probability obtained by plugging the randomly generated item parameters and ability parameters into equation (2). If the computed probability is bigger than the random number, 1 is assigned to the response, or 0 assigned otherwise.

Factors manipulated to examine the recovery ability of the program are the number of latent dimensions, the number of examinees and the number of quadrature points. For two dimensions, an additional factor, the extent of correlation between two ability dimensions, is manipulated to examine the effect of correlated dimensions on parameter recovery using the SAS MDIRT program. To evaluate the accuracy of the recovered multidimensional item characteristic curve, the root mean square deviation (RMSE) between the estimated probability and the true probability was calculated using the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{K}\sum_{j=1}^{L}(P-\hat{P})^2}{KL}}\text{, where} \qquad (46)$$

K = the number of items

L = the number of examinees.

To obtain a reliable estimate of RMSE, each condition was repeated 10 times, which was constant for all conditions.

Results

The mean and the standard deviation (in the parenthesis) of 10 estimates with the number of quadrature points fixed at 10 are shown in Table 2. The third column of Table 2 is the correlation between two dimensions, which was manipulated from 0.0 to 0.9 to examine the effect of correlated dimensions on the accuracy of the estimated item parameters. The next three columns are the correlation between the true item parameters and the recovered item parameters by SAS MDIRT program. RMSE shown in the last column was computed as an index for the agreement between the estimated probability and the true probability of a correct response.

Table 2. The means and standard deviations of RMSE between the estimated ($\hat{P}$) and true probability (P) across items and the means and standard deviations of the correlation between the true parameters and the estimated parameters from SAS MDIRT program using 10 quadrature points for two dimensions.

N	n	$r(\theta_1,\theta_2)$	$r(\alpha_1,\hat{\alpha}_1)$	$r(\alpha_2,\hat{\alpha}_2)$	$r(d,\hat{d})$	RMSE
2000	40	0.0042(0.0253)	0.989(0.004)	0.989(0.005)	0.998(0.000)	0.037(0.012)
2000	40	0.3015(0.0226)	0.974(0.019)	0.966(0.020)	0.998(0.001)	0.067(0.014)
2000	40	0.5005((0.0140)	0.941(0.042)	0.956(0.041)	0.998(0.001)	0.068(0.016)
2000	40	0.7035(0.0043)	0.891(0.058)	0.925(0.053)	0.998(0.000)	0.079(0.009)
2000	40	0.9098(0.0275)	0.797(0.120)	0.749(0.119)	0.984(0.042)	0.075(0.003)
2000	20	0.0125(0.025)	0.991(0.004)	0.989(0.004)	0.998(0.001)	0.028(0.013)
2000	20	0.3025(0.0181)	0.977(0.008)	0.983(0.009)	0.997(0.001)	0.039(0.007)
2000	20	0.5001(0.0128)	0.938(0.048)	0.946(0.028)	0.998(0.001)	0.069(0.014)
2000	20	0.7033((0.0114)	0.911(0.053)	0.883(0.067)	0.998(0.001)	0.079(0.008)
2000	20	0.9001(00025)	0.776(0.091)	0.794(0.128)	0.984(0.030)	0.078(0.007)
1000	40	-0.0125(0.0258)	0.974(0.013)	0.955(0.060)	0.995(0.000)	0.050(0.019)
1000	40	0.2828(0.0266)	0.954(0.028)	0.953(0.038)	0.994(0.007)	0.063(0.014)
1000	40	0.4973(0.0156)	0.920(0.102)	0.916(0.038)	0.991(0.013)	0.071(0.016)
1000	40	0.7067(0.0130)	0.863(0.074)	0.911(0.056)	0.996(0.001)	0.079(0.007)
1000	40	0.8996(0.0033)	0.717(0.177)	0.725(0.091)	0.995(0.004)	0.079(0.006)
1000	20	0.0009(0.0275)	0.971(0.024)	0.978(0.007)	0.995(0.002)	0.034(0.011)
1000	20	0.2846(0.0228)	0.952(0.032)	0.959(0.021)	0.994(0.003)	0.055(0.018)

1000	20	0.5013(0.0268)	0.955(0.015)	0.921(0.039)	0.997(0.001)	0.066(0.012)
1000	20	0.6978(0.0130)	0.884(0.054)	0.894(0.049)	0.994(0.002)	0.075(0.006)
1000	20	0.8944(0.0058)	0.713(0.148)	0.723(0.157)	0.976(0.055)	0.079(0.010)

RMSE shown in last column of Table 3 was computed to examine the effect of the number of trait dimensions on the accuracy of the recovered item response surface by SAS MDIRT program. For more than three dimensions, the number of quadrature points is fixed at 3 because the computational work exponentially increases depending on the number of quadrature points. However, 5 quadrature points was used to examine the effect the number of quadrature points on the accuracy of the estimated item response surface, which is done only for three-dimension.

Table 3. means and standard deviations of the RMSE between the estimated ($\hat{P}$) and true probability (P) for more than three dimensions.

# Examinees	# Items	# Dimensions	# Quadrature	RMSE
2000	40	3	5	0.0758(0.0159)
2000	40	3	3	0.0982(0.0133)
2000	40	4	3	0.1261(0.0132)
2000	40	5	3	0.1503(0.0457)
2000	20	3	5	0.0801(0.0293)
2000	20	3	3	0.0832(0.0104)
2000	20	4	3	0.1381(0.0293)
2000	20	5	3	0.1812(0.0398)
1000	40	3	5	0.0856(0.0202)
1000	40	3	3	0.0997(0.0140)
1000	40	4	3	0.1303(0.0142)
1000	40	5	3	0.1881(0.0518)
1000	20	3	5	0.0828(0.0234)
1000	20	3	3	0.1022(0.0262)
1000	20	4	3	0.1388(0.0308)
1000	20	5	3	0.1861(0.0453)

The results show that there are four factors that affect the accuracy of the SAS

MDIRT program in reproducing the multidimensional item characteristic curve (MICC): the number of examinees, the number of items, the number of quadrature points, and the degree of correlation between latent dimensions. In Table 2, the accuracy of the estimated parameters decreases as the correlation between two dimensions increases. Both Table 2 and Table 3 show that the SAS MDIRT program tends to recover MICC better when either the number of examinees or the number of items is increased. It is noticeable that the precision of the estimated MICC is mainly obtained from increasing the number of quadrature points. However, the effect of the number of quadrature points on the accuracy of recovered MICC decreases as the number of dimensions increases. For instance, using 3 quadrature points to reproduce the MICC for three dimensions does not provide the same level of accuracy when they are used for four or five dimensions. Using more quadrature points is recommended as the number of dimensions increases to compensate for the reduced accuracy.

Application of MDIRT SAS program to NLSY79 data

It is of great importance to identify the nature of items correctly because it plays a crucial role in evaluating the construct validity of a measurement. For instance, the appropriateness of practical applications of IRT, such as computer adaptive test and score equating, depends on how well the dimensionality of underlying latent variables is defined. Reckase (1985) demonstrated that math items used for the ACT, presumed to measure purely math skill, were multidimensional. In addition, he showed that item

difficulty is confounded with dimensions, which makes even more invalid the interpretation of the scores on the test.

The purpose of this empirical study is to examine the dimensionality of the PIAT mathematics test, a subtest of the Peabody Individual Achievement Test (PIAT) and to identify the confounding of item difficulty with latent dimensions. This research will provide a useful information base about the validity and dimensionality of the test, which could result in positive revision or change of content of the test.

Method

The PIAT Math Test

Biennial assessments have been conducted since 1986 to all children born to women interviewed for the National Longitudinal Survey of Youth (NLSY79), to measure the children's cognitive ability. For this purpose, the Peabody Individual Achievement Test, one of most widely used cognitive assessment instrument, known for its high test-retest reliability and concurrent validity, has been administered to the children to assess their academic achievement. In this dissertation, the PIAT math, subscale of the PIAT, was examined. It was dichotomously recorded (1 for correct response or 0 otherwise) and consisted of 84 multiple-choice items.

Since 1994 children have been administered the PIAT math test with CAPI (computer-assisted personal interview), which aimed to make the content of the interview clear, complete, and compact. Assuming that only one latent skill, math ability, is measured, all items are ranked in order of their difficulty level from easiest to hardest.

However, in the actual administration, no one was required to answer all the items of PIAT math test. A child started with a particular starting question, which was determined by the physical age and grade of the children. When a child answered a starting question wrong, he was presented the starting question of the next lower level. The basal level was established when a child made five consecutive correct responses, and the ceiling was reached when a respondent answered five out of seven items incorrectly. The final raw score was calculated by subtracting the number of incorrect responses between the basal and ceiling from the question number of the last question which was answered wrong.

Analysis

For this dissertation, the response data of children whose ages ranged from 5 to 15 on the PIAT math test were calibrated with MDIRT SAS program. Only 1998 responses (the third CAPI administration) were used. Missing responses were simply replaced with 1 before the basal 0 after the ceiling. The raw score computation implied that the PIAT math test is a Guttman scale, assuming only one latent trait is measured. Guttman's insight was that for unidimensional scales, those who make a correct response to a more difficult item will also answer all less difficult items right that preceded it. Thus, to some extent, a unidimensional scale is imposed by the administration and scoring procedure.

The NLSY children were classified in three different groups according to their age; children whose age were 5 to 7, 8 to 11 and 12 to 15 in 1998. The main reason that the children were categorized into different groups is that they are in different cognitive

developmental stages, which could affect the analysis. In addition, a different set of items was selected for each group for the stability of estimated item parameters. As mentioned, the nature of the administration didn't allow older children who established basal to answer the items below the basal while younger children who reached ceiling weren't shown to the items above the ceiling. Thus, only items in the range typically administered for the particular age were analyzed for each age group.

Results

Descriptive statistics from the selected PIAT math items for the children whose age are 5 to 7 are shown in Table 6, 8 to 11 in Table 7, and 12 to 15 in Table 8. The items are roughly in order according to their difficulty level. Item difficulty level may fluctuate from one group to another. To prevent the effect of the fluctuation of item difficulty on achievement level of examinees, basal and ceiling were established by locating five correctly answered items and five incorrectly answered items out of seven. As mentioned, responses of an examinee before basal are automatically recorded as correct while responses after ceiling recorded as incorrect, assuming a Guttman scale is applicable to the test.

Table 4. Item statistics of 30 items selected for the children whose ages were 5 to 7 in 1998

ITEM	# TRIED	# RIGHT	P
1	915	907	0.9912568
2	915	902	0.9857923
3	915	905	0.989071
4	915	900	0.9836066
5	915	898	0.9814208
6	915	796	0.8699454
7	915	835	0.9125683
8	915	882	0.9639344
9	915	804	0.8786885
10	915	855	0.9344262
11	915	598	0.6535519
12	915	645	0.704918
13	915	770	0.8415301
14	915	584	0.6382514
15	915	560	0.6120219
16	915	439	0.4797814
17	915	549	0.6
18	915	405	0.442623
19	915	415	0.4535519
20	915	421	0.4601093
21	915	409	0.4469945
22	915	456	0.4983607
23	915	356	0.389071
24	915	282	0.3081967
25	915	383	0.4185792
26	915	249	0.2721311
27	915	277	0.3027322
28	915	145	0.1584699
29	915	136	0.1486339
30	915	129	0.1409836

Table 5. Item statistics of 35 items selected for the children whose ages were 8 to 11 in 1998

ITEM	# TRIED	# RIGHT	P
21	1519	1466	0.9651086
22	1519	1445	0.9512837
23	1519	1361	0.8959842
24	1519	1345	0.885451
25	1519	1406	0.925609
26	1519	1337	0.8801843
27	1519	1357	0.8933509
28	1519	1174	0.7728769
29	1519	1192	0.7847268
30	1519	1294	0.8518762
31	1519	1239	0.8156682
32	1519	1185	0.7801185
33	1519	1241	0.8169849
34	1519	1147	0.755102
35	1519	1010	0.6649111
36	1519	1110	0.7307439
37	1519	1327	0.8736011
38	1519	806	0.5306122
39	1519	1056	0.6951942
40	1519	620	0.4081633
41	1519	1149	0.7564187
42	1519	569	0.3745885
43	1519	951	0.6260698
44	1519	950	0.6254115
45	1519	608	0.4002633
46	1519	719	0.4733377
47	1519	400	0.2633311
48	1519	829	0.5457538
49	1519	320	0.2106649
40	1519	377	0.2481896
51	1519	600	0.3949967
52	1519	337	0.2218565
53	1519	393	0.2587228
54	1519	344	0.2264648
55	1519	232	0.1527321

Table 6. Item statistics of 32 items selected for the children whose ages were 12 to 15 in 1998

ITEM	# TRIED	# RIGHT	P
38	953	819	0.8593914
39	953	885	0.9286464
40	953	688	0.7219307
41	953	916	0.9611752
42	953	752	0.7890871
43	953	841	0.8824764
44	953	845	0.8866737
45	953	657	0.6894019
46	953	708	0.7429171
47	953	531	0.5571878
48	953	764	0.8016789
49	953	384	0.4029381
50	953	504	0.5288562
51	953	543	0.5697796
52	953	484	0.5078699
53	953	523	0.5487933
54	953	403	0.4228751
55	953	345	0.3620147
56	953	271	0.2843652
57	953	459	0.4816369
58	953	337	0.3536201
59	953	258	0.270724
60	953	346	0.363064
61	953	298	0.3126967
62	953	249	0.2612802
63	953	349	0.366212
64	953	215	0.2256034
65	953	173	0.181532
66	953	179	0.1878279
67	953	132	0.13851
68	953	192	0.201469
69	953	60	0.0629591

Table 7. Item parameters for NLSY79 children whose ages were 5 to 7 in 1998

ITEM	INTERCEPT	SLOPE1	SLOPE2	MDISC	MID	DC1	DC2	A1	A2
1	6.50006	1.81034	1.73172	2.50523	-2.59459	0.72262	0.69124	43.7285	46.2715
2	8.48042	2.85889	2.31183	3.67665	-2.30656	0.77758	0.62879	38.9605	51.0395
3	4.18705	1.33533	-0.16110	1.34501	-3.11302	0.99280	0.11978	6.8793	83.1207
4	6.12678	2.40091	0.85929	2.55004	-2.40262	0.94152	0.33697	19.6924	70.3076
5	2.78529	0.67972	-0.21452	0.71277	-3.90770	0.95363	0.30097	17.5159	72.4841
6	1.49715	0.74415	-0.33561	0.81633	-1.83401	0.91158	0.41112	24.2749	65.7251
7	2.16990	0.99538	-0.61314	1.16907	-1.85610	0.85143	0.52447	31.6325	58.3675
8	4.16843	1.51584	-1.27877	1.98318	-2.10189	0.76435	0.64481	40.1512	49.8488
9	1.93716	1.03248	-0.70893	1.25243	-1.54672	0.82438	0.56604	34.4747	55.5253
10	2.48215	1.11746	-0.46653	1.21093	-2.04979	0.92281	0.38526	22.6602	67.3398
11	0.50757	0.68799	-0.30023	0.75065	-0.67618	0.91653	0.39996	23.5754	66.4246
12	0.85190	1.12518	-0.32568	1.17136	-0.72727	0.96057	0.27803	16.1427	73.8573
13	1.60957	1.05354	-0.56339	1.19472	-1.34724	0.88183	0.47157	28.1360	61.8640
14	0.55329	1.09157	-0.33264	1.14113	-0.48486	0.95657	0.29151	16.9481	73.0519
15	0.45450	0.98152	-0.60284	1.15187	-0.39457	0.85211	0.52336	31.5577	58.4423
16	0.04597	0.90608	-0.12206	0.91427	0.05028	0.99105	0.13350	7.6720	82.3280
17	0.40685	1.16448	-0.23871	1.18870	-0.34226	0.97963	0.20081	11.5845	78.4155
18	0.23363	1.38636	-0.16748	1.39644	0.16731	0.99278	0.11993	6.8881	83.1119
19	0.20085	1.48256	-0.16971	1.49225	0.13460	0.99351	0.11373	6.5301	83.4699
20	0.21384	1.74863	-0.55632	1.83499	0.11654	0.95294	0.30317	17.6484	72.3516
21	0.28349	1.77025	-0.44984	1.82651	0.15521	0.96920	0.24629	14.2578	75.7422
22	-0.06556	2.57227	-0.56310	2.63318	0.02490	0.97687	0.21385	12.3478	77.6522
23	-0.74246	2.30834	-0.42779	2.34765	0.31626	0.98326	0.18222	10.4992	79.5009
24	-1.20122	2.32262	-0.21657	2.33270	0.51495	0.99568	0.09284	5.3272	84.6728
25	-1.21409	4.04931	-0.19386	4.05395	0.29948	0.99886	0.04782	2.7410	87.2590
26	-1.28666	2.04832	-0.21455	2.05952	0.62474	0.99456	0.10417	5.9796	84.0204
27	-1.39519	2.41561	-0.89055	2.57453	0.54192	0.93827	0.34591	20.2371	69.7629
28	-1.55316	1.27095	-0.36506	1.32234	1.17455	0.96114	0.27607	16.0257	73.9743
29	-2.17931	1.93910	-0.63469	2.04033	1.06812	0.95039	0.31107	18.1239	71.8761
30	-2.13443	1.69886	-0.80037	1.87795	1.13657	0.90463	0.42619	25.2262	64.7738

(DC: directional cosine, A: angle)

Table 8. Item parameters for NLSY79 children whose ages were 8 to 11 in 1998

ITEM	INTERCEPT	SLOPE1	SLOPE2	MDISC	MID	DC1	DC2	A1	A2
21	4.4601	1.75060	-1.22669	2.13761	-2.08648	0.81895	0.57386	35.0199	54.9801
22	3.2962	1.37706	-0.83297	1.60939	-2.04809	0.85564	0.51757	31.1692	58.8308
23	2.6185	1.48466	-0.94484	1.75981	-1.48792	0.84365	0.53690	32.4728	57.5272
24	2.4119	1.41020	-0.91868	1.68304	-1.43306	0.83789	0.54585	33.0825	56.9175
25	2.1710	0.94251	-0.40147	1.02445	-2.11922	0.92001	0.39189	23.0721	66.9279
26	2.2754	1.36568	-0.82637	1.59624	-1.42549	0.85556	0.51770	31.1780	58.8220
27	2.5710	1.53443	-0.74665	1.70644	-1.50662	0.89920	0.43755	25.9474	64.0526
28	1.3301	1.31567	-0.64815	1.46666	-0.90687	0.89705	0.44192	26.2267	63.7733
29	1.6282	1.61059	-0.75762	1.77989	-0.91478	0.90489	0.42566	25.1922	64.8078
30	1.7869	1.23986	-0.41918	1.30881	-1.36529	0.94732	0.32028	18.6797	71.3203
31	1.5815	1.25896	-0.64555	1.41482	-1.11784	0.88984	0.45628	27.1470	62.8530
32	1.6010	1.63358	-0.73411	1.79094	-0.89393	0.91213	0.40990	24.1985	65.8015
33	1.7103	1.43399	-0.60225	1.55532	-1.09966	0.92199	0.38722	22.7814	67.2186
34	1.3409	1.63372	-0.42317	1.68764	-0.79456	0.96805	0.25075	14.5217	75.4783
35	0.5935	1.21295	-0.25397	1.23925	-0.47888	0.97878	0.20493	11.8257	78.1743
36	1.1183	1.59517	-0.26172	1.61650	-0.69183	0.98681	0.16190	9.3174	80.6826
37	11.4639	8.94786	-2.88224	9.40061	-1.21948	0.95184	0.30660	17.8545	72.1455
38	0.0035	0.85375	-0.12855	0.86338	-0.00400	0.98885	0.14890	8.5629	81.4371
39	0.8453	1.55441	-0.12549	1.55947	-0.54205	0.99676	0.08047	4.6156	85.3844
40	-0.4072	0.79423	0.02570	0.79464	0.51247	0.99948	0.03234	1.8531	88.1469
41	1.4576	1.89816	0.18847	1.90750	-0.76416	0.99511	0.09880	5.6703	84.3297
42	-0.6864	1.16045	-0.05924	1.16196	0.59076	0.99870	0.05098	2.9224	87.0776
43	0.3616	1.14305	0.10629	1.14798	-0.31500	0.99570	0.09259	5.3128	84.6872
44	0.3652	1.27098	0.28147	1.30177	-0.28053	0.97635	0.21622	12.4870	77.5130
45	-0.5251	0.98692	0.21345	1.00974	0.52008	0.97740	0.21139	12.2036	77.7964
46	-0.2598	1.01653	0.19437	1.03494	0.25101	0.98221	0.18781	10.8248	79.1752
47	-1.1641	1.11767	0.30325	1.15808	1.00516	0.96511	0.26186	15.1802	74.8198
48	-0.1339	2.01509	0.54590	2.08773	0.06415	0.96521	0.26148	15.1579	74.8421
49	-1.3608	1.04942	0.31890	1.09680	1.24071	0.95680	0.29075	16.9031	73.0969
50	-1.3394	1.24986	0.35438	1.29912	1.03097	0.96208	0.27278	15.8299	74.1701
51	-0.9698	1.74658	0.79694	1.91981	0.50518	0.90977	0.41512	24.5266	65.4734
52	-1.7704	1.58639	0.46606	1.65344	1.07074	0.95945	0.28187	16.3720	73.6280
53	-2.0134	2.13899	0.53838	2.20571	0.91280	0.96975	0.24408	14.1277	75.8723
54	-2.4102	2.24589	0.92934	2.43057	0.99160	0.92402	0.38235	22.4795	67.5205
55	-2.4387	1.80023	0.51524	1.87251	1.30236	0.96140	0.27516	15.9717	74.0283

(DC: directional cosine, A: angle)

Table 9. Item parameters for NLSY79 children whose ages were 12 to 15 in 1998

ITEM	INTERCEPT	SLOPE1	SLOPE2	MDISC	MID	DC1	DC2	A1	A2
38	2.43429	1.75117	0.96018	1.99713	-1.21889	0.87684	0.48078	28.7364	61.2636
39	3.28238	1.91444	0.24018	1.92945	-1.70120	0.99222	0.12448	7.1509	82.8491
40	1.20420	1.63511	0.52773	1.71816	-0.70087	0.95166	0.30715	17.8874	72.1126
41	4.47202	2.20412	0.43523	2.24668	-1.99050	0.98106	0.19372	11.1700	78.8300
42	1.48615	1.43063	0.46702	1.50493	-0.98752	0.95063	0.31032	18.0788	71.9212
43	1.92683	1.17818	0.36162	1.23242	-1.56345	0.95598	0.29342	17.0629	72.9371
44	1.94739	1.17381	0.30585	1.21300	-1.60543	0.96769	0.25214	14.6044	75.3956
45	0.72489	1.01375	0.24131	1.04208	-0.69562	0.97282	0.23156	13.3892	76.6108
46	0.83604	0.78278	0.09862	0.78897	-1.05966	0.99216	0.12500	7.1810	82.8190
47	0.21525	1.24479	0.05327	1.24593	-0.17276	0.99909	0.04276	2.4506	87.5494
48	1.13178	0.85038	0.08672	0.85479	-1.32406	0.99484	0.10146	5.8230	84.1770
49	-0.33147	0.86158	0.11755	0.86956	0.38120	0.99082	0.13518	7.7689	82.2311
50	0.07278	0.92865	-0.21138	0.95240	-0.07642	0.97506	0.22194	12.8232	77.1768
51	0.21316	0.79695	-0.09678	0.80280	-0.26552	0.99271	0.12055	6.9241	83.0759
52	-0.02752	1.20118	-0.33271	1.24640	0.02208	0.96372	0.26693	15.4818	74.5182
53	0.15286	1.31235	-0.31837	1.35041	-0.11319	0.97181	0.23576	13.6361	76.3639
54	-0.35989	1.10642	-0.33338	1.15555	0.31144	0.95748	0.28851	16.7685	73.2315
55	-0.63157	1.12122	-0.46169	1.21256	0.52086	0.92468	0.38075	22.3804	67.6196
56	-1.02594	1.19562	-0.57766	1.32786	0.77263	0.90042	0.43503	25.7872	64.2128
57	-0.33404	2.03231	-0.87816	2.21392	0.15088	0.91797	0.39665	23.3691	66.6309
58	-0.94176	1.61106	-0.72459	1.76650	0.53312	0.91200	0.41018	24.2164	65.7836
59	-1.53044	1.83598	-0.75491	1.98512	0.77095	0.92487	0.38028	22.3512	67.6488
60	-1.14596	2.06808	-0.89270	2.25252	0.50874	0.91812	0.39631	23.3478	66.6522
61	-1.74503	2.45739	-1.06241	2.67721	0.65181	0.91789	0.39683	23.3803	66.6197
62	-1.61630	1.78931	-0.95044	2.02607	0.79775	0.88314	0.46911	27.9763	62.0237
63	-4.83812	7.70700	-3.84362	8.61228	0.56177	0.89489	0.44630	26.5063	63.4937
64	-2.03246	2.04796	-1.00849	2.28281	0.89033	0.89713	0.44178	26.2172	63.7828
65	-2.27305	2.01849	-0.93616	2.22501	1.02159	0.90718	0.42074	24.8816	65.1184
66	-2.23339	2.07104	-0.81334	2.22503	1.00376	0.93079	0.36554	21.4410	68.5590
67	-1.93867	1.32361	-0.74964	1.52115	1.27448	0.87014	0.49281	29.5255	60.4745
68	-2.16636	1.95821	-1.12922	2.26047	0.95837	0.86628	0.49955	29.9704	60.0296
69	-2.79540	1.56631	-0.72102	1.72429	1.62119	0.90838	0.41815	24.7179	65.2821

(DC: directional cosine, A: angle)

Estimated parameters of the MIRT 2 parameter logistic model (2PL) using the SAS MDIRT program are presented in Table 9 for age 5 to 7 group, Table 10 for age 8 to 11 group, and Table 11 for age 12 to 15 group, successively. Unlike unidimensional IRT,

directional information for a multidimensional item is necessary to fully describe it. Direction cosines (DC1, DC2) or the angles (A1, A2) give the direction in which a multidimensional item provides the best overall information about multiple latent traits of an examinee. Given the directional information of an item, a multidimensional item discrimination parameter (MDISC) is determined at the point which gives the steepest slope. Multidimensional item difficulty (MID) is the distance from the origin to the point, where MDISC is determined in theta space.

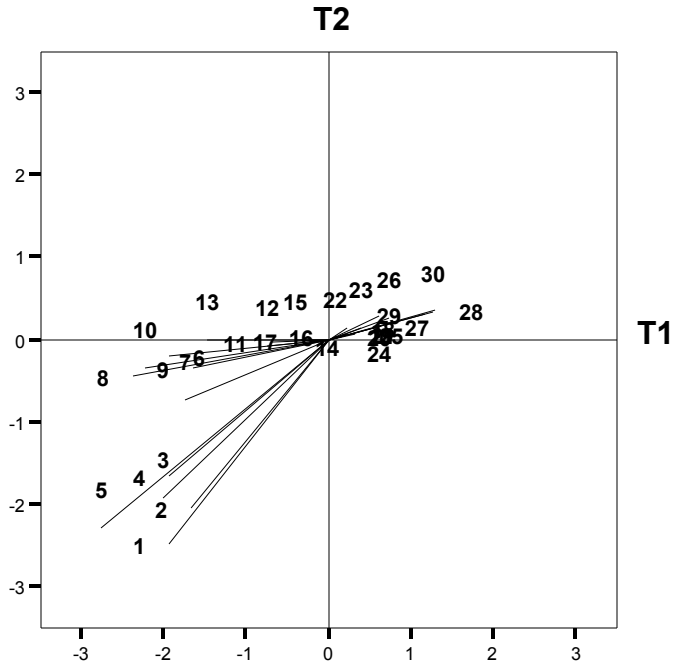

Figure 7. Graphical presentation of PIAT math items (1-30) for Children of NLSY79 whose ages were 5 to 7 using TESTFACT

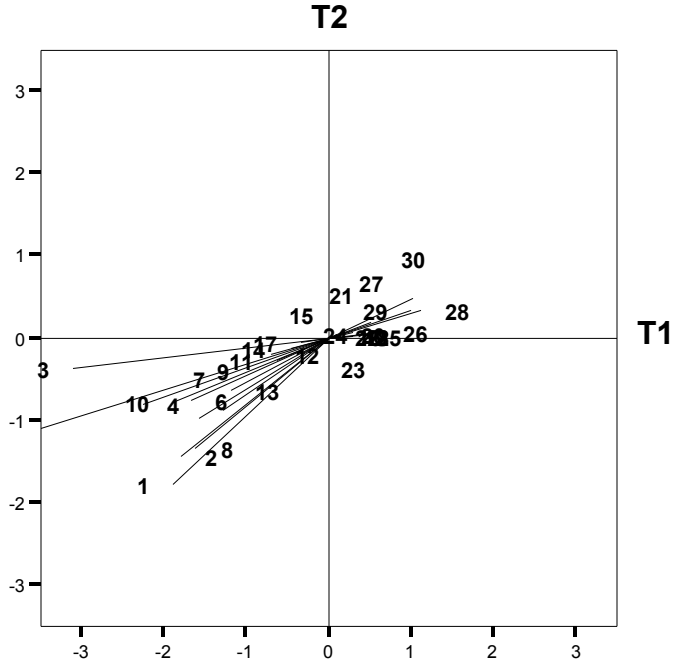

Figure 8. Graphical presentation of PIAT math items (1-30) for Children of NLSY79 whose ages were 5 to 7 using SAS MDIRT Program

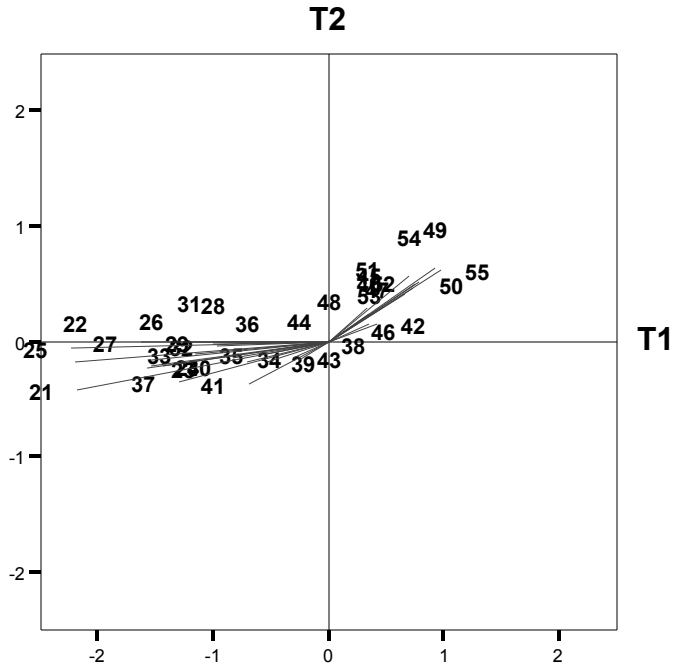

Figure 9. Graphical presentation of PIAT math items (21-55) for Children of NLSY79 whose ages were 8 to 11 using TESTFACT

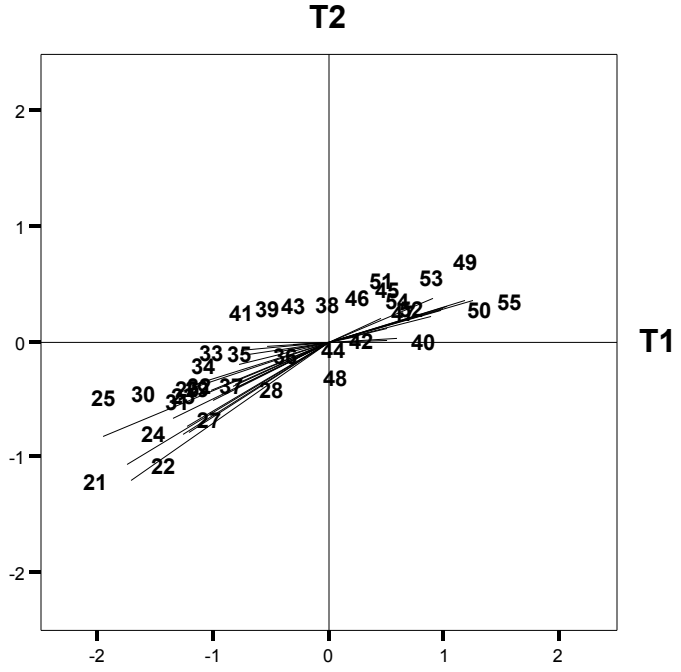

Figure 10. Graphical presentation of PIAT math items (21-55) for Children of NLSY79 whose ages were 8 to 11 using SAS MDIRT Program

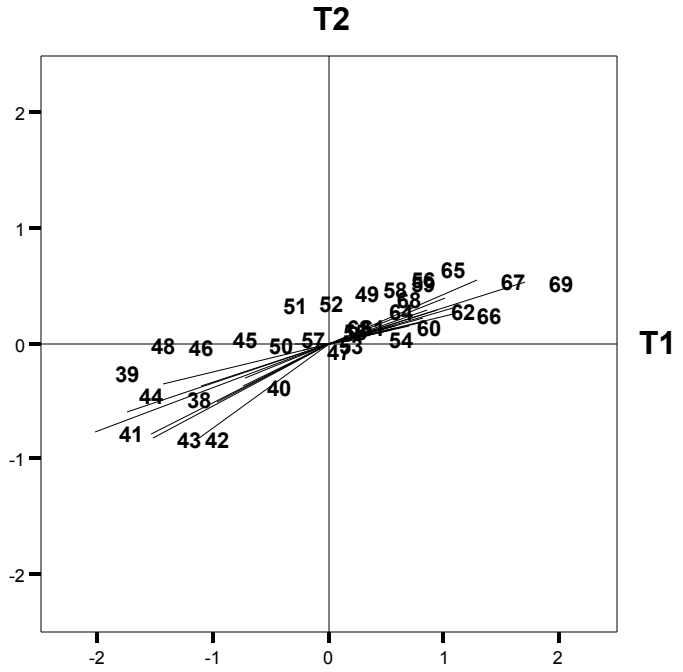

Figure 11. Graphical presentation of PIAT math items (38-69) for Children of NLSY79 whose ages were 12 to 15 using TESTFACT

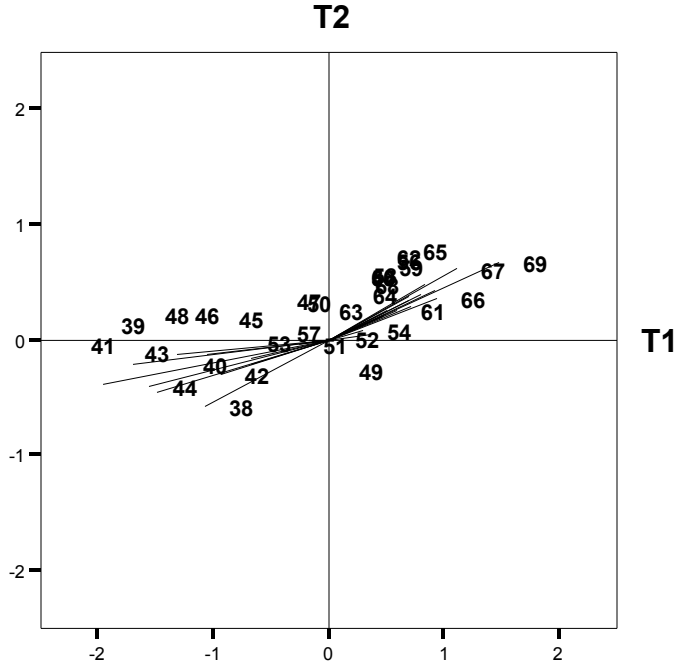

Figure 12. Graphical presentation of PIAT math items (38-69) for Children of NLSY79 whose ages were 12 to 15 using SAS MDIRT Program

Items are graphically represented in a two dimensional theta space in Figures 7 – Figure 12. Generally, most of the items are homogenous in that they are measuring roughly the same composite of two dimensions across different stages of cognitive development. It appears that most of the PIAT test items are measuring a dominant dimension and several closely correlated miscellaneous dimensions. However, those miscellaneous dimensions are not strong enough to create distinctive factors since they are highly correlated with the main factor. This result concurs with the general guideline for the use of PIAT, which warns that it is not designed to use as a diagnostic tool to measure a particular ability but designed to measure a general performance.

Generally, the estimated item parameters from the SAS MDIRT program are in agreement with those from TESTFACT (compare the pairs of MDIRT and TESTFACT graphs). The observed differences between them may come from the different methods they adopt to compute the tetrachoric correlation. TESTFACT implements Divgi's approach (1979) to computing tetrachoric correlation while SAS MDIRT employs Brown's approach (1977). In addition, TESTFACT is based on the normal probability density function while the SAS MDIRT program is approximating the normal probability function with the logistic ogive function with correction constant of 1.702.

Table 10. Comparison of factor loadings of SAS MDIRT program to those from TESTFACT after rotation for the children whose ages were 5 to 7

Item	SAS MDIRT		TESTFACT	
	Factor 1	Factor 2	Factor 1	Factor 2
1	0.45785	-0.87737	0.49068	-0.63268
2	0.55748	-0.73891	0.55577	-0.68854
3	0.72634	-0.45317	0.63517	-0.60647
4	0.73068	-0.53120	0.68129	-0.58365
5	0.56685	-0.18803	0.46592	-0.38548
6	0.63434	-0.01143	0.59778	-0.06313
7	0.75016	-0.25265	0.69924	-0.29661
8	0.84633	-0.06585	0.78432	-0.14819
9	0.76143	-0.19673	0.71181	-0.15483
10	0.77161	-0.13187	0.69667	-0.11127
11	0.60356	-0.03245	0.58447	-0.03567
12	0.77465	-0.02404	0.72637	0.07380
13	0.74996	0.01818	0.69847	-0.00215
14	0.74937	0.07531	0.70718	0.14199
15	0.72949	0.03057	0.71233	-0.00842
16	0.65226	0.12985	0.62919	0.17566
17	0.73611	0.21639	0.75233	0.00395
18	0.77758	0.22013	0.71620	0.41295
19	0.79867	0.19238	0.76719	0.21297
20	0.85984	0.16000	0.84326	0.12889
21	0.85292	0.20512	0.83168	0.18062
22	0.89618	0.27472	0.92181	0.07748
23	0.88584	0.29891	0.85084	0.28479
24	0.87864	0.38269	0.81508	0.36356
25	0.89054	0.45175	0.90352	0.15432
26	0.85795	0.25953	0.82027	0.27526
27	0.91868	0.12024	0.88793	0.21157
28	0.78434	0.10891	0.73994	0.19800
29	0.89458	0.08905	0.85404	0.21505
30	0.88780	-0.06208	0.81949	0.21786

Table 11. Comparison of factor loadings of SAS MDIRT program to those from TESTFACT after rotation for the children whose ages were 8 to 11

Item	SAS MDIRT		TESTFACT	
	Factor 1	Factor 2	Factor 1	Factor 2
21	0.91513	0.16869	0.87751	0.16759
22	0.86004	0.05880	0.81611	0.06772
23	0.87340	0.11159	0.86402	0.13259
24	0.85549	0.09166	0.84843	0.11619
25	0.73772	0.00384	0.69129	0.01806
26	0.84539	0.06903	0.83274	0.08481
27	0.86687	-0.00326	0.84452	-0.00161
28	0.83854	0.03092	0.82881	0.01734
29	0.88005	-0.01344	0.87775	0.00287
30	0.80500	-0.09886	0.76954	-0.11191
31	0.81668	-0.00488	0.80313	-0.00238
32	0.87318	-0.04023	0.86357	-0.04320
33	0.84755	-0.05862	0.82463	-0.07568
34	0.85227	-0.15880	0.84409	-0.14601
35	0.79274	-0.14263	0.78156	-0.16485
36	0.83719	-0.21442	0.82366	-0.21692
37	0.96990	-0.20931	0.96404	-0.26452
38	0.68970	-0.13590	0.68319	-0.16368
39	0.81965	-0.27297	0.80495	-0.27498
40	0.58908	-0.26248	0.59874	-0.26728
41	0.78541	-0.43264	0.75872	-0.41215
42	0.75109	-0.22909	0.74367	-0.24561
43	0.70415	-0.33910	0.69246	-0.34083
44	0.69470	-0.44404	0.68271	-0.43578
45	0.62937	-0.38835	0.62651	-0.40051
46	0.65257	-0.38423	0.65289	-0.37954
47	0.67761	-0.42297	0.66525	-0.41771
48	0.76823	-0.50625	0.75721	-0.50632
49	0.62717	-0.42028	0.62446	-0.42613
50	0.68670	-0.43999	0.68719	-0.44394
51	0.69451	-0.56295	0.68096	-0.58612
52	0.73440	-0.47417	0.73306	-0.47206
53	0.80629	-0.48535	0.78930	-0.48328
54	0.72458	-0.58082	0.72520	-0.58127
55	0.75380	-0.47803	0.75361	-0.47272

Table 12. Comparison of factor loadings of SAS MDIRT program to those from TESTFACT after rotation for the children whose ages were 12 to 15

Item	SAS MDIRT		TESTFACT	
	Factor 1	Factor 2	Factor 1	Factor 2
38	0.71155	-0.48369	0.69561	-0.50825
39	0.85248	-0.24693	0.81544	-0.27385
40	0.78183	-0.37815	0.75780	-0.37274
41	0.86191	-0.29964	0.82667	-0.31577
42	0.75541	-0.38354	0.73752	-0.39133
43	0.68813	-0.34049	0.65353	-0.35288
44	0.71504	-0.32646	0.66882	-0.34298
45	0.67762	-0.29416	0.66202	-0.28134
46	0.58214	-0.17192	0.57085	-0.19161
47	0.76763	-0.17305	0.76192	-0.17347
48	0.61606	-0.11849	0.58453	-0.14374
49	0.61391	-0.16672	0.61239	-0.16469
50	0.68998	0.01163	0.68101	0.00146
51	0.60110	0.01136	0.60392	-0.02231
52	0.76572	0.09495	0.75478	0.07880
53	0.79159	0.06670	0.77837	0.05131
54	0.73292	0.11076	0.72776	0.10174
55	0.75183	0.12521	0.73599	0.14230
56	0.75507	0.21562	0.74321	0.19636
57	0.87236	0.23209	0.85901	0.21510
58	0.83380	0.22199	0.81884	0.20014
59	0.87078	0.19182	0.85629	0.19797
60	0.87766	0.23552	0.86419	0.23179
61	0.91246	0.17332	0.89981	0.19944
62	0.84794	0.26991	0.83097	0.28762
63	0.93907	0.22919	0.92562	0.23839
64	0.87578	0.25875	0.86592	0.26692
65	0.87743	0.21898	0.86803	0.24194
66	0.89042	0.17289	0.88127	0.20173
67	0.78064	0.27200	0.76449	0.32133
68	0.84991	0.27847	0.83723	0.33089
69	0.83574	0.22742	0.80358	0.24784

The estimated factor loadings from SAS the MDIRT program are rotated to match those from TESTFACT because the estimated parameters can vary depending on their starting values. The rotated results are shown in Table 10 – Table 12 for three different age groups. The factor loadings on the first dimension from the SAS MDIRT program are well matched with those from TESTFACT, while the sign of the factor loadings on the second dimension tends to be inversely estimated. For the age 5-7 group, the factor loadings of 30 items on two dimensions are rotated 28° clockwise. The factor loadings of 35 items on two dimensions are rotated 18° clockwise for the age 8-11 group. For the age 12-15 group, the factor loadings of 32 items on two dimensions are rotated 39° clockwise. In addition, the sign of factors loadings on the second dimension is inversed for the three different age groups.

Discussion

In this paper, the accuracy of recovered parameters by SAS MDIRT macro, still relatively untested but comparable to software such as TESTFACT and NOHARM, has been verified. However, additional advanced estimation techniques need to be implemented to make it more efficient. First, an acceleration technique can be added to the SAS program to reduce computer running time in the process of estimating parameters. It is a well known property of maximum likelihood estimation method that it becomes notoriously slow as it approaches the true value. In TESTFACT, an acceleration technique suggested by Ramsey (1975) is implemented to facilitate the estimation of parameters in the maximization step of EM algorithm.

Second, an option for a prior distribution of parameters needs to be available in order to prevent item parameters from drifting toward positive or negative infinity (Heywood case). The phenomenon becomes even worse when a pseudo guessing parameter is added to the model. Various prior distributions for the uniqueness were suggested; an exponential distribution, an inverted γ prior distribution**,** β prior distribution**.** Maximizing the posterior density function gives a practical advantage over maximizing the likelihood density function by setting the bound within which discrimination parameters can be stably estimated. However, it should be remembered that an incorrectly imposed prior distribution can cause serious deterioration in the quality of estimated parameters (Baker (1992). Mislevy (1986b) pointed out that incorrectly specifying the prior distribution is likely to result in an "ensemble bias." This means that all the estimated discrimination parameters will be biased in some fashion and the statistical properties like consistency are unlikely to hold.

In addition, the stability of MDIRT SAS program in estimating parameters with skewed distribution of ability needs to be examined. Batley & Boss (1993) showed that the recovery of trait and item parameters using the MIRTE (Carlson, 1987) program is adversely affected when the range of trait level of the second dimension is restricted. In addition, the effect of correlated dimensions on the estimation of parameters and the interaction effect between the skewed ability distribution and correlated dimensions are in question. Considering that it is hardly possible to find a perfectly orthogonal latent space as far as cognitive skills are concerned, correlated latent variables in estimating parameters should be further examined.

Application of MIRT Models to many psychological phenomena that are multidimensional in nature has provided a frame in which their collective properties can be more clearly understood. Personality research is among many practical spheres to which multidimensional IRT can be applied. An attempt to measure personality with a computerized adaptive test (CAT) based on MIRT, an ultimate goal of IRT practitioners, could estimate personal traits more precisely with fewer items than traditional practice.

Examining raters' behavior using an MIRT approach is an interesting area as well. Raters' behavior can be analyzed as an item is calibrated. In fact, the application of IRT to rating data has been attempted, to give a successful index for the accuracy of raters and agreement among them. However, if we examine the inconsistency problem among raters in a multidimensional perspective, we might give a more complex but more detailed answer to the question. For example, multidimensional characteristics of raters such as the number of criteria dimensions they are using, their threshold on each dimension, and the dimension to which they are most sensitive could be investigated.

Application of MIRT to longitudinal data is an interesting area as well. We might provide deeper understanding of the famous Flynn effect (Flynn, 1984) in which the mean score on intelligence tests has been increasing over the last few decades. We cannot draw a simple conclusion about whether the difference between different age groups on the same items comes from a real difference in their intellectual ability or something else (Rodgers, 1999). If the Flynn effect exists (Rodgers and Gissberg, 2006), MIRT might help us to pinpoint exactly on which dimension it happens, because we can talk about a sub-domain score instead of a total score on the PIAT Math test. In the paradigm of MIRT,

sub-domain scores on a multidimensional test obtained at different times become comparable because we can put them on the same multidimensional scale by a multidimensional equating method.

Reckase (1997) enumerated the similarities and dissimilarities between factor analysis and item response theory. Simply stated, the difference between them is how we construct unknown latent entity (top-down process vs. bottom-up process). For example, the goal of factor analysis is to estimate parameters (factor loadings) so that the correlation between the observed covariance and the reproduced covariance is maximized. The importance of an individual item is considered in relation to its contribution to the overall fit index. Consequently, the information of an individual item is ignored when analyzing data. On the other hand, IRT emphasizes the role of each item of a test in developing its construct validity. The extent to which the construct of a test is valid depends on which items are selected and how the items are organized.

Even though we can enumerate the difference between factor analysis and MIRT, essentially the only difference existing between them is the nature of the manifest variables that are assumed to measure underlying latent variables. While manifest variables in factor analysis are continuous, response variables in item response theory are categorical (binary or polytomous). Therefore, for MIRT to make a unique contribution to the arena of test and measurement, it should find a way to get around the critical weaknesses that factor analysis suffers.

For example, the study on the dimensionality of the construct under interest and its correlational structure should precede the calibration of parameters in MIRT models,

because incorrectly determined dimensionality of a test may result in a decrease of its construct validity. Moreover, reducing dimensionality to fewer interpretable dimensions can lead to more complicated problems like test equating.

Even though the MIRT Model fits the observed data and is correctly specified, the estimated parameters are subject to a rotation for a better fit. When rotation is allowed or dimensions are permitted to correlate with each other, the estimated parameters are changed. In other words, they are uniquely determined only up to a rotation of the factor space (Bock & Aitkin, 1981). Furthermore, Gorsuch (1983) mentioned "the direction of a factor is always arbitrary so any factor with a preponderance of negative salient loadings can always be reversed" (p.181). Therefore, the direction of a factor should be determined by considering both a theoretical expectation based on accumulated knowledge about the domain under interest and empirical observations of it. MIRT shows its promising future but it still has its own limitations to overcome to achieve the ultimate goal of measurement, scoring or ranking individuals as unambiguously as possible.

Bibliography

Ackerman, T. A. (1991). The use of unidimensional item parameter estimates of Multidimensional items in adaptive testing. *Applied Psychological Measurement*, 15(1), 13-24.

Ackerman T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement,* 29(1), 67- 91.

Ackerman T.A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement,* 20, 311-329.

Baker, F. B. (1988). The item log-likelihood surface for two- and three-parameter item characteristic curve model. *Applied Psychological Measurement*, 12, 387-395.

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.

Batley, R. M. & Boss, M. W. (1993). The effects on parameter estimation of correlated dimensions and a distribution-restricted trait in a multidimensional item response theory. *Applied Psychological Measurement,* 17(2), 131-141.

Bock, R. D. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46(4), 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12, 261-280.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice,* 16 (4), 21-32.

Brown, M.B. (1977). Algorithm AS 116: The tetrachoric correlation and its standard error. *Applied Statistics*, 26, 343-351.

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

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

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data Via the EM algorithm. *Journal of the Royal Statistical Society*, Series B 39, 1-38.

Divgi, D. R. (1979). Calculation of the tetrachoric correlation coefficient. *Psychometrika*, 4, 169-172

Embretson, S. and Reise, S. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Erlbaum.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, 95, 29-51

Fraser, C. (1987). NOHARM II: A Fortran Program for Fitting Both Unidimensional and Multidimensional Normal Ogive Model of Latent Trait Theory. Armidale, Australia: The University Of New England,

Haley (1952). *Estimation of the Dosage Mortality Relationship When the Dose is Subject to Error*, (Techniccal Report No. 15). Stanford, CA: Stanford University, Applied Mathematics and Statistics Laboratory.

Hambleton, R. K., van der Linden, W. J. (1982). Advances in item response theory and

applications : An introduction. *Applied Psychological Measurement*, 6, 373-378.

Hambleton, R. K., Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Norwell, MA: Kluwar Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc.

Harman, H.H. (1976). *Modern Factor Analysis, 3rd ed*. Chicago University Press.

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

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

McDonald, R. (1985). *Factor Analysis and Related Methods*. Hillsdale, NJ: Erlbaum.

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

Mislevy, R. J. & Bock, R. D. (1986). PC-BILOG: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software Inc.

Mislevy, R. J. (1986b). Bayes modal estimation in item response models. *Psychometrika*, 51, 177-195.

Muraki, E., & Bock, R. D. (1997). PARSCALE 3: IRT Based Test Scoring and Item Analysis for Graded Items and Rating Scales. Chicago: Scientific Software International, Inc.

Neyman, J., and Scott, E. L. (1948). Consistent estimates based on partially consistent

observations. *Econometrika*, 16(1), 1-32.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3, 1-18.

Oshima, T. C. & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement*, 16, 237-248.

Ramsey, J. O. (1975). Solving implicit equations in psychometric data analysis. *Psychometrika*, Vol. 40, 337-360.

Reckase, M. D. (1985) The difficulty of items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.

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

Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.

Rodgers, J. L. (1999). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, 26, 337–356.

Rodgers, J. L. & Gissberg, Linda (2006). Identification of a Flynn Effect in the NLSY:Moving from the center to the boundaries, *Intelligence*, in Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph Supplement No. 17.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331-354.

Steven, P. R & Henson, J. M. (2000). Computerization and adaptive administration of

the NEO PI-R. *Assessment*, 7(4), 347-364.

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

Thissen, D. (1991). MULTILOG: Multiple Category Item Analysis and Test Scoring Using Item Response Theory. Chicago: Scientific Software International, Inc.

Thissen, D. & Wainer, H. (Eds) (2001) *Test Scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thurstone, L. L. (1925). A Method of scaling psychological and educational tests. *Journal of Educational Psycholog,* 16. 433-451.

Wilson, D. T., Wood, R., & Gibbons, R. (1984). TESTFACT: Test Scoring, Item Statistics, and Item Factor Analysis. Mooresville, IN: Scientific Software.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of Item Response Theory (pp. 45-56. Vancouver, BC: Educational Research Institute of British Columbia.