*Handbook of Statistics, Vol. 26*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(06)26018-8



# Multidimensional Item Response Theory

*Mark D. Reckase*

## 1. Introduction

Psychometric theory deals with the way that information in persons' responses to a series of test tasks, usually called test items, is used to estimate locations of the persons (examinees) on a continuum[1] of performance, aptitude, or other psychological dimension. By continuum in this case is meant an ordering from low to high on the characteristic of interest. One subclass of theories under the general heading of psychometric theory focuses on the relationship between responses to individual items and the locations of persons on the relevant continuum. This subclass is called item response theory (IRT). IRT hypothesizes that a particular mathematical form describes the relationship. A mathematical expression relates the probability of a particular response, or score, for an item to the location on the continuum. Once a mathematical form is hypothesized, the characteristics of the continuum for locating individuals have properties in addition to order – distances between locations become meaningful. Many different mathematical forms have been proposed to represent the connection between the probability of a response to the location on the continuum and it is unclear if each of these mathematical forms constitutes a single theory or if the general approach of using such models is the theory. In general, however, hypothesizing a particular mathematical form to describe the relationship between the probability of a response and the location of the examinee on the continuum is called item response theory.

There is extensive literature on IRT including several excellent textbooks (e.g., Fischer and Molenaar, 1995; Hambleton and Swaminathan, 1985; Hulin et al., 1983; Lord, 1980). Most of the models discussed in the literature on IRT make the assumption that the goal of the measurement task is to locate the examinee on a single continuum. This is usually called the unidimensionality assumption. This means that if the location on this single continuum is known and the hypothesized mathematical form accurately describes the relationship between the probability of a response and the location, then no other information will improve the accuracy of the description of the response process. This kind of assumption is common in many areas of science even though it likely is a simplification of the real state of the world. For example, the speed of a falling object

[1] Psychometric methods also exist for estimating the classification of individuals into discrete classes. These methods are beyond the scope of this chapter.

is usually assumed to be solely a function of the strength of the force of gravity. However, the actual speed of a falling object is influenced by the resistance of the air it falls through, the shape of the object, and other factors. Physics texts are careful to indicate that the equations that describe the speed of a falling object only apply in a vacuum, but the equation for the speed of a falling object is regularly used for practical applications even though it is not absolutely correct for the real life situation.

The unidimensional assumption of most IRT models is similarly a simplification of reality. However, the IRT models might still be useful in the same way that the mathematical equations from Newtonian physics are useful even though they are not perfectly accurate. Test items are actually quite complex tasks that require multiple skills and knowledge to generate a particular response. An example of this complexity is a mathematics test item that presents a practical problem to be solved. Reading skills are required to understand the description of the task. The mathematics task requires problem solving and computational skills as well. Thus, several continua are needed to describe the location of the examinee on the skills required by this item rather than a single one. For these complex cases, the location of the examinee is thought of as a point in a multidimensional space rather then as a point along a single continuum.

Because of the complexity of real test items and the multiple skills and knowledge that examinees use when responding to items, a model based on multiple dimensions is needed. A more realistic mathematical model of the relationship between the probability of response and the location of a person uses a vector to represent the location of the examinee in a multidimensional space. The location of examinee $j$ is represented by the vector $\boldsymbol{\theta}_j = [\theta_{j1}, \theta_{j2}, \ldots, \theta_{jM}]'$, where $M$ is the number of continua (dimensions) on which the examinees need to be ordered to describe the capabilities that come to bear when responding to the test item. The elements of the vector provide coordinates for the location of the examinee in a multidimensional space. Because the location of the individual is given in a multidimensional space, the mathematical models that are used to relate the probability of the response to the location of the examinee are called multidimensional item response theory (MIRT) models. The mathematical expressions for MIRT models will be described in Section 3. A general conceptual framework for the use of MIRT models is given here.

In MIRT, examinees are assumed to be complex entities that differ in many different ways. To totally describe the dimensions on which examinees vary, a large number of dimensions would be needed. A vector whose elements are the coordinates on these dimensions would describe an examinee in this high order space. When an examinee responds to a test item, he or she will draw on the relevant dimensions to determine the response to the item. In theory, if the examinee's location in the full multidimensional space were known, it might be possible to predict the response to a test item with a high degree of certainty. However, the location in the fully specified multidimensional space is never known in practice. At best, some relatively small subset of the dimensions is available for the MIRT model. As a result the relationship between the response and the location in the space is not a deterministic one with a specific form. Instead, the best that can be done is to relate the probability of the response to the estimated location in the multidimensional space and to the estimated test item characteristics. MIRT specifies

the form of the mathematical relationship and uses it to estimate the locations of the examinees and the characteristics of the test items.

MIRT was an outgrowth of two other statistics-based methodologies. One has already been mentioned – unidimensional item response theory. The other methodology is factor analysis – a method for identifying the hypothetical dimensions that underlie the relationships among variables in a set of data. MIRT differs from the traditional applications of factor analysis in a number of important ways. First, factor analysis is typically used to analyze a matrix of correlations. Because correlations are a unit free statistic, analyzing a correlation matrix is effectively the same as analyzing variables in $z$-score form, variables with mean 0 and standard deviation 1. MIRT does not standardize the variables because the differences in item scores provide important information about the characteristics of test items. In a sense, MIRT is a special form of unstandardized factor analysis for special kinds of discrete variables – item scores. McDonald (2000) emphasizes the relationships between the two methodologies.

Second, MIRT was designed for use with variables that are discrete transformations of continuous variables such as the 0 and 1 scores for test items. The use of traditional factor analysis techniques for data of this type has resulted in spurious difficulty factors (Wherry and Gaylord, 1944), especially when the Pearson product-moment correlation is used as the basis for the analysis. The early work of McDonald (1962) in particular focused on devising ways to use factor analytic type methodologies on discrete data. MIRT is an outgrowth of that work.

A third difference between MIRT and the exploratory version of factor analysis is that the goal of exploratory factor analysis is often to identify the minimum number of hypothetical factors that will recover the relationships in the data. Parsimony in description has been a goal of this type of analysis. MIRT procedures have a different goal – the accurate representation of the form of the relationship between the probability of the response and the location in the multidimensional space. The emphasis has not been to minimize the number of dimensions. In fact, Reckase (1997b) analyzed item response data with high numbers of dimensions (25 or more). However, most of the current applications of MIRT have used fewer dimensions because of limitations in estimation programs or because it is useful to graph the results and it is difficult to graph more than three dimensions.

Many of the statistical estimation procedures used for MIRT are very similar to those used for factor analysis. In fact, the commonly used MIRT programs (e.g., NOHARM and TESTFACT) provide results in a form similar to factor analysis programs. In general, however, the results from the MIRT part of the analysis are quite different from those from traditional factor analysis mainly because the variables being analyzed are not standardized to have a mean of zero and a variance of one.

## 2. General forms of MIRT models

The basic premise of MIRT is that persons vary on many different dimensions. If full knowledge is available on all of the dimensions of variation, a person's characteristics can be represented as a point in a high-dimensional space. The coordinates of the point

are given by a vector of real numbers, $\boldsymbol{\theta}_j = [\theta_{j1}, \theta_{j2}, \ldots, \theta_{jM}]'$, where $j$ is the index for the person and $M$ is the number of dimensions in the fully defined space. Differences between persons are indicated by differences in locations in the space. Two assumptions that are often unstated but that are critical to the use of the elements of the $\boldsymbol{\theta}_j$-vector are that the vector indicates a point in a Euclidean space and the elements represent coordinates in a Cartesian coordinate system. The latter point indicates that the elements of the vector represent values on orthogonal axes. These two assumptions allow the use of the standard Euclidean distance formula to determine the distance between the locations of two persons, $j$ and $k$,

$$\Delta_{jk} = \sqrt{\sum_{g=1}^{M} (\theta_{jg} - \theta_{kg})^2}, \tag{1}$$

where $\Delta_{jk}$ is the distance between the two points and the other symbols are previously defined.

Without further assumptions or constraints on the process for estimating the elements of the $\boldsymbol{\theta}_j$-vector, the solution is indeterminate. The dimensions do not have zero points (i.e., no origin) and units of measurement have not been defined. As a result, the outcome of any order-preserving transformation of the points in the space represents the relationships in the data as well as the outcome of any other such transformation. To define the specific characteristics of the space of point estimates of persons' locations, typically called "defining the metric of the space," three constraints are used. The first two constraints are that the mean and variance of estimates of the coordinates for a sample of persons are set to specific values. Any values can be used, but most computer programs set the mean and variance to 0 and 1, respectively. Setting these values fixes the origin and unit of measurement for the estimates of the elements of the $\boldsymbol{\theta}_j$-vector. The unit of measurement and origin could also be set by fixing the parameters of an item, but this is a less commonly used approach.

The third constraint specifies the orientation of the coordinate axes in the multidimensional space. The orientation of the coordinate axes is sometimes fixed by specifying that the direction of best measurement of a test item coincides with the direction specified by an axis. Different items fix the orientation of different axes. Other approaches are to specify a statistical criterion such as maximizing the number of test items that have monotonically increasing probabilities of correct response corresponding to increases in coordinate dimensions.

For realistic situations of modeling the relationship between item responses and persons' locations, the full vector of coordinates is not known. Instead, some smaller number of coordinates is considered, $m < M$, with the value of $m$ based on practical considerations. Often $m$ is 2 or 3 to allow graphic presentation of results. Seldom is $m$ chosen to be more than 5, although there are a few notable examples using $m$ as large as 50 (e.g., Reckase et al., 2000). Because each person is not located in the full $M$-dimensional proficiency space, it is not possible to develop a functional relationship between the actual response to an item and the location of the person. Instead, MIRT models the probability of response rather then the actual response to reflect that there are likely sources of variation in people that are not included in the model.

Additional constraints that specify the metric of the space are placed on the form of the relationship between the probability of correct response and the location in the $m$-dimensional coordinate space. Through the specification of the form of the relationship, the specific characteristics of the metric for the space are defined.

The possible forms for the relationship between the probabilities of a correct response to an item and the locations of persons in the space mathematical forms have been proposed. However, for MIRT models for achievement or aptitude test items, the mathematical form is often limited by an additional constraint. The constraint is that the relationship between probabilities of correct response and locations in the space should be monotonically increasing. That is, for an increase in the value of any of the coordinates, the probability of correct response should increase.

MIRT models are typically models for the relationship between the probability of a response to one item and the location of a person in the coordinate space. The model only contains parameters for the characteristics of one item and the location for one person. The item parameters are assumed to be constant. The probability of correct response to the item changes with the location of the person. The model is not for sets of items, although a number of items are analyzed at the same time during estimation of item parameters and tests of fit of the models to item response data. An often unstated assumption is that the model will simultaneously apply to all of the items. It is always possible to fit a model to the data from one item. The simultaneously fitting of the model to multiple items is a stringent test of the usefulness of the model.

Another assumption is often included as part of MIRT. The assumption is that the probability of response to one item conditional on a person's location is independent of the probability of the response to another item conditional on the same location. This is expressed mathematically as

$$P(u_1, u_2, \ldots, u_n|\boldsymbol{\theta}_j) = \prod_{i=1}^{n} P(u_i|\boldsymbol{\theta}_j), \tag{2}$$

where $u_i$ is the response to item $i$, $n$ is the number of items on the test, and the other symbols are as previously defined. This assumption is important for the estimation of parameters and it is assumed to hold only if $\boldsymbol{\theta}_j$ contains all dimensions needed to respond to all of the items. If too few dimensions are included in $\boldsymbol{\theta}_j$, the assumption of local independence will not likely hold. This provides a means for testing if a sufficient number of dimensions have been included in the modeling of the responses to the items.

The commonly used MIRT models were partly developed as extensions of item response theory (IRT) models that had only a single parameter describing the locations of the persons (see Reckase, 1997a for a brief summary of the development of the models). That is, the persons are assumed to only vary along a single continuum. Accordingly, most MIRT models have these "unidimensional" models as special cases. This feature of the models will be described when specific forms of the models are described.

## 3. Common forms of MIRT models

Although many types of MIRT models have been proposed, two basic forms appear most frequently in the research literature and in applications of the methodology. The first of these two models is usually called the *compensatory* MIRT model because high values for one set of coordinates for a person's location can offset low values of other coordinates when computing the probability of correct response to an item. That is, persons with quite different locations in the multidimensional space can have the same probability of correct response to an item.

The logistic form of the compensatory MIRT model is given by

$$P(u_i = 1|\boldsymbol{\theta}_j) = \frac{e^{\mathbf{a}_i'\boldsymbol{\theta}_j + d_i}}{1 + e^{\mathbf{a}_i'\boldsymbol{\theta}_j + d_i}}, \tag{3}$$

where e is the mathematical constant, 2.718..., that is the base of the natural logarithms, $\mathbf{a}_i$ is an $m$-element vector of item parameters that indicates the rate the probability of a correct response changes as the location on the corresponding coordinate dimension changes and $d_i$ is a scalar item parameter that is related to the difficulty of the test question. More intuitive meanings for these parameters will be provided in Section 4.

The exponent in Eq. (3) is of the form:

$$a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + \cdots + a_{im}\theta_{jm} + d_i, \tag{4}$$

where the second subscript indicates the coordinate dimension from 1 to $m$, $i$ is the index for items, and $j$ is the index for persons. From this representation, it is clear that if this sum is equal to a particular value, $s$, there are many combinations of $\theta$-values that will lead to the same sum, even if the $\mathbf{a}$- and $d$-parameters are held constant. It is this property of the exponent of the model that gives it the label of compensatory. Reducing the value of $\theta_{jv}$ can be offset by raising the value of another person coordinate value, $\theta_{jw}$, to yield the same probability of correct response.

The general form of the relationship between the probability of a correct response, $u_i = 1$, and the location of a person in the $\boldsymbol{\theta}$-space can be seen from a plot of the model when $m = 2$. The plot for the case when $a_{i1} = 1.5$, $a_{i2} = .5$, and $d_i = .7$ is presented in Figure 1.

Several important features of the compensatory MIRT model can be noted from the graphs in Figure 1. First, the item response surface increases more quickly along the $\theta_1$-dimension than along the $\theta_2$-dimension. This is a result of the differences in the elements of the $\mathbf{a}$-parameter vector. Also, note from the contour plot that the equiprobable contours are straight lines. This is the result of the form of the exponent of e in the model equation. For the example given here, the exponent of e is of the following form:

$$1.5\theta_{j1} + .5\theta_{j2} + .7 = s. \tag{5}$$

If $s$ is equal to 0, then the probability of a correct response is .5 and any combination of $\theta$s that result in a value of the expression being 0 define the coordinates of the equiprobable contour of .5. It is clear from Eq. (5) that this is the equation for a

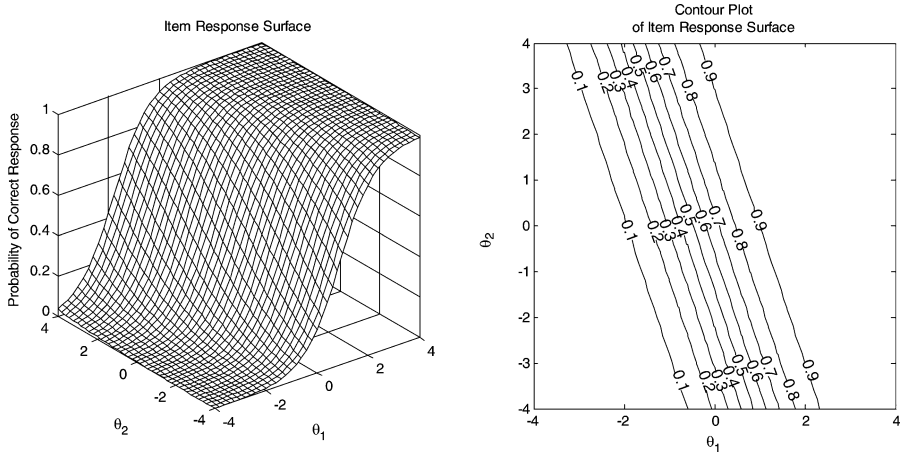


Fig. 1. Graphic representations of the compensatory model – item response surface and equiprobable contours for an item with $a_{i1} = 1.5$, $a_{i2} = .5$, and $d_i = .7$.

straight line. Different equiprobable contours are specified by changing the value of $s$. The orientation of the equiprobable contours is dependent on the **a**-parameter vector for the item and the distance the lines are from the origin is dependent on the $d$- and **a**-parameters.

Two variations of the model given in Eq. (3) are often used. One variation includes a parameter that indicates the probability of a correct response for persons who are very low on all of the $\boldsymbol{\theta}$-coordinates. This model is specifically designed for use with multiple-choice items. Such items have a non-zero probability of correct response even when persons taking the items have very little skill or knowledge because it is possible to guess at the answer. The parameter is usually labeled $c$ and the form of the model is given in Eq. (6).

$$P(u_i = 1|\boldsymbol{\theta}_j) = c_i + (1 - c_i)\frac{e^{\mathbf{a}_i'\boldsymbol{\theta}_j + d_i}}{1 + e^{\mathbf{a}_i'\boldsymbol{\theta}_j + d_i}}. \tag{6}$$

For this model, even if the term on the right (which is the same as that in Eq. (3)) is very close to 0, the probability of correct response will still be $c_i$ or slightly greater. The value of $c_i$ is a lower asymptote for the probability of correct response for the item. The $c$-parameter is a scalar rather than a vector because the process used to respond to items when a person has limited skills or knowledge is not expected to be related to any specific ability. Although the actual psychological processes persons use for responding to items when they have limited ability are not known, they are assumed to be related to guessing. However, the $c$-parameter estimates are often lower than the chance guessing level for an item. Therefore, the parameter is often called the pseudo-guessing parameter.

The second variation of the model given in Eq. (3) is derived from a family of models that have observable sufficient statistics for the item and person parameters. These

models were developed based on the work of Rasch (1960, 1961). A typical characteristic of the Rasch family of models is that the $a$-parameters for the set of items being analyzed are assumed to have the same value, usually 1. However, if the $\mathbf{a}$-parameters in the model presented in Eq. (3) are assumed to be equal, the exponent of e simplifies to

$$a_i(\theta_{j1} + \theta_{j2} + \cdots + \theta_{jm}) + d_i. \tag{7}$$

If the elements of $\mathbf{a}_i$ are equal for all dimensions, there is only an observable sufficient statistic for the sum of the elements of the $\boldsymbol{\theta}$-vector, or alternatively, the sufficient statistic for each of the elements is the same. Therefore, it is not possible to get a unique estimate for each element of the $\boldsymbol{\theta}$-vector. Because unique estimates of the $\boldsymbol{\theta}$s are not available, this approach to creating a simplified version of the model in Eq. (3) is not viable.

The approach that has been used is an extension of a general formulation of a model with observable sufficient statistics that was presented by Rasch (1961). The model can be applied to a wide variety of testing situations, but only the multidimensional version for dichotomous items is presented here. The full version of the model is presented in Adams et al. (1997) as well as other publications. The model using the same symbols as defined above is given by:

$$P(u_i = 1|\boldsymbol{\theta}_j) = \frac{e^{\mathbf{a}_i'\boldsymbol{\theta}_j + \mathbf{w}_i'\mathbf{d}}}{1 + e^{\mathbf{a}_i'\boldsymbol{\theta}_j + \mathbf{w}_i\mathbf{d}}}, \tag{8}$$

where $\mathbf{a}_i$ is a $m \times 1$ vector of scoring weights that are not estimated but are specified in advance by the person using the model; $\mathbf{d}$ is a $p$-dimensional vector of item characteristics that is constant across all items; and $\mathbf{w}_i$ is a $p \times 1$ set of weights that are set in advance that indicate the way the item characteristics influence the items.

The weight vectors for this model are usually set based on a logical analysis of the content of the items. In many cases, the elements of $\mathbf{a}_i$ and $\mathbf{w}_i$ are 0s or 1s and the vectors serve to select the person dimensions and item characteristics that apply for a specific item. It is the fact that the $\mathbf{a}_i$ and $\mathbf{w}_i$ vectors are set in advance that results in the observable sufficient statistics property of these models. Because the $\mathbf{a}_i$ and $\mathbf{w}_i$ vectors are set in advance, the use of this model has close similarities to confirmatory factor analysis and structural equation modeling. The predicted probabilities of correct response can be compared to probabilities estimated from the observed data to get a fit statistic. That statistic can be used to test hypotheses about the reasonableness of the pre-specified weights. Briggs and Wilson (2003) provide an example of this type of analysis.

The unidimensional logistic model is a special case of this model when $\theta$ is a scalar rather than a vector. The model simplifies to $a_i\theta_j + d_i = a_i(\theta_j - b_i)$ when $d_i = -a_i b_i$. The right side of the first equation in the previous sentence is the usual exponent for the unidimensional two-parameter logistic model with difficulty parameter $b_i$.

To this point, the logistic form of the compensatory MIRT model has been presented. This form of the model is used because it is more mathematically tractable. However, instead of using the cumulative logistic function as the basis of the model, the cumulative normal distribution function could be used. In fact, the early development

of these models assumed a cumulative normal form for the test characteristic surface (e.g., McDonald, 1967). The equation for this compensatory MIRT model based on the normal distribution is given by

$$P(u_{ij} = 1|\theta_j) = \int_{-\infty}^{\mathbf{a}_i'\theta_j + d_i} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\, \mathrm{d}t, \tag{9}$$

where all of the symbols have been defined previously.

The normal ogive model in (9) and the logistic model in (3) are very similar in form. Birnbaum (1968, p. 399) indicated that the probability of correct response specified by the two models with the same parameter values will differ by less than .01 over the entire range of $\boldsymbol{\theta}$ if the exponents of e in the logistic model in (3) are multiplied by 1.7. When the constant 1.7 is included in the expression for the logistic model, parameters estimated assuming the model in (9) are often used as estimates for the parameters of the model in (3). The practical implication of this similarity of models is that more computer programs are available for the estimation of model parameters.

The second of the two basic forms of the MIRT models is often called the *non-compensatory* model, although it is more correctly labeled as the *partially compensatory* model because it does not totally remove the compensation effect present in the model in (3). The partially compensatory nature of the model will be described in more detail below. The form of this model was first proposed by Sympson (1978) as an alternative to the model given in (6). The mathematical form of the model is given in Eq. (10).

$$P(u_i = 1|\theta_j) = c_i + (1 - c_i) \prod_{\ell=1}^{m} \frac{\mathrm{e}^{1.7a_{i\ell}(\theta_{j\ell} - b_{i\ell})}}{1 + \mathrm{e}^{1.7a_{i\ell}(\theta_{j\ell} - b_{i\ell})}}, \tag{10}$$

where $\ell$ is the index for coordinate dimensions, $b_{i\ell}$ is a parameter that indicates the difficulty of performing the tasks related to dimension $\ell$. Note that the initial version of the model included the constant 1.7 so that each term in the product would be similar to a normal ogive model. The item response surface and contour plot for this model are presented in Figure 2 for the two-dimensional case with $a_{i1} = 1.5$, $a_{i2} = .5$, $b_{i1} = -1$, and $b_{i2} = 0$. The $c$-parameter for the figure has been set to 0 so that it could be more easily compared to Figure 1.

The partially compensatory nature of this model is shown by the curvature of the surface. When the $c$-parameter is 0, the probability of correct response for an item based on this model can never be higher than the lowest value in the product. An increase in one of the coordinates can improve the overall probability somewhat, but only up to the limit set by the lowest term in the product.

If the terms in the product are represented in this two-dimensional case as $p_1$ and $p_2$, it is easy to see that the curves that define the equiprobable contours are hyperbolas in terms of these probability values, $h = p_1 p_2$, where $h$ is the value for any one of the contours, and $p_1$ and $p_2$ are the terms in the product in (10). The actual curves plotted in Figure 2 are not hyperbolas because they are mapped through the logistic functions to the $\theta$ coordinate system. However, like the hyperbola, they asymptote to a value as the elements of $\boldsymbol{\theta}$ approach infinity.

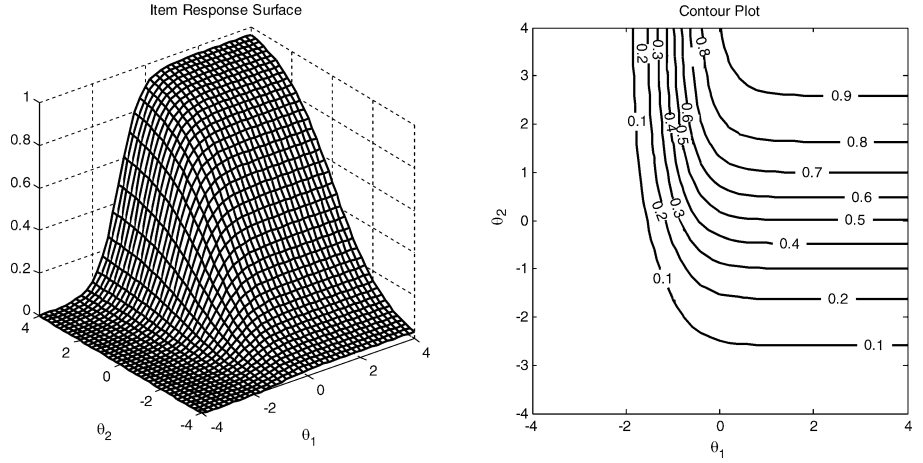


Fig. 2. Graphic representation of the partially compensatory model – item response surface and equiprobable contours for an item with $a_{i1} = 1.5$, $a_{i2} = .5$, $b_{i1} = -1$, $b_{i2} = 0$ and $c_i = 0$.

The model in (10) can be simplified by setting $c_i$ to zero. Then the model parallels the compensatory model given in (3). A further simplification of this model can be obtained by setting all of the $a_{i\ell}$ equal to 1. This model is given by

$$P(u_{ij} = 1|\boldsymbol{\theta}_j) = \prod_{\ell=1}^{m} \frac{\mathrm{e}^{(\theta_{j\ell} - b_{i\ell})}}{1 + \mathrm{e}^{(\theta_{j\ell} - b_{i\ell})}}. \tag{11}$$

Whitely (1980) suggested this model as a way to describe the interactions among the various cognitive components that enter into the performance on an item. In this case, the terms in the product follow the Rasch model form with only a difficulty parameter for each dimension. Note that if $m$ is 1 in Eqs. (10) and (11), the special cases are the unidimensional three-parameter logistic model and Rasch models, respectively.

The compensatory and partially compensatory forms of MIRT models have quite different properties. For the compensatory model, as any element of $\boldsymbol{\theta}_j$ approaches positive infinity, the value of the exponent approaches positive infinity even if the other elements of the $\boldsymbol{\theta}$-vector are very small negative values. Therefore, as any element of the $\boldsymbol{\theta}$-vector increases, the probability of correct response approaches 1.0. For the partially compensatory model, the product in (10) can never be greater than its lowest value term. If one term of the product is .1 and all other elements of the $\boldsymbol{\theta}$-vector increase so the other terms approach 1.0, the product will only approach .1. Therefore, for the partially compensatory model, the probability of correct response to the item as a whole is limited by the lowest probability of success on a component of the item.

Another difference between the two models is the influence of the number of dimensions in the model on the scaling of the coordinate space. Suppose that the $c_i$ and $d_i$ parameters in (6) are all zero. Also, suppose that a person has elements of the $\boldsymbol{\theta}$-vector that are all zero as well. The probability of correct response in that case is .5. Adding another dimension with a $\theta_{j\ell} = 0$ does not change the probability at all because the

exponent remains zero. However, consider the same case for (11) with the $b$-parameters equal to zero. For one dimension, the probability of correct response is .5, for two dimensions .25, for three dimensions .125, etc. In general, if the probability of correct response for a test item that has a constant probability of mastery for each component of the model, $p_c$, then probability of correct response will be $p_c^m$ where $m$ is the number of dimensions. If the observed proportion of correct responses for 100 persons at the same point in the $\theta$-space is .5, the values of the item parameters needed to give that proportion correct will change depending on the number of dimensions. That is, the estimates of $\boldsymbol{\theta}$ for a two-dimensional solution will be quite different in magnitude from estimates of $\boldsymbol{\theta}$ for a three-dimensional solution. This complicates the interpretation of the elements of the $\boldsymbol{\theta}$-vector for the partially compensatory model. The range of values that is considered high or low changes depending on the number of dimensions used in the model.

Although the compensatory and partially compensatory MIRT models are quite different in mathematical form and philosophical position on the way that skills and knowledge interact when applied to test items, there is some evidence that the two models fit real test data equally well. Spray et al. (1990) identified item parameters for the two models that gave similar classical item statistics. Although the item response surfaces were very different in some regions of the $\boldsymbol{\theta}$-space, for regions where the density of an assumed bivariate normal distribution was the greatest, the surfaces were similar. Figure 3 shows item response surfaces for the two models that resulted in similar proportions of correct response when a standard bivariate normal distribution of examinees was assumed. The parameters for the two models are given in Table 1.

The parameter values in Table 1 show that the parameters are quite different for the two models even though they appear to have similar functions. The $a$-parameters for the partially compensatory model are larger than those for the compensatory model. This is an effect of the multiplicative nature of the partially compensatory model.

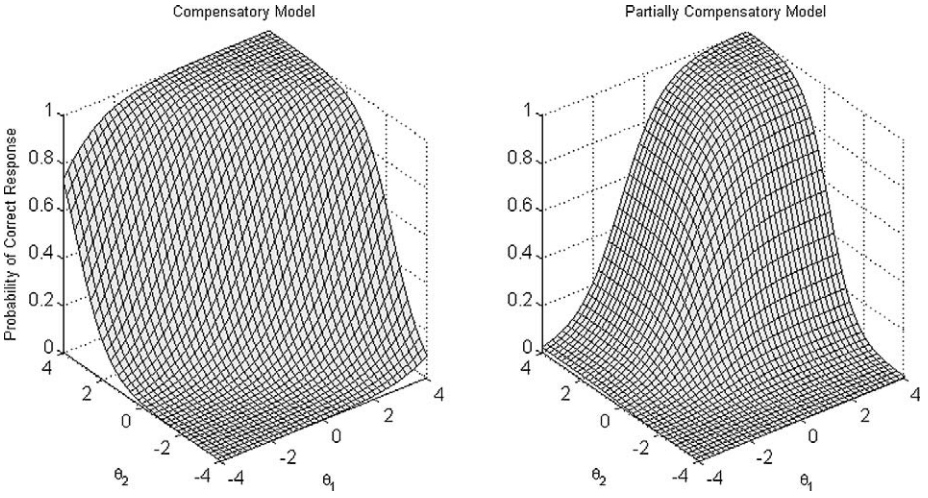


Fig. 3. Compensatory and partially compensatory models matched to yield approximately the same proportion correct assuming a standard bivariate normal distribution with $\rho = 0$.

Table 1
Parameters for matched compensatory and partially compensatory models

| Compensatory model | | Partially compensatory model | |
|---|---|---|---|
| Parameters | Values | Parameters | Values |
| $a_1$ | .90 | $a_1$ | 1.26 |
| $a_2$ | 1.31 | $a_2$ | 1.60 |
| $d$ | −.67 | $b_1$ | −.92 |
| | | $b_2$ | −.15 |

Note: The $a$-parameters for the two models do not have exactly the same meaning and are not comparable. The $d$-parameter for the compensatory model is on approximately the same scale as $-ab$ for the partially compensatory model.

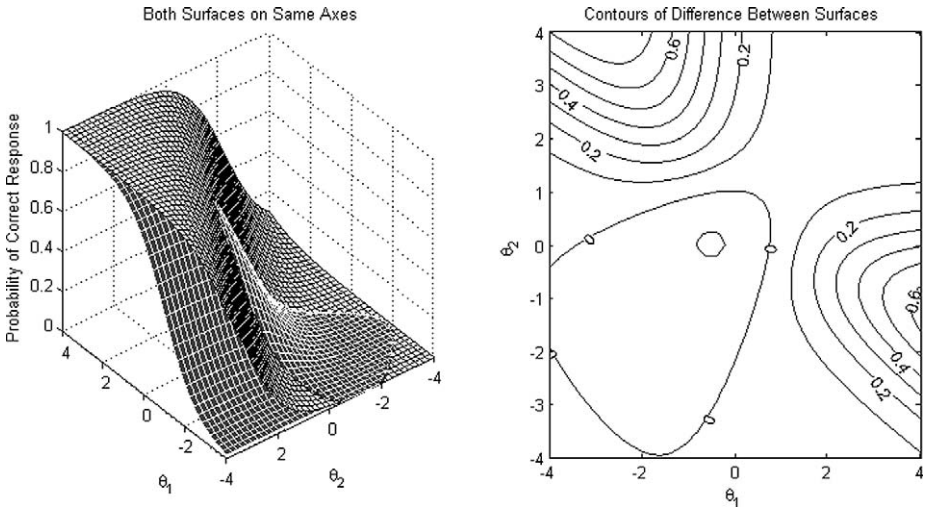


Fig. 4. Compensatory and partially compensatory surfaces on same axes (left panel) and contour plot of the differences between the two surfaces (right panel). Note that the surfaces on the left panel have been rotated to make the intersection of the surfaces more evident.

Figure 4 presents two graphs. The graph on the left shows the surfaces for both models on the same set of axes so that they can be compared. The right pane of the figure shows the contour plot for the difference between the two surfaces. The left panel in Figure 4 shows that the two surfaces intersect in such a way that they are similar near the origin of the $\boldsymbol{\theta}$-space. The surfaces diverge the most when the $\theta_1$ and $\theta_2$ coordinates are opposite in sign and indicate points that are far from the origin. This can be seen clearly in the contour plot in the right pane. In the region around the $(0, 0)$ point in the $\boldsymbol{\theta}$-space, the differences between the two surfaces are close to 0 in probability. However, in the regions around $(-2, 4)$ and $(4, -1)$ the differences in probability are quite large, approximately .7 and .6 respectively.

The form of the contour plot for the difference in the two surfaces shows why the two models yield similar results on classical test statistics (proportion correct of .40 and .38 for the compensatory and partially compensatory, respectively). The large differences in the surfaces occur in regions where the density of the bivariate normal distribution is very small. Further, most skills and knowledge measured by cognitive tests are positively correlated, making persons at locations $(-2, 4)$ even less likely. The similarity of the predicted probabilities for the high density region of the normal distribution and the fact that the partially compensatory model is less mathematically tractable have led to dominance of the compensatory model in the research literature.

## 4. Descriptions of item characteristics

If the MIRT models provide a reasonable representation of the data obtained from the interaction of persons with a test item, then the characteristics of the test item are described by the surface defined by the model and they are summarized by the parameters of the model. For example, if the compensatory model given in (6) represents the relationship between locations in the $\boldsymbol{\theta}$-space and the proportion of correct responses to an item for a sample of examinees, then the characteristics of the item can be summarized by the $\mathbf{a}$-vector, and the scalar $c$ and $d$ parameters.

Unfortunately, the model parameters do not have intuitive meaning. To help in the interpretation of the MIRT models, several statistics have been derived from the parameters to help describe the workings of the models. These statistics indicate the combination of skills best measured by the item, the difficulty of the item, and the usefulness of the item for differentiating between persons at different points in the $\boldsymbol{\theta}$-space.

Generally, the goal of measurement is to differentiate among objects using a particular characteristic of the objects. In this case the objects are people who are located at different places in the $\boldsymbol{\theta}$-space. These locations are represented by vectors of coordinates (e.g., $\boldsymbol{\theta}_j$, $\boldsymbol{\theta}_k$, etc.). Test items are used to gain information about the differences in locations of the persons. Assuming the MIRT model describes the functioning of the test item, the probability computed from the model gives the likelihood of the item score given the persons location in the $\boldsymbol{\theta}$-space. The .5 equiprobable contour is an important feature of an item because the principle of maximum likelihood indicates that if persons respond correctly to an item, they are estimated to be located on the side of the .5 contour that has the higher probability of a correct response. If there is an incorrect response to the item, the person is estimated to be on the side of the .5 contour that has the lower probabilities of correct response. Therefore, the location of the .5 contour is an important characteristic of the item.

Another important characteristic of a test item is how well it differentiates between two persons located at different points in the $\boldsymbol{\theta}$-space. In the context of MIRT models, differentiation is indicated by a difference in the probability of correct response to the item. If the probability of correct response to the item for the locations of two persons is the same, the item provides no information about whether the persons are at the same point or different points. However, if the difference in probability of correct response is large, then it is very likely that the persons are located at different points in the $\boldsymbol{\theta}$-space.

Differences in probability of correct response for an item are largest where the slope of the item response surface is greatest, and when points in the space differ in a way that is perpendicular to the equiprobable contours for the item response surface.

These uses of a test item and the characteristics of the item response surface suggest three descriptive measures for an item: (1) the direction of steepest slope for an item; (2) the distance of the point of steepest slope from the origin of the space; and (3) the magnitude of the slope at the point of steepest slope. These three descriptive measures can easily be derived from the mathematical form of the item response surface.

Measures of direction and distance need to be taken from a particular location. For convenience, the location that is used for these measures is the origin of the space – the $\boldsymbol{\theta}$-vector with all zero elements. Also for mathematical convenience, the development of these measures is done in polar coordinate system so that the direction is easily specified as angles from a particular axis. The compensatory model with the $c$-parameter set to zero will be used for this development.

The elements of the $\boldsymbol{\theta}_j$ vector in (6), $\theta_{j\ell}$, represent the coordinates of a point in an $m$-dimensional space. Let $\theta_j$ represent the distance from the origin to the point specified by the $\boldsymbol{\theta}_j$ vector. The point and distance are shown in Figure 5 for the two-dimensional case. The angle between the line connecting the point to the origin and coordinate axis $\ell$

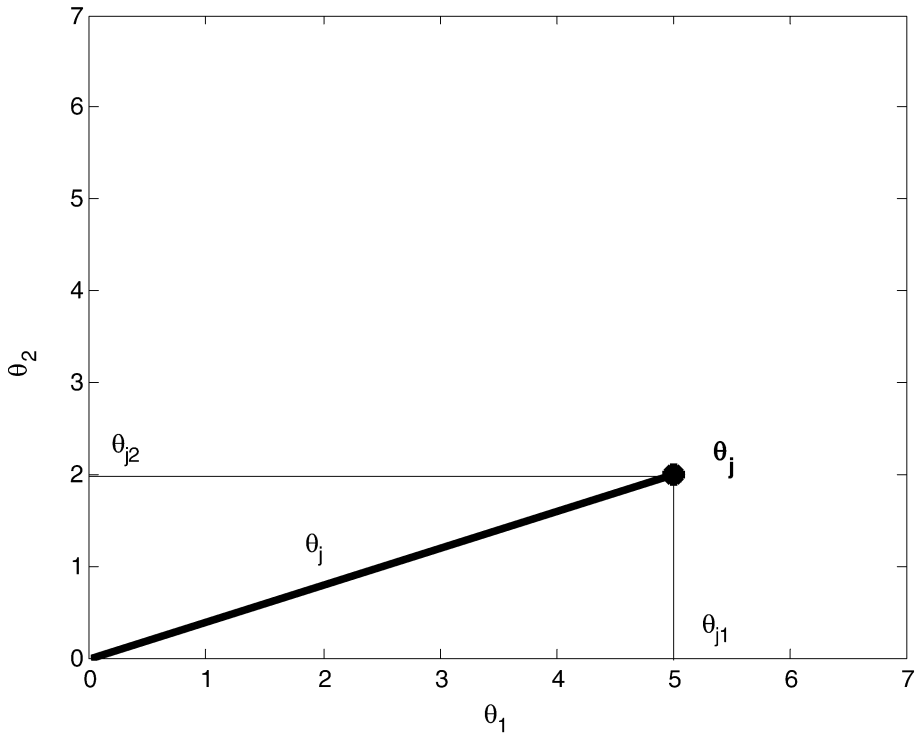


Fig. 5. Polar coordinate representation of person location $\boldsymbol{\theta}_j$.

is given by $\alpha_{j\ell}$. From right triangle trigonometry, the cosine of angle $\alpha_{j\ell} = \theta_{j\ell}/\theta_j$. Solving for $\theta_{j\ell}$ yields $\theta_j \cos\alpha_{j\ell}$. Thus, each element of the coordinate vector can be represented by the distance from the origin times the cosine of the angle between the line to the point and the coordinate axis.

In this two dimensional case, $\alpha_{j2} = 90 - \alpha_{j1}$. More generally, the relationship among the angles between the coordinate axes and the line connecting the origin of the space to the $\boldsymbol{\theta}_j$-point is given by

$$\sum_{\ell=1}^{m} (\cos\alpha_{j\ell})^2 = 1. \tag{12}$$

This relationship is a general property of the relationships of angles with a line represented in an orthogonal coordinate space.

The model in (6) can be reparameterized by substituting a vector composed of elements $[\theta_j \cos\alpha_{j\ell}]$, the scalar value $\theta_j$ times the vector of cosines $\cos\boldsymbol{\alpha}_j$, for the vector of coordinates $\boldsymbol{\theta}_j$. After this reparameterization, the MIRT model takes the following form

$$P(u_i = 1|\theta_j, \boldsymbol{\alpha}_j) = \frac{e^{\mathbf{a}_i'\theta_j \cos\boldsymbol{\alpha}_j + d_i}}{1 + e^{\mathbf{a}_i'\theta_j \cos\boldsymbol{\alpha}_j + d_i}}, \tag{13}$$

where $\theta_j$ is a scalar parameter that gives the distance of person $j$ from the origin of the space and $\boldsymbol{\alpha}_j$ is a vector of angles between the coordinate axes and the line from the origin to the person's location in the space. This formulation of the model specifies the probability of correct response for any angle from the axes given by $\boldsymbol{\alpha}_g$ and any distance from the origin given by $\theta_g$. The slope in any direction specified by $\boldsymbol{\alpha}_g$ can be determined by taking the first derivative of (13) with respect to $\theta_g$,

$$\frac{\partial P(u_i = 1|\theta_g, \boldsymbol{\alpha}_g)}{\partial\theta_g} = P_{ig}(1 - P_{ig})\mathbf{a}_i' \cos\boldsymbol{\alpha}_g, \tag{14}$$

where $P_{ig} = P(u_i = 1|\theta_g, \boldsymbol{\alpha}_g)$, $\theta_g$ is a point on the line from the origin specified by the angles in $\boldsymbol{\alpha}_g$, and $g$ is the index for the specific angle from the origin that is being considered. The subscript for $\theta$ and $\boldsymbol{\alpha}$ has been changed from $j$ in (12) to $g$ in (14) because $\theta_j$ referred to a specific individual and $\theta_g$ refers to a point along a line in direction $\boldsymbol{\alpha}_g$.

The maximum slope in the direction specified by $\boldsymbol{\alpha}_g$ can be determined by taking the second derivative of the item response function and solving for zero. The second derivative is given by

$$\frac{\partial^2 P(u_i = 1|\theta_g, \boldsymbol{\alpha}_g)}{\partial\theta_g^2} = P_{ig}\big(1 - 3P_{ig} + 2P_{ig}^2\big)(\mathbf{a}_i' \cos\boldsymbol{\alpha}_g)^2. \tag{15}$$

There are three solutions when (15) is set to zero, but two of them are for $P_{jg}$ equal to 0 or 1 that lead to values of $\theta_g$ that are negative or positive infinity, respectively. The only finite solution for 0 is $P_{jg} = .5$. This means that the slope is steepest along the .5 contour for all directions from the origin. The expression for the steepest slope in direction $\boldsymbol{\alpha}_g$ is simply ¼ $\mathbf{a}_i' \cos\boldsymbol{\alpha}_g$. The slope of the surface along one of the coordinate

axes is simply ¼ $a_{i\ell}$, where $\ell$ indicates the index for the axis, because the cosine is 1 when the angle is 0°and the cosine is 0 when the angle is 90°.

To determine the direction that gives the steepest slope overall, the expression in (14) for the .5 contour is differentiated with respect to $\boldsymbol{\alpha}_g$ and solved for 0. However, before performing the differentiation, the last element in the vector of cosines, $\cos\alpha_{gm}$, is replaced with the following term

$$\cos\alpha_{gm} = \sqrt{1 - \sum_{h=1}^{m-1} \cos^2\alpha_{gh}}. \tag{16}$$

Adding this constraint insures that the sum of the squared cosines will be equal to 1. This has the effect of making the coordinate axes orthogonal to each other.

Differentiating the expression with the added constraint with respect to each of the angles in the $\boldsymbol{\alpha}_g$-vector, rearranging terms and simplifying results in $m - 1$ expressions of the form

$$a_{ih} - a_{im}\left(\frac{\cos\alpha_{gh}}{\cos\alpha_{gm}}\right), \quad h = 1, \ldots, m-1. \tag{17}$$

Setting these expressions equal to zero and solving for each $\cos\alpha_{gh}$ yields $m$ expressions of the form

$$\cos\alpha_{ih} = \frac{a_{ih}}{\sqrt{\sum_{h=1}^{m} a_{ih}^2}}. \tag{18}$$

Note that in this equation, the subscript on $\alpha$ has been changed from $g$ to $i$ to indicate that this is now a feature of the item. This expression gives the cosine of the angle between each of the axes and the line that connects the origin to the point of steepest slope on the item response surface. The cosines are called the *directions cosines* for that line. This direction indicates the particular direction between two points in the space that the test item provides the most information about differences in the points.

Another useful piece of information is the distance from the origin to the point of steepest slope in the direction defined by (18). That distance can be determined by substituting the vector of direction cosines from (18) into (14) and then solving for $\theta_g$. The result is the distance from the origin to the point of steepest slope. The equation for the distance is:

$$D_i = \frac{-d_i}{\sqrt{\sum_{h=1}^{m} a_{ih}^2}}. \tag{19}$$

$D_i$ is often called the multidimensional difficulty, or MDIFF, of the item because it is on a scale that is analogous to that used in unidimensional IRT models. In those models, the exponent of e is of the form $a(\theta - b) = a\theta - ab$. The $ab$ term is analogous to the $d$-parameter in the MIRT model, with the sign reversed. To get the equivalent of the $b$-parameter, the last term is divided by the discrimination parameter, $a$.

The slope at the point of steepest slope in the direction specified by (18) can be obtained by substituting (18) into (15) with the probability specified as .5. The result is

$$\text{Slope} = \frac{1}{4}\sqrt{\sum_{h=1}^{m} a_{ih}^2}. \tag{20}$$

This expression without the constant ¼, $\sqrt{\sum_{h=1}^{m} a_{ih}^2}$, is referred to as the multidimensional discrimination (MDISC) for the item. Note that MDISC is in the denominator of (18) and (19). The results in (18), (19), and (20) also apply to the model in (6) with $c_i > 0$. Changing that parameter only raises the item response surface, but does not change the location of the equiprobable contour that has the steepest slope. However, the probability value for that contour is $(c_i + 1)/2$ rather than .5.

The values given in (18), (19), and (20) provide a convenient way to represent items for two or three dimensional cases. Each item can be represented as a vector that points in the direction of increasing probability for the item response surface. The base of the vector is on the equiprobable contour that is the line of steepest slope. The base is a distance MDIFF away from the origin. The length of the vector is a scaled value of MDISC. The direction the vector is pointing is given by (18). An example of this vector representation of an item is given on the contour plot of an item response surface in Figure 6. Further discussion of the development of these measures is given in Reckase (1985) and Reckase and Hirsch (1991).

The vector plot for the item with the parameters specified in the legend for Figure 6 has an MDIFF value of .66, and MDISC value of 1.26, and angles with the two axes of 18° and 72°, respectively. The use of item vectors allows many items to be graphically represented at the same time. It is also possible to show items in three dimensions, something that is not possible with response surfaces or contour plots. Figure 7 shows items from a science test using three dimensions. This graph shows that the items measure in different directions as the MDIFF values for the items change. Relatively easy items measure along $\theta_1$, while more difficult items measure along $\theta_2$ and $\theta_3$.

Another way of representing the way that an item functions is its capability to differentiate between two points in the space. This capability is usually labeled as the *information* provided about the location of the examinee by the item. It is the ratio of the rate of change of the item response surface over the variance of error of estimates at that point in the space. The general equation for information (Lord, 1980) for item $i$ is given by

$$I_i(\theta) = \frac{[\frac{\partial P_i(\theta)}{\partial \theta}]^2}{P_i(\theta)[1 - P_i(\theta)]}. \tag{21}$$

When $\boldsymbol{\theta}$ is a vector rather than a scalar, the information at the $\boldsymbol{\theta}$-point depends on the direction taken because the slope of the surface depends on the direction, as was shown above. Therefore, the numerator of (21) is changed to include the directional derivative in angle $\boldsymbol{\alpha}$ rather than the standard derivative:

$$I_\alpha(\theta) = \frac{[\nabla_\alpha P_i(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}, \tag{22}$$

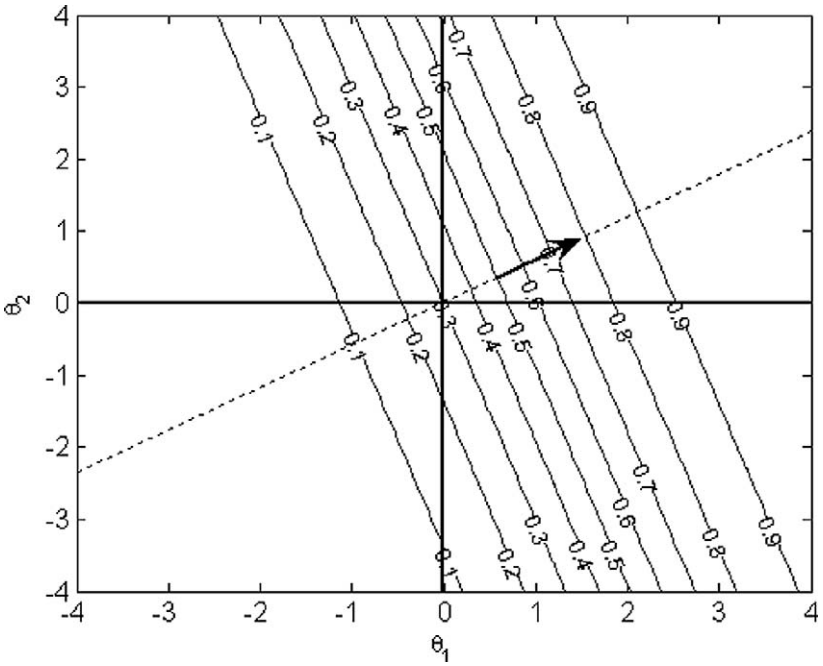


Fig. 6. Item vector representing MDIFF, MDISC, and the direction of best measurement of an item. ($a_1 = 1.2,\ a_2 = .4,\ d = -.84$.)

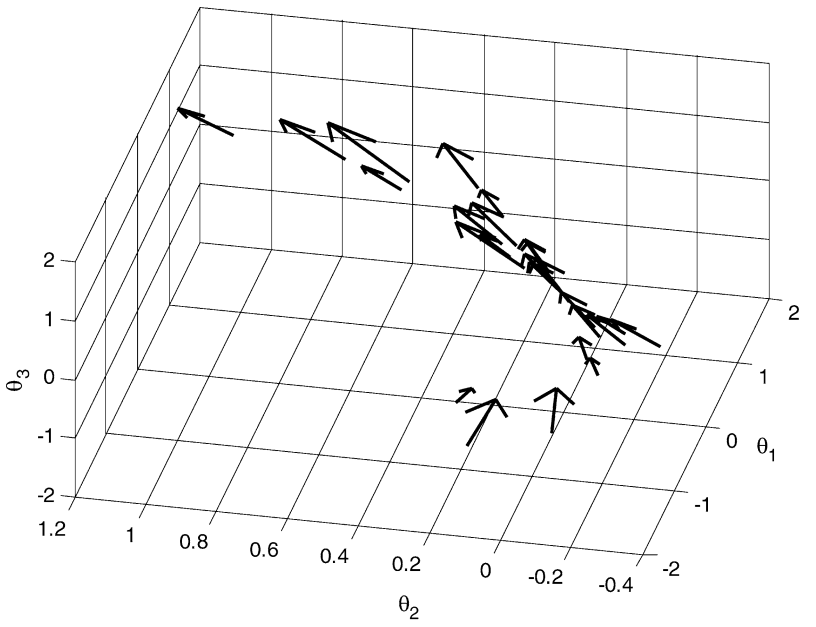


Fig. 7. Vector plot of items in three dimensions.

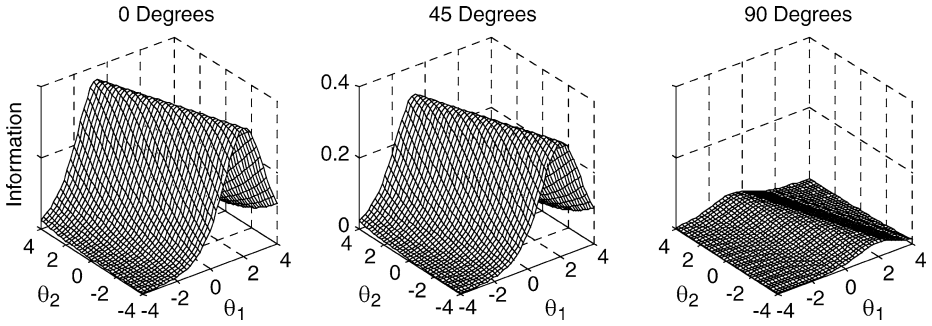


Fig. 8. Information functions for an item in three different directions.

where the directional derivative is defined as follows

$$\nabla_\alpha P_i(\theta) = \frac{\partial P_i(\theta)}{\partial \theta_1}\cos\alpha_1 + \frac{\partial P_i(\theta)}{\partial \theta_2}\cos\alpha_2 + \cdots + \frac{\partial P_i(\theta)}{\partial \theta_m}\cos\alpha_m. \tag{23}$$

Figure 8 shows the information for an item with the same parameters as used in Figure 6 for directions that are 0°, 45°, and 90° from the $\theta_1$ axis. The information functions have the same orientation over the $\boldsymbol{\theta}$-plane, but the heights differ. The orientation has the highest point of the function over the .5-contour of the surface shown in Figure 6. The angle of steepest slope for the surface is 18° which is closest to the 0° direction. Therefore, the information is highest in the left most panel of Figure 8. The 90° direction has substantially lower information than the others because that direction is almost parallel to the equiprobable contours for the surface.

## 5. Descriptions of test characteristics

IRT provides a number of useful ways to describe the characteristics of a test by combining the characteristics of items. For example, because the probability of correct response to an item is the same as the expected score on an item and because the expected value of a sum is the sum of the expected values, a test characteristic surface can be computed by summing the item characteristic surfaces. The item parameters for 20 items that can be modeled using a two-dimensional coordinate system are given in Table 2. For each of the items on this test, the item characteristic surface was computed. These 20 surfaces were summed to form the test characteristic surface. That test characteristic surface is shown in Figure 9.

The height of the surface represents the estimated true score in the number-correct metric for the test. For examinees who are high on both $\theta$-values, the estimated true score is near the maximum summed score on the test of 20. Those low on both $\theta$-values have estimated true scores near zero. The surface does not have linear equal score contours because of the combination of multiple items. Overall, the surface shows how number-correct scores are expected to increase with increases in the coordinate dimensions.

Table 2
Item parameters for a 20 item test

| Item number | $a_1$ | $a_2$ | $d$ | Item number | $a_1$ | $a_2$ | $d$ |
|---|---|---|---|---|---|---|---|
| 1 | 1.81 | .86 | 1.46 | 11 | .24 | 1.14 | −.95 |
| 2 | 1.22 | .02 | .17 | 12 | .51 | 1.21 | −1.00 |
| 3 | 1.57 | .36 | .67 | 13 | .76 | .59 | −.96 |
| 4 | .71 | .53 | .44 | 14 | .01 | 1.94 | −1.92 |
| 5 | .86 | .19 | .10 | 15 | .39 | 1.77 | −1.57 |
| 6 | 1.72 | .18 | .44 | 16 | .76 | .99 | −1.36 |
| 7 | 1.86 | .29 | .38 | 17 | .49 | 1.10 | −.81 |
| 8 | 1.33 | .34 | .69 | 18 | .29 | 1.10 | −.99 |
| 9 | 1.19 | 1.57 | .17 | 19 | .48 | 1.00 | −1.56 |
| 10 | 2.00 | .00 | .38 | 20 | .42 | .75 | −1.61 |

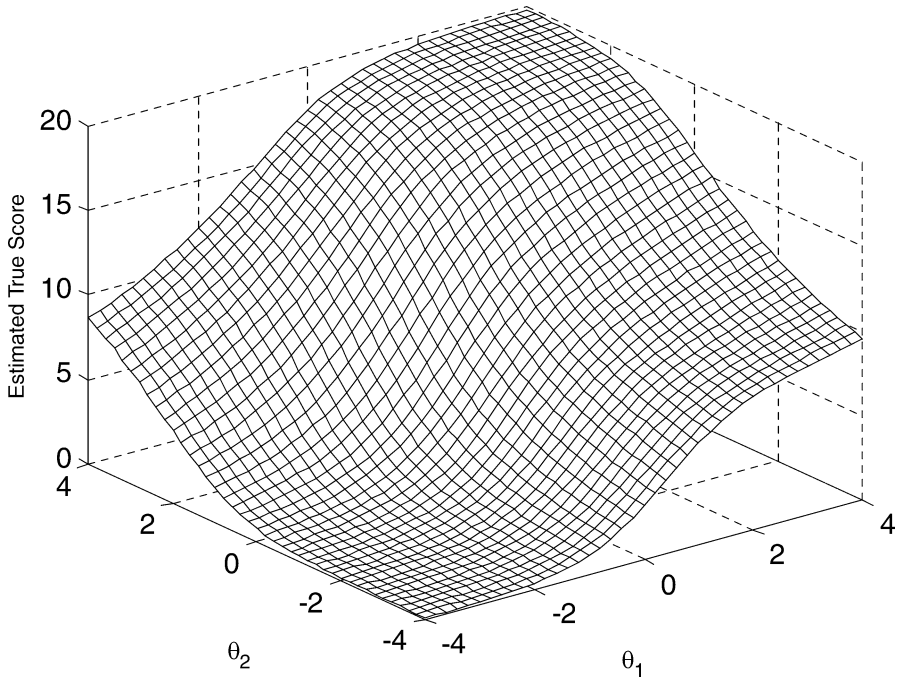


Fig. 9. Test characteristic surface for items in Table 2.

The relationship between item information and the information for the full test is similar to the relationship between the item characteristic surface and the test characteristic surface. However, note that the test information is the sum of the item information values in *a particular direction*. Figure 10 shows the test information for the set of items given in Table 2 in the 0°, 45°, and 90° directions from the $\theta_1$ axis. The different test

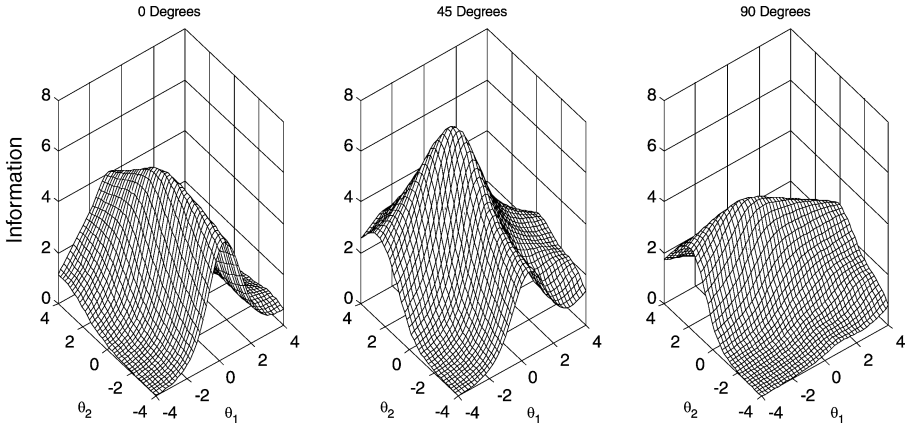


Fig. 10. Test information functions in directions 0°, 45°, and 90° from the $\theta_1$ axis.

information functions show that the form of the information surfaces is not the same for the different directions. The set of items provides more information near the origin of the space in a 45° direction than the other two directions. There are other more subtle differences as well, such as having information higher along the $\theta_2$ dimension when measuring in the 90° direction.

The amount of information provided by a test is related to the direction in the space that the test best differentiates between nearby points. The test shown in Figure 10 is very good at distinguishing between person locations between 0 and 1 on the $\theta_1$ dimension in a direction parallel to that dimension, but is best at distinguishing persons located between 2 and 3 along the $\theta_2$ dimension in the direction parallel to that dimension.

The single direction best measured by this set of items was derived by Wang (1985). She showed that the direction could be determined from the eigenvector of the $\mathbf{a}'\mathbf{a}$ matrix that corresponds to the largest eigenvalue from the eigenvalues/eigenvector decomposition of that matrix. The values of the eigenvector are the equivalent of the elements of the **a**-parameter vector for an item and can be used to find the direction of measurement for a unidimensional calibration of the test. Wang (1985) labeled the line defined by this eigenvector the "reference composite" for the test. In the case of the set of items in Table 2, the reference composite is a line through the origin that has a 37° angle with the $\theta_1$ axis. The reference composite is shown by the long dashed vector in Figure 11.

The concept of a reference composite can be used with sets of items as well as the full test. For example, if the first ten items in the test measure one content area and the second ten items measure a different content area, reference composites can be determined for each set of items and the directions of the unidimensional scale for those two item sets can be determined. Those two reference composites are shown in Figure 12 by the vectors drawn with thick lines. Note that the item vectors for the first ten items, the easier items, are much more tightly grouped around their reference composite than those for the second set of ten items, the harder items. When the item vectors for a set of items are all pointing in the same direction, that set of items can be

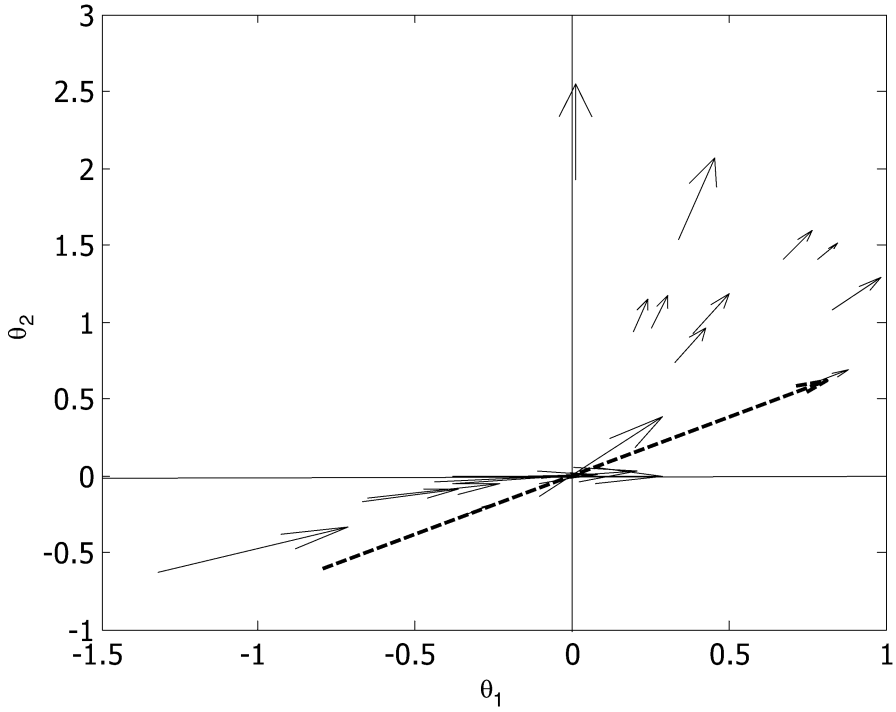


Fig. 11. Item vectors and reference composite vector for the items in Table 2.

represented using a unidimensional model. The first ten items approximately meet that condition.

The second ten items are much more widely dispersed in their directions of best measurement. Although the reference composite for those items indicates the direction of measurement for a unidimensional analysis of that set of items, the null hypothesis that the item responses from this set of items were generated by a unidimensional IRT model would likely be rejected.[2]

The reference composites in Figure 12 show an important relationship between dimensions of sensitivity for test items and the coordinate axes that define the multidimensional space. The coordinate axes for this analysis are a convenient orthogonal frame of reference for representing the item vectors. The reference composites indicate the directions of best measurement for the sets of test items. The dimensions of measurement of sets of items generally do not correspond to the coordinate axes and they are usually not orthogonal. The composites measured by sets of items are often fairly highly correlated. When simulation studies are used to generate data of this type, it is often suggested that the coordinate axes need to be non-orthogonal (i.e., correlated) to

[2] Statistical procedures such as DIMTEST (Stout et al., 1996) are available to test the null hypothesis that the data were generated using a model with a single person parameter. Such procedures were not used for this example because the analysis of a ten item set would not give much power to test the null hypothesis.

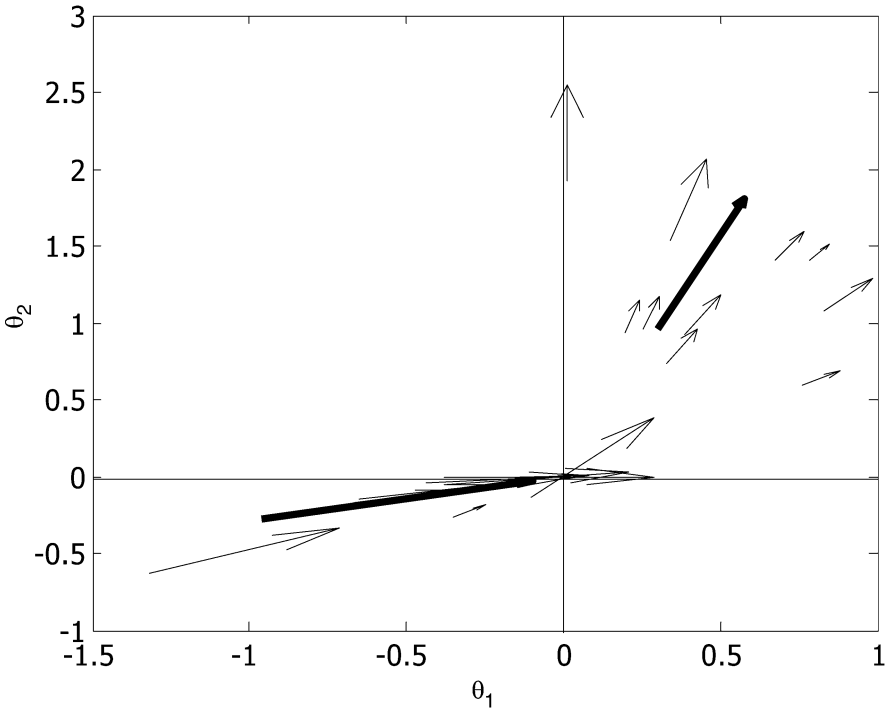


Fig. 12. Reference composites for the first ten and second ten items for the items in Table 2.

represent the correlation between the traits being measured. That is not the case. The dimensions of sensitivity for sets of items in the test can be correlated while the coordinate axes are orthogonal. These two conditions can easily hold at the same time. A common example is a scatter plot between height and weight. The coordinate axes for height and weight are drawn orthogonal to each other. The correlation between height and weight is shown by the form of the scatter plot. The angle between the coordinate axes is not used to represent the correlation between the variables that are represented in the scatter plot.

## 6. The estimation of model parameters

The most intuitively appealing and creative models are not very useful unless the parameters in the model can be estimated fairly accurately. Estimation procedures for compensatory MIRT models have been available for some time based on the normal ogive version of the compensatory model (see Eq. (9)). Because of the similarity of the logistic model to the normal ogive model, the results of the normal ogive-based programs are often used as estimates of the parameters of the logistic models as well. This is because the increase in estimation error caused by model misspecification is believed to be small relative to other sources of estimation error.

Initial estimation procedures for MIRT were implemented in two computer programs, NOHARM and TESTFACT. NOHARM (Fraser, 1988) is based on the work by McDonald (1967, 1981) to estimate the parameters of the normal ogive model. The method programmed in NOHARM uses a four term polynomial approximation to the normal ogive model. The approximation is used to predict the joint frequency of item scores on pairs of items using least squares polynomial regression. The item response data from the test is tallied to provide the data in the cells of a two-by-two table of frequencies of response combinations for each pair of items (Table 3). In the table, $n_{ijuv}$ gives the frequency of the response pair $u, v$ where $u$ and $v$ can take on the values of 0 and 1 for items $i$ and $j$. The program uses ordinary least squares to estimate the parameters $\mathbf{a}$ and $d$ that minimizes

$$\sum_{i,j,u,v} (n_{ijuv} - \hat{n}_{ijuv})^2, \tag{24}$$

where $\hat{n}_{ijuv}$ is the value predicted from the polynomial equation. McDonald (1982) and others have found this procedure to be computationally fast and robust. It has been used to estimate parameters for up to 50 dimensions. However, this approach is only used to estimate the item parameters. The $\boldsymbol{\theta}$-vectors are not estimated, but are assumed to be distributed as multivariate normal with a specified variance covariance matrix. The program allows $c$- and $\mathbf{a}$-parameters to be fixed to specified values to support confirmatory analyses.

The other commonly used program for the estimation of parameters is TESTFACT (du Toit, 2003, Chapter 5). This program uses methodology developed by Bock et al. (1988). Like the methodology used in NOHARM, this program assumes the normal ogive formulation of the MIRT model. However, rather than predicting the elements of the $2 \times 2$ table of frequencies, TESTFACT determines the parameter estimates that maximize the probabilities of the observed frequencies of the full response pattern on all of the items for the examinees. Because the full response pattern is used, this methodology has been labeled "full information factor analysis." The maximization is done assuming a multivariate normal distribution of proficiencies. The basic equation for the estimation is given by

$$P(\mathbf{U} = \mathbf{U}_s) = \int_{\theta} L_s(\theta) g(\theta)\, \mathrm{d}\theta, \tag{25}$$

where $\mathbf{U}_s$ is a particular binary response pattern for a set of items, $L_s(\boldsymbol{\theta})$ is the likelihood of response string at a particular point in the $\theta$-space, and $g(\boldsymbol{\theta})$ is the multivariate

Table 3
Frequency of item score pairs for two items

| | | Item $j$ | |
|---|---|---|---|
| | | 0 | 1 |
| Item $i$ | 0 | $n_{ij00}$ | $n_{ij01}$ |
| | 1 | $n_{ij10}$ | $n_{ij11}$ |

normal density for the proficiency values. This integral is approximated using $m$-fold Gauss–Hermite quadrature. Because estimation is done assuming a known marginal distribution of proficiencies, this procedure is called marginal maximum likelihood (MML) estimation. The actual implementation of the estimation procedure uses the EM algorithm (Dempster et al., 1977). The details of the process are given in Bock et al. (1988).

Recently, there has been substantial interest in using Markov chain Monte Carlo (MCMC) methods for the estimation of the parameters of MIRT models. At the time this chapter was written (i.e., 2005), MCMC for MIRT estimation were relatively new, so the properties of those estimates were not well documented. Bolt and Lall (2003) used MCMC methods based on the Metropolis–Hastings algorithm to estimate the parameters of both the compensatory and partially compensatory logistic models presented in Eqs. (8) and (10). Generally, the procedure was able to recover the model parameters in a simulation study, although the recovery was better for the compensatory model than the partially compensatory model. A negative aspect of these methods was that estimation took approximately 40 hours for 31 items and 3000 examinees. This should not be taken as too serious a problem because computers are already substantially faster than those used in 2002. Also, a number of researchers are currently working on the use of this methodology. By the time this chapter is published, there likely will be a number of additional journal articles that update the progress of this methodology. The reader is encouraged to read current publications on these methodologies to get up to date information on their usefulness for MIRT.

## 7. Applications

There have been relatively few applications of MIRT procedures to practical testing problems, at least as compared to unidimensional IRT procedures. For example, as of 2005, no large scale testing program uses MIRT for test equating while several use unidimensional IRT procedures for that purpose. There is a growing body of research on such applications, but most of the research is on evaluating new procedures or checking methodology through simulation. There are also some confirmatory analyses related to the sensitivity of items to differences along specified dimensions. No doubt, the number of operational applications will increase in the coming years. A few of the areas where applications are actively being developed will be highlighted here. As with any dynamic research area, the current research literature should be checked for the newest developments in this field.

The three areas that have received the most attention to date are (1) the detailed analysis of test content, (2) equating/linking of test calibrations to put them into the same solution space, and (3) computerized adaptive testing. Each of these applications is described in turn.

The MIRT estimation of item parameters and the calculation of the direction of item vectors give a convenient way to study the structure of test content. Item vectors that point in the same direction indicate that the items measure the same composite of the coordinate dimensions. Reckase et al. (1988) showed that when sets of items have item vectors that point in the same direction they can be modeled using a unidimensional

IRT model. Further, Reckase and Stout (1995) determined the general conditions when sets of items that require multiple skills and knowledge to specify a correct response can yield item response data that are well fit by unidimensional models. These findings suggest that if items are clustered into sets with minimal angles between item pairs, the resulting clusters are sets of items that are sensitive to differences on the same composite of skills and knowledge. Miller and Hirsch (1992) demonstrated the process of identifying sets of items sensitive to the same set of skills and knowledge using cluster analysis procedures.

Miller and Hirsch (1992) used NOHARM to estimate the item parameters and then computed the cosine of the angle between two item vectors using the equation

$$\cos\alpha_{ij} = \frac{\mathbf{a}_i'\mathbf{a}_j}{\sqrt{\sum_{k=1}^{m} a_{ik}^2 \sum_{k=1}^{m} a_{jk}^2}} = \sum_{k=1}^{m} \cos\alpha_{ik}\cos\alpha_{jk}. \tag{26}$$

The matrix of cosines was then converted to the angles between the item directions. The angles were used as the dissimilarity measure for a hierarchical cluster analysis using Ward's method (Ward, 1963). The number of clusters is difficult to determine with precision, but considering a combination of the substantive meaning of clusters and the standard errors of the parameter estimates can give reasonable guidance to the number of clusters.

The parameters given in Table 2 are used to demonstrate this procedure. The angles between the items were computed using Eq. (26) and the results were clustered using Ward's method. The results are shown in the dendrogram in Figure 13. In this diagram, the item numbers that are connected together near the horizontal axis have small angles between them. The horizontal bar connecting the items is at the level showing the angular distance on the vertical axis. The dendrogram shows that the 20 items given in Table 2 are generally of two different types. These two types generally are consistent with the two reference composites in Figure 12.

The method of clustering using the angles between items has been used on a wide variety of test data and it has been shown to yield a very fine grained summary of item content (Reckase et al., 1997). The results in Miller and Hirsch (1992) show that interpretable clusters can be identified even when there is a relatively small angle between clusters. Other analyses at ACT, Inc. (reported in Reckase, 1997c) were able to detect item content clusters based on as few as four items. Traditional factor analysis approaches would not detect these item clusters because they did not account for very much variance out of a 60 item test. However, the clustering based on angular distances was quite stable, being replicated on a number of parallel forms of the test.

A second area of the application of MIRT that has received substantial research effort is the linking of MIRT calibrations. The linking of calibrations is important for equating and vertical scaling of test forms and the development of item pools for computerized adaptive testing. The methods for computerized adaptive testing with MIRT also depend on the linking of calibrations for the development of item pools of sufficient size to support sound implementation.

When items are calibrated using any one of the calibration approaches that are available, certain aspects of the calibration are not statistically determined from the data

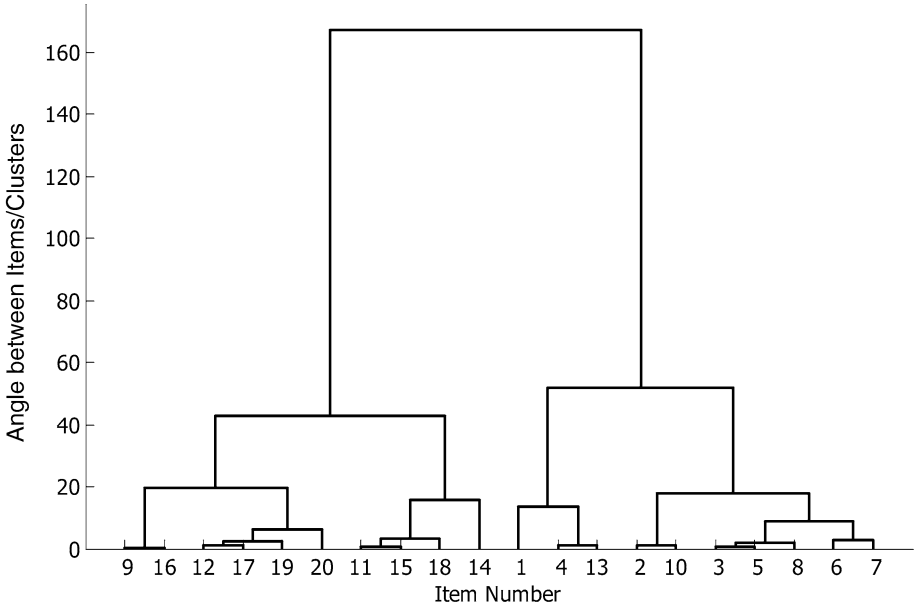


Fig. 13. Dendrogram of item clusters based on item parameters in Table 2.

and these indeterminacies are usually addressed by fixing/constraining certain features of the parameter estimates. Li and Lissitz (2000) lay out these issues and indicate that there are three types of indeterminacy in the solutions. First, the origin of the coordinate system is not defined. This is usually addressed by setting the mean vector of the estimated proficiencies to a vector of zeros. Second, the units along each of the coordinate axes are not determined. This is usually addressed by setting the unit as the standard deviation of the observed proficiencies. These first two types of indeterminacy also exist for unidimensional IRT models so a variety of procedures have been developed to address those issues.

The third type of indeterminacy is the orientation of the axes of the coordinate system in the space defined by the proficiency estimates. NOHARM addresses this type of indeterminacy by fixing the coordinate axes to run parallel to the item vectors for pre-selected items. The solution can later be changed using a variety of rotation procedures to set the orientation of the coordinate axes according to other criteria. It is clear, however, that using a different set of items to anchor the coordinate axes will give a solution that appears different, even though it may be possible to make two solutions look the same after rotation to a common criterion. Other estimation methods use different constraints to set the orientation of the coordinate axes. This means that analyzing the same set of data with two different programs will likely give different "looking" results, different parameter estimates, even though they might be very nearly the same after an appropriate rotation.

Because of these indeterminacies and the different ways that they are addressed by different estimation programs, the estimates of parameters obtained from the programs

are to some extent sample specific. That is, if a set of items are calibrated using samples that have different mean vectors and different variance/covariance matrices, the procedures for resolving indeterminacies will cause the parameter estimates to be different when, in fact, they are simple transformations of each other. The goal of MIRT linking/equating procedures is to determine the transformations that will convert a set of item parameter estimates from one sample to estimates of the same parameters from another sample. The transformation determined for items can be used to determine the comparable transformation for abilities. That transformation will keep the person locations in the same place relative to each other, but convert the locations to a different coordinate system.

To make the discussion of the indeterminacy of the solutions concrete, the parameters in Table 2 were used to generate item response data according to the compensatory MIRT model based on two different $\boldsymbol{\theta}$-distributions. In both cases, 2000 simulated student response vectors were generated based on each assumed distribution of $\boldsymbol{\theta}$. The first distribution has a mean vector of $(0, 0)$ and the identity matrix for the variance–covariance matrix. The second distribution has a mean vector of $(.5, -.75)$ and variance–covariance matrix $\left[\begin{smallmatrix} 2.25 & .45 \\ .45 & .5625 \end{smallmatrix}\right]$. The first distribution is consistent with that specified in the TESTFACT estimation program. However, that distribution has a zero intercorrelations between the $\theta_1$ and $\theta_2$ coordinates of the examinees' locations in the space. A zero correlation between coordinates is unlikely in practice because most cognitive proficiencies are related to each other for typical testing populations.

The second distribution specifies a correlation between the $\theta$-coordinates of .4 and a standard deviation for the first set of coordinates that is twice that for the second set of coordinates. Because of the close match to the assumptions built into TESTFACT, the analysis of data generated from the item parameters and the first proficiency distribution should result in parameter estimates that are similar to the generating parameters. The analysis of the data generated from the second distribution should give results that are quite different because the TESTFACT will set the estimated mean vector to a default value of $(0, 0)$ and the default variance covariance matrix to the identity matrix. Because the person parameters have are estimated on a rotated and translated set of coordinates, the item parameters are estimated with a compensating translation and non-orthogonal rotation. For the second set of generated data, the item parameter estimates will not appear similar to the generating parameters. However, the linking of the calibrations to a common set of coordinate axes should show that these two sets of parameter estimates are estimates of the same true parameters.

Table 4 gives the generating parameters and the two sets of estimated parameters side by side so they can be conveniently compared. The estimates from TESTFACT have been multiplied by 1.7 to put them on the logistic model metric because TESTFACT assumes the normal ogive model when estimating parameters. The values are those obtained using the Varimax rotation option from the TESTFACT program. The values labeled Est 1 in the table come from the estimation of the generated data assuming the standard normal distribution of $\boldsymbol{\theta}$s with the identity matrix for the variance covariance matrix. The values labeled Est 2 are for the estimates based on the distribution described above. It is clear from a comparison of the parameter estimates that the Est 1 values are fairly close to the values used to generate the data. The Est 2 values are quite different,

Table 4
Item parameter estimate comparison

| Item number | $a_1$-parameter | | | $a_2$-parameter | | | $d$-parameter | | |
|---|---|---|---|---|---|---|---|---|---|
| | True | Est 1 | Est 2 | True | Est 1 | Est 2 | True | Est 1 | Est 2 |
| 1 | 1.81 | 1.51 | 2.28 | 0.86 | 0.86 | 1.51 | 1.46 | 1.27 | 1.81 |
| 2 | 1.22 | 1.20 | 1.36 | 0.02 | 0.07 | 0.74 | 0.17 | 0.17 | 0.77 |
| 3 | 1.57 | 1.37 | 1.98 | 0.36 | 0.46 | 1.04 | 0.67 | 0.55 | 1.19 |
| 4 | 0.71 | 0.60 | 0.98 | 0.53 | 0.63 | 0.82 | 0.44 | 0.44 | 0.39 |
| 5 | 0.86 | 0.85 | 1.01 | 0.19 | 0.22 | 0.56 | 0.10 | 0.17 | 0.44 |
| 6 | 1.72 | 1.67 | 2.28 | 0.18 | 0.34 | 1.11 | 0.44 | 0.43 | 1.34 |
| 7 | 1.86 | 1.66 | 2.33 | 0.29 | 0.36 | 1.18 | 0.38 | 0.43 | 1.38 |
| 8 | 1.33 | 1.19 | 1.81 | 0.34 | 0.43 | 1.00 | 0.69 | 0.75 | 1.20 |
| 9 | 1.19 | 1.04 | 1.56 | 1.57 | 1.53 | 1.97 | 0.17 | 0.08 | −0.33 |
| 10 | 2.00 | 1.98 | 2.93 | 0.00 | 0.15 | 0.91 | 0.38 | 0.50 | 1.77 |
| 11 | 0.24 | 0.18 | 0.29 | 1.14 | 1.07 | 1.07 | −0.95 | −0.92 | −1.64 |
| 12 | 0.51 | 0.36 | 0.69 | 1.21 | 1.39 | 1.33 | −1.00 | −1.06 | −1.64 |
| 13 | 0.76 | 0.64 | 0.99 | 0.59 | 0.67 | 0.79 | −0.96 | −0.99 | −0.96 |
| 14 | 0.01 | −0.07 | 0.04 | 1.94 | 1.64 | 1.31 | −1.92 | −1.81 | −3.13 |
| 15 | 0.39 | 0.25 | 0.43 | 1.77 | 1.55 | 1.53 | −1.57 | −1.45 | −2.70 |
| 16 | 0.76 | 0.66 | 0.92 | 0.99 | 1.05 | 1.14 | −1.36 | −1.54 | −1.63 |
| 17 | 0.49 | 0.48 | 0.61 | 1.10 | 0.94 | 1.19 | −0.81 | −0.76 | −1.36 |
| 18 | 0.29 | 0.15 | 0.48 | 1.10 | 1.08 | 1.00 | −0.99 | −1.01 | −1.66 |
| 19 | 0.48 | 0.49 | 0.54 | 1.00 | 0.92 | 0.99 | −1.56 | −1.68 | −1.97 |
| 20 | 0.42 | 0.37 | 0.56 | 0.75 | 0.68 | 0.67 | −1.61 | −1.66 | −2.00 |
| Mean | 0.93 | 0.83 | 1.20 | 0.80 | 0.80 | 1.09 | −0.39 | −0.40 | −0.44 |
| Standard deviation | 0.62 | 0.59 | 0.82 | 0.57 | 0.48 | 0.33 | 0.98 | 0.97 | 1.60 |

with larger mean $a$-parameters and larger standard deviations for $d$-parameters. This is not a surprising result because the second data set did not have characteristics that matched the default origin and unit of measurement used in the TESTFACT program. Because TESTFACT constrains the $\boldsymbol{\theta}$-estimates to be uncorrelated, the correlation between the $\boldsymbol{\theta}$-vectors used to generate the data must be accounted for in other ways. The parameter estimates show that adjustment needed to account for the correlation is an increase in the $a$-parameter estimates and an increase in variance for the $d$-parameter estimates. However, the parameter estimates from both data sets represent the same correlational structure among the items. To show that, the parameter estimates need to be rotated and translated to be put on the same coordinate system as the generating values.

Approaches to determining transformations for putting all of the item parameters on the same coordinate system have been developed by Li and Lissitz (2000) and Min (2003). Both use Procrustes methods to determine a rotation matrix that rotates the coordinate axes so that the item vectors point in the same directions in the coordinate space. They then solve for translation constants that minimize the difference between the target and estimated $d$-parameters. The approach presented here is an extension of

those methods based on work by Martineau (2004) and Reckase and Martineau (2004). This approach is more general than previously published methods.

The rotation matrix for the coordinate axes is computed using the oblique Procrustes method as specified by Mulaik (1972, p. 297). That method determines the rotation matrix from the following equation

$$\mathbf{Rot} = (\mathbf{a}_a'\mathbf{a}_a)^{-1}\mathbf{a}_a'\mathbf{a}_b, \tag{27}$$

where $\mathbf{a}_b$ is the $n \times m$ matrix of base form discrimination parameters that are the target for the transformation, $\mathbf{a}_a$ is the $n \times m$ matrix of discrimination parameters for the same items on the alternate form, and **Rot** is the $m \times m$ rotation matrix for the discrimination parameters.

For the examples given in Table 4, using the generating **a**-matrix as the target, the rotation matrix for Est 1 is $\left[\begin{smallmatrix} 1.06 & -.09 \\ .07 & 1.09 \end{smallmatrix}\right]$ and that for Est 2 is $\left[\begin{smallmatrix} .75 & -.49 \\ .03 & 1.26 \end{smallmatrix}\right]$. The rotation matrix for Est 1 is very close to an identity matrix. This is because the data were generated according to the same model as is assumed by the TESTFACT program. The rotation matrix for Est 2 is quite different than that for Est 1 because it must account for the differences in variances of the generating $\theta$ parameters and the covariance between those parameters. The difference in the diagonal values account for the different variances for the generating $\theta$s and non-zero off-diagonal value accounts for the covariance between the $\theta$s. The **a**-matrix on the coordinate system for the target set of parameters is given by

$$\widehat{\mathbf{a}}_b = \mathbf{a}_a\mathbf{Rot}, \tag{28}$$

where $\widehat{\mathbf{a}}_b$ is the estimate of the $a$-parameters on the base form metric. The rotations for Est 1 and Est 2 were applied to the estimated **a**-matrices and the results are presented in Table 5 with the generating parameters. It is clear that the parameters after rotation are much closer than those in Table 4. The means and standard deviations of the $a$-parameters for the two dimensions after rotation are very similar to statistics for the parameters used to generate the data.

The transformation of the $d$-parameters from the alternate form to the metric of the base form is given by

$$\mathbf{Trans} = \mathbf{a}_a(\mathbf{a}_a'\mathbf{a}_a)^{-1}\mathbf{a}_a'(\mathbf{d}_b - \mathbf{d}_a), \tag{29}$$

where $\mathbf{d}_b$ is the $n \times 1$ vector of $d$-parameters for the base form, $\mathbf{d}_a$ is the $n \times 1$ vector of $d$-parameters for the alternate form, and **Trans** is the $n \times 1$ transformation vector for the $d$-parameters. The other symbols have been defined above. The estimate of the $d$-parameters on the base form from those on the alternate form is given by

$$\widehat{\mathbf{d}}_b = \mathbf{d}_a + \mathbf{Trans}. \tag{30}$$

The correlations between the parameter estimates and the true values are also very high. For example, the correlation between the generating $d$-parameter and Est 2 is .994.

The transformation for the estimates of $\boldsymbol{\theta}$ from the alternate form to the base form metric is given by

$$\widehat{\boldsymbol{\theta}}_b' = \mathbf{Rot}^{-1}\boldsymbol{\theta}_a' + (\mathbf{a}_b'\mathbf{a}_b)^{-1}\mathbf{a}_b'(\mathbf{d}_a - \mathbf{d}_b), \tag{31}$$

Table 5
Item parameter estimate comparison after rotation

| Item number | $a_1$-parameter | | | $a_2$-parameter | | | $d$-parameter | | |
|---|---|---|---|---|---|---|---|---|---|
| | True | Est 1 | Est 2 | True | Est 1 | Est 2 | True | Est 1 | Est 2 |
| 1 | 1.81 | 1.66 | 1.74 | 0.86 | 0.80 | 0.79 | 1.46 | 1.29 | 1.37 |
| 2 | 1.22 | 1.28 | 1.04 | 0.02 | −0.03 | 0.27 | 0.17 | 0.18 | 0.35 |
| 3 | 1.57 | 1.48 | 1.50 | 0.36 | 0.37 | 0.35 | 0.67 | 0.57 | 0.55 |
| 4 | 0.71 | 0.68 | 0.75 | 0.53 | 0.64 | 0.56 | 0.44 | 0.45 | 0.35 |
| 5 | 0.86 | 0.92 | 0.77 | 0.19 | 0.16 | 0.21 | 0.10 | 0.18 | 0.14 |
| 6 | 1.72 | 1.79 | 1.73 | 0.18 | 0.22 | 0.29 | 0.44 | 0.45 | 0.53 |
| 7 | 1.86 | 1.78 | 1.77 | 0.29 | 0.25 | 0.35 | 0.38 | 0.45 | 0.57 |
| 8 | 1.33 | 1.29 | 1.38 | 0.34 | 0.37 | 0.38 | 0.69 | 0.76 | 0.67 |
| 9 | 1.19 | 1.21 | 1.22 | 1.57 | 1.57 | 1.72 | 0.17 | 0.11 | 0.24 |
| 10 | 2.00 | 2.11 | 2.21 | 0.00 | −0.01 | −0.28 | 0.38 | 0.52 | 0.25 |
| 11 | 0.24 | 0.27 | 0.24 | 1.14 | 1.15 | 1.20 | −0.95 | −0.90 | −0.88 |
| 12 | 0.51 | 0.48 | 0.55 | 1.21 | 1.48 | 1.34 | −1.00 | −1.04 | −0.97 |
| 13 | 0.76 | 0.72 | 0.76 | 0.59 | 0.67 | 0.51 | −0.96 | −0.98 | −1.03 |
| 14 | 0.01 | 0.04 | 0.07 | 1.94 | 1.79 | 1.64 | −1.92 | −1.79 | −1.95 |
| 15 | 0.39 | 0.38 | 0.37 | 1.77 | 1.67 | 1.72 | −1.57 | −1.43 | −1.63 |
| 16 | 0.76 | 0.77 | 0.72 | 0.99 | 1.08 | 0.99 | −1.36 | −1.52 | −1.31 |
| 17 | 0.49 | 0.58 | 0.49 | 1.10 | 0.98 | 1.20 | −0.81 | −0.75 | −0.76 |
| 18 | 0.29 | 0.24 | 0.38 | 1.10 | 1.16 | 1.02 | −0.99 | −1.00 | −1.12 |
| 19 | 0.48 | 0.58 | 0.43 | 1.00 | 0.95 | 0.90 | −1.56 | −1.66 | −1.49 |
| 20 | 0.42 | 0.44 | 0.44 | 0.75 | 0.71 | 0.58 | −1.61 | −1.65 | −1.82 |
| Mean | 0.93 | 0.94 | 0.93 | 0.80 | 0.80 | 0.79 | −0.39 | −.39 | −.40 |
| Standard deviation | 0.62 | 0.60 | 0.61 | 0.57 | 0.56 | .56 | 0.98 | .97 | 1.00 |

where $\boldsymbol{\theta}_a$ is a $1 \times m$ vector of estimates from the alternate form calibration, and $\widehat{\theta}'_b$ is the $1 \times m$ parameter estimate vector after transformation to the coordinate system from the base form.

The transformations presented here are useful for linking calibrations of distinct sets of items to form an item pool for computerized adaptive testing. They can also be used for multidimensional generalizations of horizontal and vertical equating (see Reckase and Martineau, 2004, for an example). In all of these cases, the goals of the analysis can be accomplished by including a common set of items, often called an anchor set, in the tests administered to different groups of examinees. When the groups are at the same educational or grade level, the result is horizontal equating. When the groups differ in educational or grade level, and the non-common items differ in difficulty, the analysis is called vertical scaling. In any case, the set of common items must be sensitive to all of the proficiency dimensions that are of interest. In many cases, the total number of proficiency dimensions is not known. In that case, Reckase and Hirsch (1991) showed that overestimating the number of dimensions was less of a problem than underestimating the number of dimensions. In the latter case, dimensions were projected on top of each

other, losing important information. Therefore, calibrations should use enough dimensions to capture all of the meaningful capabilities in all of the tests being administered. This is especially important for vertical scaling because tests of differing levels of difficulty may assess different skills and knowledge. The number of dimensions for analysis should be the union of the dimensions in all of the tests being calibrated.

One of the more interesting applications of MIRT is as the basis for computerized adaptive testing (CAT). CAT is a testing methodology that has strong connections to the sequential design of experiments. In this case, the experiment is the administration of one test item and the result is the response to that test item. The design of a CAT addresses the problem of how to design the following experiments after obtaining the results of prior experiments. The design of subsequent experiments requires the selection of the next test item with the goal of giving results that can be used to determine the location of the examinee in the latent space.

The design of the $n + 1$ experiments is essentially the selection of the item that has characteristics that will reduce the error in estimation of the examinee's location in the latent space by the greatest amount. In order to select the item, that is, to design the next experiment, information must be available about the characteristics of the items that could be used for the experiment. This information is obtained through the calibration of test items using the procedures described above. The calibration results in a data file of prospective test items along with estimates of the parameters that describe the functioning of the item. This data file is typically called an item pool or item bank.

Numerous criteria have been suggested for selecting the next item when CAT is implemented using MIRT models. The most frequently used method for unidimensional versions of CAT is to select the item that provides the most information at the current estimate of $\boldsymbol{\theta}$. In the multidimensional case, the information provided by the item is given by Eq. (22). That formulation of information requires the specification of a direction that is the focus of the assessment. That direction is not easy to specify for a MIRT CAT so other criteria for selecting items have been considered. The method for item selection described here is directly related to the procedure for estimating the $\boldsymbol{\theta}$-vector so that methodology will be described first.

Maximum likelihood estimation is frequently used to obtain the estimate of $\theta$ when unidimensional models are used as the basis for CAT. In the unidimensional case, at least one correct and incorrect response is needed before a finite maximum likelihood estimate can be obtained. Before at least one of each response type is obtained, heuristics are used to update the $\theta$-estimates. For MIRT based CAT, the number of types of response strings that do not lead to finite $\boldsymbol{\theta}$-vector estimates is greater than for the unidimensional case, so more elaborate heuristics are required. Acceptable heuristics have not yet been developed. Segall (2000) recommends using Bayesian estimation procedures as an alternative to maximum information item selection and maximum likelihood $\boldsymbol{\theta}$ estimation because it resolves the problem of infinite estimates and it makes the estimation more efficient by using prior information about the distribution of $\boldsymbol{\theta}$. The recommended methods are briefly described below.

The prior distribution for the $\boldsymbol{\theta}$-estimates is assumed to be multivariate normal with a specified mean vector and variance–covariance matrix:

$$f(\theta) = (2\pi)^{-\frac{m}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}-\boldsymbol{\mu})}, \tag{32}$$

where $\boldsymbol{\Sigma}$ is the variance/covariance matrix for the $\boldsymbol{\theta}$-values, $\boldsymbol{\mu}$ is a vector of means for the $\boldsymbol{\theta}$-values.

The other variables have been defined earlier. When Segall specifies the prior it is common to use a vector of zeros as the mean vector and to set the variances to 1.0. The covariances are set to values that are expected for the dimensions in the test data. Because most cognitive dimensions are correlated in an unselected population, using a value of .5 or .6 as prior correlations for the variables may be reasonable.

The posterior distribution of the $\boldsymbol{\theta}$-estimate is determined by multiplying the prior distribution by the likelihood of the observed response vector given $\boldsymbol{\theta}$ and then standardizing the distribution to have a total area of 1.0. The likelihood of the observed response vector is given by Eq. (2) above assuming the compensatory MIRT model. When $\boldsymbol{\theta}$ is the parameter of interest and the vector of responses $\mathbf{u}$ is considered fixed, the joint probability is represented by $L(\mathbf{u}|\boldsymbol{\theta})$. The posterior density of $\boldsymbol{\theta}$ given the response vector $\mathbf{u}$ is given by

$$f(\theta|\mathbf{u}) = \frac{L(\mathbf{u}|\theta)f(\theta)}{f(\mathbf{u})}, \tag{33}$$

where $f(\mathbf{u})$ is the marginal probability of $\mathbf{u}$ given by

$$f(\mathbf{u}) = \int_{-\infty}^{\infty} L(\mathbf{u}|\theta)f(\theta)\,\mathrm{d}(\theta). \tag{34}$$

The mode of the posterior density is used as the estimate for $\boldsymbol{\theta}$ because it is more convenient to compute that value than the mean or the median, especially when the number of coordinate dimensions is high. The differences in the various methods for summarizing the posterior distribution are small when a reasonable number of items have been administered.

Segall (2000) recommends using the characteristics of the posterior distribution for $\boldsymbol{\theta}$ to determine the next item to administer. Specifically, the item is selected that will result in the "largest decrement in the size of the posterior credibility region" (p. 62). The details of the mathematical derivation of the practical implementation of this criterion would require more space than is available in this chapter. The interested reader should refer to Segall (2000) for the details. He shows that maximizing the posterior information matrix provides the most decrement in the posterior credibility region. The information is computed for the already administered items, $S_{k-1}$, and each of the possible items that are candidates for administration, $i$. The criterion is evaluated at the most recent estimate of the $\boldsymbol{\theta}$ vector. The posterior information matrix is given by

$$I_{i,S_{k-1}}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\,\partial\theta'}\ln f(\theta|\mathbf{u}_k)\right]. \tag{35}$$

When this value is determined in practice, a number of simplifying assumptions are made. The posterior distribution is assumed to be normally distributed even though the actual posterior distribution is not normal. The mean of the normal approximation is set to the mode of the actual posterior.

When MIRT CAT is implemented in practice, the number of items administered reaches a preset value, or the process continues until the posterior credibility region is

smaller than some threshold value. When the stopping criterion has been reached, the estimated $\boldsymbol{\theta}$-vector is assumed to indicate the location of the examinee in the multidimensional space. The elements of the vector can be reported to provide results related to the coordinate axes, or the point can be projected onto reference composites to give estimates of performance on constructs defined by sets of items.

## 8. Discussion and conclusions

MIRT is a theory that indicates that the interaction between a person and a test item can be represented by probabilistic models that represent the person by a vector of coordinates in a multidimensional space. The application of that theory is through the development and evaluation of the forms of the probabilistic models and the development of statistical methodology for estimating the parameters of the models. If the MIRT model provides a reasonable representation of the item response data from the interaction of persons and test items, then the collection of related statistical methods can be applied to challenging problems in educational and psychological measurement.

The most immediate problem is dealing with the fact that most educational and psychological measurement instruments are not designed to be consistent with unidimensional IRT models. Even those that report single scores and use unidimensional IRT models for reporting scores typically have complex content specifications that imply that multiple skills and knowledge are needed to successfully respond to the test items. These specifications imply that multidimensional models are needed to properly model the item response data. Robustness of the unidimensional models and essential unidimensionality (Stout, 1987) are often used to support the use of the unidimensional models, but there is also an extensive literature checking whether or not unidimensionality assumptions are reasonable for the data from a particular test.

My own perspective on the dimensionality problem is that all matrices of test data need multidimensional models to accurately represent the interaction of persons and test items, but sometimes the multiple dimensions are highly correlated in the examinee sample or the set of items in the test are not very sensitive to differences in many of the dimensions. In the latter cases, the use of unidimensional models may be more efficient ways of getting at the major reference composite for the test, but it is still helpful to understand that other dimensions are in play. That understanding helps put differential item functioning into a formal theoretical context and helps explain the formal requirements for test parallelism.

MIRT is still in its infancy as a psychometric methodology. While there are few applications of MIRT within large scale testing programs as of 2005, there is a growing body of research related to the development of practical methods. At the time of the writing of this chapter, estimation procedures are fairly well developed and the practical issues of linking calibrations and representing analysis results are being addressed. A few applications of MIRT to test design and analysis are beginning to appear.

It is tempting to predict that MIRT procedures will be widely used in large scale testing programs within the next few years. Such predictions are difficult to support. The history of the implementation of computerized adaptive testing shows that very

elegant solutions to testing problems do not necessarily get put into practice because of practical issues related to computer availability and test security. It appears that the use of MIRT will improve test design and development, the reporting of test results, and the efficiency of measurement. Whether or not the promise of this theory and methodology becomes a reality will depend largely the level of motivation of researchers in this field. It is hoped that this chapter will provide stimulus for progress in the practical application of this MIRT methodology.

## References


Adams, R.J., Wilson, M., Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement* **21**, 1–23.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M., Novick, M.R. (Eds.), *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

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

Bolt, D.M., Lall, V.F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement* **27**, 395–414.

Briggs, D.C., Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement* **4**, 87–100.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Scientific Software International, Lincolnwood, IL.

Fischer, G.H., Molenaar, I.W. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*. Springer, New York.

Fraser, C. (1988). NOHARM II: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. The University of New England, Armidale, Australia.

Hambleton, R.K., Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer–Nijhoff, Boston, MA.

Hulin, C.L., Drasgow, F., Parsons, C.K. (1983). *Item Response Theory*. Dow Jones-Irwin, Homewood, IL.

Li, Y.H., Lissitz, R.W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement* **24**, 115–138.

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

Martineau, J.A. (2004). The effect of construct shift on growth and accountability models. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

McDonald, R.P. (1962). A general approach to nonlinear factor analysis. *Psychometrika* **27**, 397–415.

McDonald, R.P. (1967). *Nonlinear Factor Analysis*. Psychometric Monographs, No. 15.

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

McDonald, R.P. (1982). Some alternative approaches to the improvement of measurement in education and psychology: Fitting latent model. In: Spearritt, D. (Ed.), *The Improvement of Measurement in Education and Psychology: Contributions of Latent Trait Theories*. Australian Council for Educational Research.

McDonald, R.P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement* **24**, 99–114.

Miller, T.R., Hirsch, T.M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory. *Applied Measurement in Education* **5**, 193–211.

Min, K.-S. (2003). The impact of scale dilation on the quality of the linking of multidimensional item response theory calibrations. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Mulaik, S.A. (1972). *The Foundations of Factor Analysis*. McGraw-Hill, New York.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4, pp. 321–334.

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

Reckase, M.D. (1997a). The past and future of multidimensional item response theory. *Applied Psychological Measurement* **21** (1), 25–36.

Reckase, M.D. (1997b). High dimensional analysis of the contents of an achievement test battery: Are 50 dimensions too many? Paper presented at the meeting of the Society for Multivariate Experimental Psychology, Scottsdale, AZ, October.

Reckase, M.D. (1997c). A linear logistic multidimensional model for dichotomous item response data. In: van der Linden, W.J., Hambleton, R.K. (Eds.), *Handbook of Modern Item Response Theory*. Springer, New York.

Reckase, M.D., Hirsch, T.M. (1991). Interpretation of number correct scores when the true number of dimensions assessed by a test is greater than two. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, April.

Reckase, M.D., Martineau, J. (2004). Vertical scaling of science achievement tests. Paper commissioned by the Committee on Test Design for K-12 Science Achievement, Center for Education, National Research Council, Washington, DC, October.

Reckase, M.D., Stout, W. (1995). Conditions under which items that assess multiple abilities will be fit by unidimensional IRT models. Paper presented at the European meeting of the Psychometric Society, Leiden, The Netherlands, July.

Reckase, M.D., Ackerman, T.A., Carlson, J.E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement* **25**, 193–203.

Reckase, M.D., Thompson, T., Nering, M. (1997). Identifying similar item content clusters on multiple test forms. Paper presented at the meeting of the Psychometric Society, Gatlinburg, Tennessee, June.

Reckase, M.D., Martineau, J., Kim, J. (2000). A vector approach to determining the number of dimensions needed to represent a set of variables. Paper presented at the annual meeting of the Psychometric Society, Vancouver, Canada, July.

Segall, D.O. (2000). Principles of multidimensional adaptive testing. In: van der Linden, W.J., Glas, C.A.W. (Eds.), *Computerized Adaptive Testing: Theory and Practice*. Kluwer, Dordrecht, The Netherlands.

Spray, J.A., Davey, T.C., Reckase, M.D., Ackerman, T.A., Carlson, J.E. (1990). Comparison of two logistic multidimensional item response theory models. Research Report ONR90-8, ACT, Inc., Iowa City, IA, October.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika* **52**, 589–617.

Stout, W., Habing, B., Douglas, J., Kim, H.R., Roussos, L., Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement* **20**, 331–354.

Sympson, J.B. (1978). A model for testing with multidimensional items. In: *Proceedings of the 1977 Computerized Adaptive Testing Conference*, University of Minnesota, Minneapolis.

Wang, M.-m. (1985). Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecifications on the application of IRT. Unpublished manuscript.

Ward, J.H. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association* **58**, 234–244.

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

Whitely, S.E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika* **45**, 479–494.