Multidimensional Item Response Theory

Alan Socha

James Madison University
**Background**

Item response theory (IRT) is a system of models that explains the relationship between persons and test items. Most of the time a one-parameter (1PL), two-parameter (2PL), or three-parameter (3PL) unidimensional model is calibrated. The 1PL and 2PL models can be seen as special cases of the 3PL model. The 3PL model contains parameters for item difficulty ($b$ or $\delta$), item discrimination ($a$ or $\alpha$), and a lower-asymptote ($c$ or $\chi$) used to estimate guessing. In the 2PL model the lower-asymptote is fixed to 0. In the 1PL model the lower-asymptote is fixed like in the 2PL model, and the item discrimination parameter is fixed to be equal across all items. All three models estimate a unidimensional latent variable/construct ($\theta$) representing *ability*, *demonstrated proficiency*, or *latent trait*. The term demonstrated proficiency will be used here to represent all of these.

A unidimensional latent variable is not always appropriate though, as most sets of test items may measure several demonstrated proficiencies instead of a single one (Ackerman, 1994; Reckase, 1985; Reckase, Ackerman, & Carlson, 1988). In these cases, when the assumption of unidimensionality is violated, multidimensional item response theory (MIRT) models should be used. MIRT models describe the interaction between test items and a person when the characteristics of a person are described using a vector of hypothetical constructs instead of a single construct (Reckase, 1997b). MIRT can be used to serve a number of different purposes. For instance, MIRT provides a way for describing differential item functioning (DIF; Reckase, 2002) and to support item selection for a test so that unidimensional models can be used to describe the item responses (Reckase, 2002).
Assumptions

As with any model, the assumptions underlying the model must be investigated. These assumptions are very similar to those of unidimensional IRT models. Unidimensional models have the assumptions of conditional independence, functional form, and, of course, unidimensionality.

When the responses to an item are independent of the responses to any other item, conditional on the person’s demonstrated proficiency, conditional independence is met. In multidimensional models, this definition is expanded to mean that for any group of individuals that are characterized by the same values of $\theta_1, \theta_2, \ldots, \theta_F$, where $F$ is the number of dimensions, the conditional distributions of the item responses are all independent of each other (de Ayala, 2009). In other words, responses to an item are independent of the responses to any other item, conditional on all of the person’s demonstrated proficiencies.

The functional form assumption means that the data follow the function specified by the model. For example, in a 1PL model this would mean that there is no guessing and that the item discriminations really are equal throughout the entire test. This assumption is basically the same in multidimensional models and is assessed by investigating goodness-of-fit. Residuals can also be useful in determining if the model is inappropriate (Reckase, 1997a).

The assumption of unidimensionality assumes that the responses are a manifestation of only one latent dimension. Violation of this assumption is why we choose to use MIRT models. McDonald’s nonlinear factor analyses, Holland and Rosenbaum’s conditional association approach, and Nandakumar and Stout’s DIMTEST procedure can be used to test the assumption of unidimensionality (Ackerman, 1994). Residuals are also useful in determining whether too few dimensions are specified in the model (Reckase, 1997a). In MIRT models, the assumption
changes to a general dimensionality assumption. In other words, the number of dimensions in the model is correctly specified. The question, of course, is how many dimensions actually exist when the unidimensionality assumption is violated. A scree plot of the eigenvalues obtained from a principal factor analysis is one method for determining the number of dimensions (Ackerman, 1994).

Several things need to be carefully considered regarding dimensionality. For example, even though most items may be multidimensional, the strength of each dimension may not be strong and a unidimensional model may still be appropriate (Reckase, 1985). Tests that contain items that measure the same weighted composite of multiple dimensions may also still meet the requirements of a unidimensional model, despite requiring several demonstrated proficiencies to answer the items correctly (Reckase et al., 1988). Other situations that can be modeled unidimensionally are when the items are capable of measuring different skills but the examinees only vary in their level of proficiency on one of those skills, when the items are measuring levels of only one skill and examinees vary in their levels of proficiency of each of those skills, and when the items collectively distinguish between levels of several skills and the examinees differ in their levels of proficiencies on more than one of these skills (Ackerman, 1994).

**Item and Person Parameters**

Multidimensional item parameters can be interpreted in a similar way to unidimensional item parameters except that they are vectors and the direction in $\theta$-space influences the interpretation. The lower-asymptote is the exception, which is not a vector and has the same meaning in the 3PL model as it does in a multidimensional model. Person parameters are elements of the vector of demonstrated proficiency, $\theta_i$. The dimensions in this vector can be orthogonal, but do not have to be.
The discrimination parameters are elements of the $\alpha$-vector. The direction in the $\theta$-space determines the discriminating power of an item (Reckase, 1997a). In other words, the $\alpha$-vector slopes in the direction of a particular $\theta$-axis indicate the sensitivity the item is to differences in demonstrated proficiency along that axis. For example, if the direction is parallel to the surface, the slope will be zero and the item non-discriminating. Chances are, an item will not be a pure measure of a particular dimension, but instead will be more or less discriminating for combinations of dimensions.

The multidimensional discrimination (MDISC or $A_i$) statistic is an overall measure of the capability of an item to distinguish between individuals that are in different locations in $\theta$-space (Reckase & McKinley, 1991). $A_i$ is conditional on the direction in $\theta$-space, and, if this direction is chosen to match that of the coordinate axes, $A_i$ would give information about how well the item measures a particular dimension. The direction of the maximum slope, $\omega_{\alpha}$, indicates the weighted composite of demonstrated proficiencies that is best measured by an item (Reckase et al., 1988). This direction is a function of an item’s discrimination, $\alpha_{\alpha}$, and the multidimensional item discrimination parameter, $A_i$. The multidimensional discrimination parameter and maximum slope are defined as:

$$A_i = \sqrt{\sum_{j=1}^{F} \alpha_{\alpha j}^2} \quad \omega_{\alpha} = \arccos \left( \frac{\alpha_{\alpha}}{A_i} \right)$$

$A_i$s can only be compared when they are measuring the same direction, so a directional discrimination, $A_\omega$ (where $\omega$ is the common direction), can be calculated to compare the discrimination capacity of items that are measuring in different directions. The directional discrimination is defined as:

$$A_\omega = \sum_{j=1}^{F} \alpha_{\alpha j} \cos(\omega_{\alpha j})$$
The multidimensional item difficulty (MID or $\Delta_i$) describes the difficulty of an item as a direction and a distance in the latent space (Reckase, 1985). In order to specify the $\Delta_i$, $n$ statistics are needed, where $n$ is the number of dimensions. These statistics are $n-1$ angles and a distance. For example, suppose we have two dimensions. In this case, the angle of one of the axes and the distance along the vector to the point of maximum discrimination are needed to specify $\Delta_i$. The direction that gives the steepest slope for $A_i$ will be the same as the direction specified for $\Delta_i$. The same direction must be used if the difficulty will be compared across items, since the direction specifies the composite of demonstrated proficiencies. $\Delta_i$ is defined as:

$$\Delta_i = -\frac{\gamma_i}{A_i} \text{ where } \gamma_i \text{ is the difficulty}$$

**Item and Test Characteristic Surfaces and Information**

The item characteristic surface (ICS) is the multidimensional analog to the item characteristic curve (ICC). In the ICC there is one $\theta$ for a given item parameter set that will result in the logit equaling 0.0, but in the ICS there can be many. This is because a different slope can be calculated for each direction one might use to traverse the ICS.

Vector graphs can simplify the presentation of multidimensional item characteristics (de Ayala, 2009). Vector graphs present the item’s location, how well it discriminates, and which dimension, if any, it measures best. The starting point ($\Delta_i$), length ($A_i$), and direction in multidimensional space ($\omega_i$) are needed in order to graph an item vector.

The test characteristic surface (TCS) is the multidimensional analog to the test characteristic curve (TCC; de Ayala, 2009). The TCS tells us how much information an item can provide for a given direction. de Ayala (2009) suggested that it might be easier to choose the direction consistent with the item’s vector, since there are multiple item information surfaces. Comparing the TCSs for different test forms, which can be done graphically by plotting the
surface that represents the difference of the two TCSs, is a useful way to investigate whether multiple forms of a test are parallel (de Ayala, 2009). The total multidimensional information can also be calculated, but is also contingent on the direction. The same direction must be used for all items in calculating the total multidimensional information. The multidimensional item information can be given as:

\[
I_{\omega}(\theta) = p_i(1 - p_i)\left(\sum_{f} \alpha_{\omega f} \cos \omega_f\right)^2
\]

**Types of MIRT Models**

There are two main types of MIRT models: noncompensatory models and compensatory models. Noncompensatory models are the simpler of the two. Such models are used when a respondent’s location on one dimension does not compensate for his/her location on the other dimension(s) (de Ayala, 2009). In other words, a deficit in \( \theta \) on one dimension cannot overcome a deficit in \( \theta \) on another dimension (Reckase, 1997b). The \( m \)-dimensional noncompensatory two-parameter logistic (MNC2PL) model is given as:

\[
P(x_i = 1) = \prod_{k=1}^{m} \frac{e^{\alpha_k (\theta - \gamma_{ik})}}{1 + e^{\alpha_k (\theta - \gamma_{ik})}}, \text{ where } \gamma_i = -\Delta_i \Delta_j.
\]

This is basically the product of two or more 2PL models. The multiplicative nature prohibits an examinee from compensating for low demonstrated proficiency on one dimension by being high on another. The compensatory model, on the other hand, is additive. The additive nature allows the examinee to be able to compensate for low demonstrated proficiency on one dimension by being high on another. In other words, a correct response can be obtained even with a low \( \theta \)-value on some dimensions (Reckase, 2002). An example of this is a mathematical word problem. Someone with low mathematics proficiency can still get an item correct by
compensating with high verbal proficiency. The $m$-dimensional compensatory two-parameter logistic (MC2PL) and $m$-dimensional compensatory three-parameter logistic (MC3PL) models are defined as (where $\gamma_i$ is defined the same as it is for the MNC2PL model):

$$\text{MC2PL: } P(x_i = 1) = \frac{\sum_{k=1}^{m} \alpha_k \theta_k + \gamma_i}{1 + e^{\sum_{k=1}^{m} \alpha_k \theta_k + \gamma_i}}$$

$$\text{MC3PL: } P(x_i = 1) = \chi_i + (1 - \chi_i) \frac{\sum_{k=1}^{m} \alpha_k \theta_k + \gamma_i}{1 + e^{\sum_{k=1}^{m} \alpha_k \theta_k + \gamma_i}}$$

Similar to unidimensional models, joint maximum likelihood estimation (JMLE) and maximum marginal likelihood estimation (MMLE) can be used to calibrate the item parameters. Several software packages exist for calibrating MIRT models. LOGIST is one such program that utilizes JMLE (Reckase, 1997a). MULTIDIM (Reckase, 1997a), NOHARM (Reckase, 2002), and MAXLOG (McKinley & Reckase, 1983) each uses MMLE. There is also TESTFACT, which utilizes least squares estimation (Reckase, 2002).

**Future Research**

Much needs to be done in MIRT. First, it should be noted that no software packages currently exist to estimate the parameters in noncompensatory models (Ackerman, 1994) and thus should be developed, with the exception of some multidimensional Rasch models. Reckase (2002) suggested that sample sizes of over 1,000 and fairly long tests are needed to produce stable parameter estimates while Ackerman (1994) suggested that at least 2,000 examinees were needed. Therefore, one final need is for more knowledge concerning the data requirements needed to support high-dimensional spaces (Reckase, 2002).
References


