An NCME Instructional Module on

Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests

Terry A. Ackerman, *University of North Carolina-Greensboro*
Mark J. Gierl, *University of Alberta*
Cindy M. Walker, *University of Wisconsin-Milwaukee*

Many educational and psychological tests are inherently multidimensional, meaning these tests measure two or more dimensions or constructs. The purpose of this module is to illustrate how test practitioners and researchers can apply multidimensional item response theory (MIRT) to understand better what their tests are measuring, how accurately the different composites of ability are being assessed, and how this information can be cycled back into the test development process. Procedures for conducting MIRT analyses—from obtaining evidence that the test is multidimensional, to modeling the test as multidimensional, to illustrating the properties of multidimensional items graphically—are described from both a theoretical and a substantive basis. This module also illustrates these procedures using data from a ninth-grade mathematics achievement test. It concludes with a discussion of future directions in MIRT research.

Keywords: dimensionality, multidimensional item response theory, test development and analysis

The purpose of this module is to illustrate how test practitioners and researchers can apply multidimensional item response theory (MIRT) to understand better what their tests are measuring, how accurately the different composites of ability are being assessed, and how this information can be cycled back into the test development process. MIRT is used to model the relationship between *two or more* unobservable variables, conceptualized as constructs or dimensions, and the probability of the examinee correctly answering any particular test item. Conversely, unidimensional item response theory is used to model the relationship between *one* unobserved construct or dimension and the probability of the examinee correctly responding to any particular test item. That is, we present the multidimensional analogue to unidimensional item response theory (IRT), extending and expanding the *Educational Measurement: Issues and Practice* module presented by Harris (1989 Module 7). Note, however, that many unidimensional concepts become increasingly complex as one considers multiple dimensions. For example, the unidimensional item characteristic curve becomes an item response surface

Terry A. Ackerman is Professor, University of North Carolina–Greensboro, 207 Curry Building, UNCG, P.O. Box 26170, Greensboro, NC 27402-6170; e-mail: taackerm@uncg.edu. His areas of specialization are item response theory, multidimensional item response theory, differential item/test functioning, and test dimensionality.

Mark J. Gierl is Associate Professor, Centre for Research in Applied Measurement and Evaluation. 6-110 Education North, Faculty of Education, University of Alberta, Edmonton, AB, Canada T6G 2G5; e-mail: mark.gierl@ualberta.ca. His areas of specialization are educational and psychological measurement, focusing on differential item and bundle functioning, cognitively diagnostic assessment, unidimensional and multidimensional item response theory, and test translation and adaptation.

Cindy M. Walker is Assistant Professor, Department of Educational Psychology, University of Wisconsin–Milwaukee, P.O. Box 413, Enderis Hall 785, Milwaukee, WI 53201; e-mail: cmwalker@uwm.edu. Her areas of specialization are educational and psychological measurement, unidimensional and multidimensional item response theory, differential item functioning, dimensionality assessment, and assessment of mathematical understanding and ability.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. This module may be photocopied without permission if reproduced in its entirety and used for instructional purposes only. The complete series is available at www.ncme.org.

in two dimensions. One of our goals with this mod-ule is to guide the reader through this transition.

The module is divided into four sections. In the first section we discuss dimensionality and present methods used to verify that meaningful, replicable dimensions are, in fact, being assessed on a particular test. In the second section we develop one multidimensional item response theory model and show how it relates to the two-parameter unidimensional IRT model. In the third section we illustrate how to use the computer program NOHARM for MIRT item parameter calibration and how to graphically represent item characteristics in a multidimensional space. In the fourth section we present future directions for MIRT research. Throughout the module concepts are illustrated using ninth-grade mathematics achievement test data from an examination administered in the Canadian province of Alberta. Although MIRT can be used to model many dimensions, to simplify the illustrations in this module, we will only consider two-dimensional data using dichotomously scored items.

At the outset, though, some words of caution are in order. Although multidimensional statistical and graphical results may appear to present evidence of analytic rigor, they have little utility unless they can help the practitioner and researcher better understand what the test scores represent. MIRT analyses should not only provide validity evidence but also insight that can be cycled back into the test development process (Ackerman, 1994, 1996). Too often psychometric analyses are solely quantitative in nature and focus primarily on examinees' correct responses, ignoring the actual test content and cognitive processes required by examinees to solve items on the test. It is our conviction that psychometric analyses should always be guided by substantive hypotheses, analyses, and interpretations. Moreover, whenever quantitative analyses are conducted based on substantive considerations, they should be confirmatory in nature and motivated by a thorough review of the items, keeping in mind what the test was designed to measure. This approach is essential for unidimensional and multidimensional analyses, and it is the approach taken in this module. Finally, there is an axiom that the reader must attend to: Working with real data is never easy and rarely are the interpretations straightforward. As a result, a complete test analysis should be viewed as an iterative craft, not just a straightforward application of measurement principles and formulas.

Assessing Dimensionality

Overview

Many educational and psychological tests are inherently multidimensional, meaning these tests measure two or more constructs or dimensions.[1] A construct is a theoretical representation of the underlying trait, concept, attribute, process, and/or structure that the test is designed to measure (Messick, 1989). The items on a *factorially simple* test measure *one* underlying dimension (McDonald, 1999). For example, one might believe that a mathematics test is measuring one identifiable construct—algebra. However, the dimensional structure of most real testing data is much more complex. The items on a *factorially complex* test measure *two or more* underlying dimensions (McDonald, 1999). For example, one might suspect a mathematics test is measuring algebra and geometry. In this case, a subset of test items with algebra content might be considered a measure of the first dimension, whereas the remaining items with geometry content might be considered a measure of the second dimension.

In this module, we focus on factorially complex tests where some item responses are influenced by two underlying dimensions. When all items lie along the two-dimensional coordinate axes (see section "Estimating and Representing MIRT Item Parameters") the test displays *simple structure*[2] (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996). Items can also lie in a narrow sector around the two-dimensional coordinate axes. In this case, the test displays *approximate simple structure.* When items lie throughout the two-dimensional latent space (i.e., items measure a range of skills in the $\theta_1\theta_2$ composite) the test displays *complex structure.* Traditionally, most dimensionality analyses have focused on identifying simple structure because it is easy to interpret; in fact, all of the dimensionality analyses discussed in this module are designed to assess testing data that approximate simple structure.

Recently, Zhang and Stout (1999) noted that "a certain pattern of separated clusters of items about the test composite should typically result from the categorical nature of many test specifications" (p. 214). This statement suggests that dimensionality and MIRT analyses should be supported by, and perhaps even preempted with, substantive judgment. Specifically, a thorough analysis of the content areas and cognitive skills needed to successfully respond to each item should be conducted. It might be helpful to conduct this substantive analysis by referring to the test specifications with the aid of specialists who have extensive knowledge of the content and the examinees' cognitive skills. If subsets of items measure different content areas and/or cognitive skills, then these items have the potential to represent distinct dimensions. Hence, the first step in any MIRT analysis is to determine whether the data are indeed multidimensional.

Traditionally, linear factor analysis, using tetrachoric correlations (polychoric correlations in the case of polytomous data), has been used to assess the dimensionality of test data. However, there are problems with this approach. For example, the relationship between item performance and the underlying latent ability is often nonlinear (Hattie, 1984). Nonlinearity can result in a mismatch between the model and the data. Dimensionality can also be confounded with item difficulty, such that the factors represent items with comparable difficulty levels as opposed to items that measure distinct dimensions. Moreover, the mathematics underlying factor analysis requires the matrix of tetrachoric correlations to be positive semidefinite, a condition that is not always satisfied with real data. Finally, there is a no standard approach for determining the number of meaningful factors (Mislevy, 1986).

Many other empirical methods have been proposed to investigate the dimensionality of test data (e.g., Hambleton & Rovinelli, 1986; Hattie, 1984, 1985), including, more recently, the development of several nonparametric tests based on Stout's (1987) theory of *essential unidimensionality.* Essential dimensionality is based on the assumption that there is only one dominant latent ability that influences examinees' responses to items (Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar, 1991; Nandakumar & Stout, 1993; Stout et al., 1996; Zhang & Stout, 1999). Unfortunately, most dimensionality analyses are exploratory in nature and many of these procedures produce results that contradict substantive dimensionality hypotheses.

As we advocated previously, substantive judgment should guide the assessment of dimensionality in at least three different ways. First, *test specifications* can guide the assessment of dimensionality. Test specifications outline the achievement domain and help test developers obtain a representative sample of items from this domain. The specifications also guide item writing and help structure the final form of the test based on the content *and* cognitive domain that the test is designed to measure. Second, a *content analysis* can guide the assessment of dimensionality. For example, test development specialists can review items and identify dimensions based on item content. A content analysis is guided by the professional experience of the reviewers. Two variations of content review can be used: (a) specialists may use their experience and judgment to identify dimensions during an item review or (b) content-based judgments can be found in the educational and psychological literature to guide interpretation using well known tests. Third, *psychological analyses* can guide dimensionality assessment when the hypothesized item structure is formulated from a psychological perspective. For example, a cognitive task analysis could be used to identify skills that characterize mathematics performance (e.g., Gallagher, 1998; Gallagher, De Lisi, Holst, McGillcuddy-De Lisi, Morely, and Cahalan, 2000). These cognitive skills could be identified and operationalized using test items to inform a dimensionality assessment.

To illustrate a systematic approach for investigating the dimensionality of test data we used the results from the 1996 ninth-grade mathematics achievement test administered in the Canadian province of Alberta. For the purposes of this instructional module, a 35-item multiple-choice test was analyzed using data from 6,000 examinees randomly selected from the original database. When one has access to large databases in actual test analyses, it is often prudent to cross-validate the results of dimensionality analyses using different subsets of data (e.g., different samples from the same population).

In our example, a substantive analysis of the item content was conducted, as we previously advocated. Two content reviewers, who were familiar with the content area, had several years of math teaching experience, and had participated in the process of standardized test development, evaluated the 35-item mathematics test. These reviewers concluded that two distinct abilities were being measured by the test: general mathematics ability and spatial ability. Furthermore, the reviewers believed that all of the items were measuring general mathematics ability while only six items were measuring spatial ability. Consequently, the data were sorted such that the first six items in the 35-item test were hypothesized to measure both spatial ability and general mathematics ability whereas the remaining 29 items were hypothesized to measure only general mathematical ability. A discussion and illustration of some of the more recent dimensionality tests used to evaluate these substantive hypotheses are presented next.

Methods for Assessing Dimensionality

When using any of the methods in this section to assess dimensionality, one must remember that these procedures are only tools. Whereas promising results for these procedures have been found in simulation studies, relatively few published studies have supported the procedures using actual test data. Moreover, the dimensionality analyses conducted for the purpose of this module did not always provide strong support for the hypothesized underlying structure. The reader should note that dimensionality assessment is a large and encompassing area in psychometric research. In fact, a separate instructional module could be devoted to this topic alone. Our intention is not to review the vast number of procedures available for dimensionality analyses but only to survey a small number of the more promising techniques. Our goal is to describe these procedures so the reader has a *conceptual* understanding of how they work. The reader is referred to the original references for a more mathematical and theoretical account. In this module, we describe three popular methods that can be used to assess dimensionality: hierarchical cluster analysis, DETECT, and DIMTEST. Hierarchical cluster analysis and DETECT can aid practitioners and researchers in assessing the dimensionality of test data by employing statistical techniques that may be helpful in forming sets of mutually exclusive items. Once this information is gathered, the computer program DIMTEST can be used to conduct a statistical test to determine if the groups of items are distinct dimensions. The software to conduct these analyses can be obtained from Assessment Systems Corporation (www.assess.com). In describing these procedures, we are merely reinforcing the importance of evaluating the dimensionality of the test data prior to performing multidimensional IRT analyses.

Hierarchical Cluster Analysis. Roussos (1992) developed two computer programs, CCPROX and HCA, which are used in conjunction with one another to conduct hierarchical agglomerative cluster analysis. This procedure is used to cluster items into progressively larger groups deemed to be dimensionally homogeneous starting with each item constituting its own cluster and concluding with all items in one cluster. The program can be run with 120 dichotomously scored items and with no restriction on the number of examinees. Prior to conducting a cluster analysis, a measure of proximity between all possible pairs of items must be found. CCPROX allows the user to choose from several different measures of proximity, one of which is the estimated conditional covariance between pairs of items, conditioning on an examinee's score using the remaining items (Douglas, Kim, Roussos, Stout, & Zhang, 1999). In simulation studies, this measure of proximity was demonstrated to be sensitive to multidimensionality (Douglas et al., 1999; Hartz, Roussos, & Stout, 2000) and was the measure used in our example.

The proximity matrix serves as input to the program HCA that will subsequently form the various item clusters. At the start of the cluster analysis procedure, each item forms its own cluster. The algorithm completes $k-1$ iterations for a k-item test and at each iteration of the algorithm the two clusters that are closest in proximity are joined together. In the final stage of the HCA algorithm, all items are combined forming a single cluster. Within the HCA program there are many options for determining the proximity of clusters. Results from simulation studies suggest that using the unweighted pair group method of averages (UPGMA) for proximity produces the most accurate classification of items when approximate simple structure exists (Douglas et al., 1999) and was the approach used in this example. It is possible for more than one pair of clusters to be closest in proximity. If this happens, one can combine the first or last pair. Because this decision is arbitrary, Roussos (1992) suggested that the pro-

gram be run twice, employing each of the tie-breaking mechanisms to ensure that the results are not affected.

Hierarchical cluster analysis can aid the practitioner and researcher who suspects multidimensionality but is unsure of the underlying structure. However, this type of analysis always results in a set of distinct clusters, regardless of whether or not the data are multidimensional. Therefore substantive judgment should be used when interpreting the output. Items are never removed from a cluster once it is formed. Furthermore, the solution for each successive iteration of the algorithm depends on the previous solution, hence the solution attained at one level or more may not be optimal.

This possible outcome may help explain the solution we found using the actual test data. In our example, Items 1 through 6 were thought to measure the spatial ability construct. Five of these items tended to cluster together at several different levels. However, they never formed a distinct cluster of their own. For example, Item 1 did not join a cluster until Level 12, and the cluster it joined contained Item 5. However, Item 5 previously formed a cluster with Item 16 at Level 5, which was not expected. In addition, Items 2, 3, and 4 formed a cluster at Level 18 of the analyses and remained a distinct cluster until Level 30; however, in Level 30, these items were joined with a cluster consisting of items that were not suspected of measuring the spatial ability construct. Likewise, the cluster that contained Items 1 and 5 joined items not suspected of measuring spatial ability at Level 27. Item 6 appeared to be an anomaly, standing alone until Level 24 and never joining the cluster of items suspected of measuring the spatial ability construct.

DETECT. DETECT is an exploratory nonparametric dimensionality assessment procedure that estimates the number of dominant dimensions present in a data set and the magnitude of the departure from unidimensionality. DETECT also identifies the dominant dimension measured by each item (Roussos, Reese, & Harris, 1997). It can be run with 120 dichotomously scored items with up to 6,000 examinees. This procedure produces mutually exclusive, dimensionally homogeneous clusters of items. The user specifies the maximum number of dimensions. Because the clusters of items are mutually exclusive, this procedure is most useful when a researcher suspects approximate simple structure. Although the procedure can still be informative when simple structure fails to hold, different clusters of items identified by the procedure in this case may actually be quite homogeneous (Zhang & Stout, 1999).

The main objective of DETECT is to identify clusters that maximizes the value of the DETECT index. This index represents the magnitude of departure from unidimensionality (Douglas et al., 1999). Homogeneous clusters of items that are as distinct as possible are found using a genetic algorithm. The DETECT index is created by computing all item covariances, conditioning on examinees' scores using the remaining items (Zhang & Stout, 1999). When the data are unidimensional, clusters of items will be found that are not particularly homogeneous. In this case, the conditional covariance will be positive for some pairs of items and negative for other pairs of items resulting in a DETECT index that is near zero. If, however, the underlying structure of the data is multidimensional, clusters of items will be found that have positive within-cluster conditional covariances and negative between-cluster conditional co- variances, resulting in a large DETECT index.

Based on results from simulation studies, Kim (1994) suggested that when the DETECT index is less than 0.10, the data can be considered unidimensional; an index between 0.10 and 0.50 can be considered a weak amount of dimensionality; an index between 0.51 and 1.00 can be considered a moderate amount of dimensionality; and an index greater than 1.00 can be considered a strong amount of dimensionality. Using DETECT on the data considered in this module and requesting two dimensions resulted in a DETECT index of 0.13, signaling a relatively weak amount of multidimensionality. Similar to the results obtained using HCA, one cluster consisted of Items 2, 3, and 4. Unlike the results obtained using HCA, this cluster also included Item 6, an item suspected of being dimensionally comparable to Items 1 through 5. Once again, Items 1 through 6 never formed a distinct group and Items 1 and 5 clustered with items that were not suspected of measuring the spatial ability construct.

DETECT also provides the user with an index, r, representing how well the underlying structure of the data approximates simple structure. Values of r that are greater than 0.80 suggest that approximate simple structure of the data holds. For the data considered in this module $r = 0.46$, suggesting that approximate simple structure does not hold as we originally predicted.

DIMTEST. One of the most promising dimensionality analyses available to practitioners and researchers is the latest version of DIMTEST (Froelich, 2000; Froelich & Habing, 2001; Stout, Froelich, & Gao, 2001). DIMTEST is a nonparametric statistical procedure that conducts a hypothesis test to assess the presence of multidimensionality. Similar to the previous procedures described, DIMTEST assesses the relationship between subsets of items based on conditional item covariances. However, unlike previous procedures described, DIMTEST allows the user to conduct confirmatory analyses. It can be run with 120 dichotomously scored items with up to 6,000 examinees. The most recent version of DIMTEST only requires the user to select subsets of test items that measure the same dominant dimension. These subsets of items can be identified using any dimensionality assessment such as substantive judgment, hierarchical cluster analysis, or DETECT.

The test statistic, T, calculated by DIMTEST represents the degree of dimensional distinctiveness of two clusters of items. T is distributed normally with an expected value of zero. Based upon the substantive a priori hypothesis of the spatial and general math dimensions, DIMTEST was used to determine whether Items 1 to 6 were dimensionally distinct from the remaining items. The resulting T statistic estimated by DIMTEST was 2.69 ($p = .004$), suggesting the spatial items are dimensionality distinct from the remaining items on the mathematics test.

In summary, the results from DIMTEST suggest that the spatial items identified by the content reviewers measure a distinct dimension when compared with the remaining items on the mathematics test. However, the results from DETECT indicate that the spatial items only signal a weak amount of dimensionality due, perhaps, to the inclusion of Items 1 and 5 that were not found to cluster with the remaining spatial items in the hierarchical cluster analysis. Clearly, these three methods offer different lenses from which to view dimensionality and, typical of most multi-

method studies, the results from these procedures vary in their ability to discern the signal of the valid skills from construct irrelevant noise leaving the researcher to resolve the different results. It is important to remember that these three procedures were designed to assess approximate simple structure. The results from the substantive reviews and the graphical representations of the six spatial items, presented in the section "Estimating and Representing MIRT Item Parameters," suggest that our data have a more complex structure.

Introduction to Multidimensional Item Response Theory

Overview

The main focus of this module is on the two-parameter compensatory MIRT model because it has been extensively developed, studied, and applied to practical testing problems. The model is described as *compensatory* due to the additive nature of the logit (as presented in the next section in Equation 4). This feature makes it possible for an examinee with low ability on one dimension to compensate by having a higher level of ability on the second dimension. In this section, a brief theoretical explanation of the two-parameter compensatory MIRT model is presented. Relationships between the two-parameter unidimensional IRT model and the two-dimensional MIRT model are also described. This section concludes with a brief overview of two other multidimensional IRT models.

Factor Analytic Multidimensional Model for Compensatory Abilities

One way to understand the factor analytic approach to the parameterization of multiple dimensions is to review the work of Christoffersson (1975; also see McDonald, 1997, 1999). He defined a set of unobservable variables, $\boldsymbol{v}$, that follow the multiple common factor model. Thus, item i can be expressed as

$$\boldsymbol{v}_i = \boldsymbol{\lambda}'\mathrm{f} + \delta_i, \tag{1}$$

where $\boldsymbol{\lambda}' = [\lambda_1, \lambda_2, \ldots, \lambda_n]$ is the matrix of common factor loadings, $\mathbf{f}$ is a vector of common factors, and δ_i is the ith unique factor. The model assumes for each item i there is a latent ability $\boldsymbol{v}_i$ that is required to correctly answer the item. This latent ability is assumed to be continuous and normally distributed. If the examinees' proficiency is beyond a given threshold, t_i, they will get the item correct. If not, their effort will result in an incorrect response. For each dichotomous item i, an examinee's response, U_i, can be expressed as:

$$U_i = 1 \text{ if } \boldsymbol{v}_i \geq t_i \tag{2}$$

and

$$U_i = 0 \text{ if } \boldsymbol{v}_i < t_i.$$

The proportion of examinees correctly responding to item i (i.e., the p value or difficulty level) can be expressed as the proportion of area under a normal curve beyond the threshold t_i as

$$p_i = N(t_i), \tag{3}$$

where N denotes the normal ogive function.

This outcome leads to the formulation of the k-dimensional normal ogive model that can be expressed as

$$\begin{aligned} P &= \{U_i = 1 \mid \theta_1 \ldots \theta_k\} = N\{\beta_{i0} + \boldsymbol{\beta}'\boldsymbol{\theta}_i\} \\ &= N\{\beta_{i0} + \beta_{i1}\theta_1 + \beta_{i2}\theta_2 + \cdots \\ &\qquad + \beta_{ik}\theta_k\}, \end{aligned} \tag{4}$$

where

$$\beta_{i0} = \frac{t_i}{\sqrt{\psi_i}} \tag{5}$$

and for the kth dimension

$$\boldsymbol{\beta}_i = \frac{\boldsymbol{\lambda}_i}{\sqrt{\psi_i}}. \tag{6}$$

ψ_i is the explained item variance or 1 minus the communality, given as

$$\psi_i = 1 - \boldsymbol{\lambda}_i'\mathbf{P}\boldsymbol{\lambda}_i, \tag{7}$$

where $\mathbf{P}$ is the covariance matrix of latent abilities. In Equation 4, N is the normal ogive function and θ is the latent ability vector. It is helpful to note that this is a direct extension of Lord's (1980) parameterization of the unidimensional model,

$$\begin{aligned} P\{U_i = 1 \mid \theta\} &= N\{a_i(\theta - b_i)\} \\ &= N\{a_i\theta - a_i b_i\}. \end{aligned} \tag{8}$$

Comparing the two models for the unidimensional case, β_{i0} is analogous to $-a_i b_i$ and thus cannot be interpreted as simply the difficulty or location parameter, b_i. However, β_i does correspond to a_i, the discrimination parameter.

To understand further the link between Lord's parameterization and the factor analytic model, Equations 5 and 6 provide a link between the classical test theory (CTT) difficulty and discrimination parameters (i.e., p value and biserial correlation) with their unidimensional IRT counterparts, the a and b parameters. Specifically, the IRT a and b parameters can be linked to their CTT counterparts by the equations

$$a_i \approx \frac{r_{\text{bis}}}{\sqrt{1 - r_{\text{bis}}^2}} \tag{9}$$

and

$$b_i \approx \frac{z(p_i)}{r_{\text{bis}}}, \tag{10}$$

where r_{bis} is the biserial correlation for item i and $z(p_i)$ is the z value that corresponds to the threshold point for item i. Note that the area under the marginal normal curve that contains the threshold value is equal to the proportion of examinees answering the item correctly (i.e., the p value of the item). Conceptually, Equations 6 and 9 are equivalent, where λ_i, the factor loading, represents the correlation between the item score and factor. The item uniqueness in the denominator of Equation 6 is directly related to the square root of one minus the proportion of explained item variance or $\sqrt{1 - \psi_i}$.

Conceptually, Equation 5 is directly analogous to the negative product of Equations 9 and 10,

$$\begin{aligned} -a_i b_i &= \left[\frac{r_{\text{bis}}}{\sqrt{1 - r_{\text{bis}}^2}}\right]\left[\frac{z(p_i)}{r_{\text{bis}}}\right] \\ &= \frac{z(p_i)}{\sqrt{1 - r_{\text{bis}}^2}}. \end{aligned} \tag{11}$$

In the unidimensional case, the relationship between the p value, biserial correlation, and the item characteristic curve (ICC) can be easily shown graphically. In Figure 1, the top graph represents conditional distributions of the biserial relationship between the latent ability (θ) and the continuous scale of knowledge Y_i for a particular item i. For all examinees with ability θ, there is assumed to be a conditional normal distribution of the latent variable Y that determines how the examinees will answer item i. The probability that examinees at this θ level will get the item

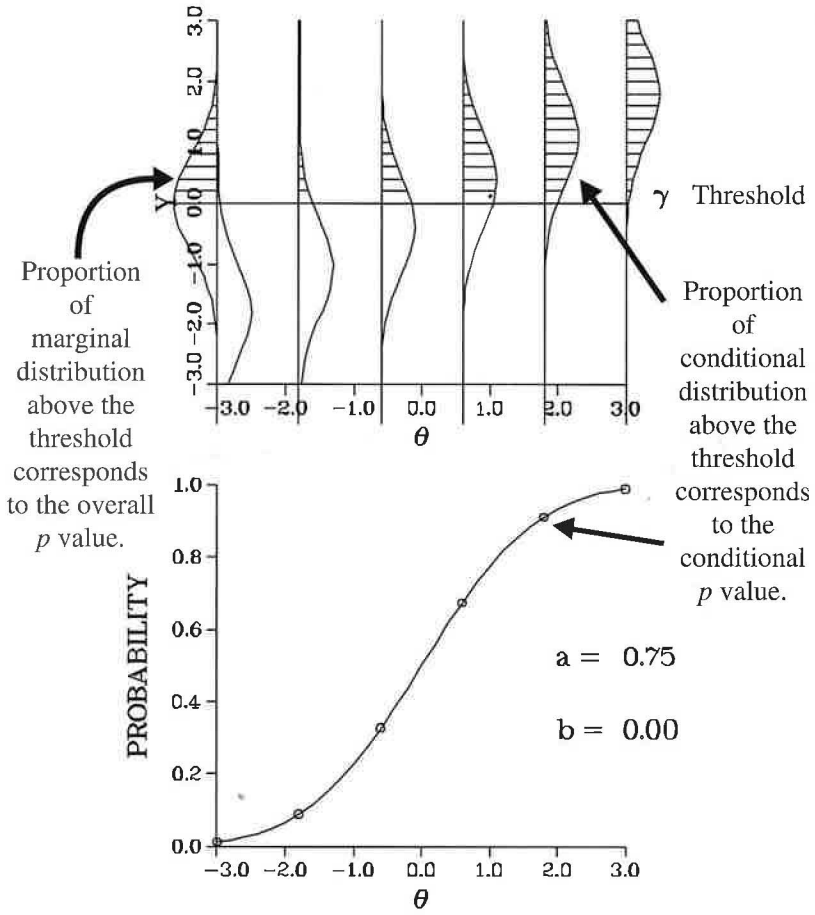


FIGURE 1. The conditional distributions of the biserial relationship between latent ability θ and the continuous scale of knowledge Y_i for item i and the corresponding unidimensional item characteristic curve. (Biserial = 0.6; p value = 0.5.)

correct is equal to the proportion of the conditional distribution that lies above t_i. Thus, at each level of θ, the proportion of the conditional distribution that lies above the threshold t_i corresponds to the probability of correct response. Collectively, these probabilities form the unidimensional ICC. Note also that the proportion of the marginal Y distribution above this threshold corresponds to the p value of the item. An analogous illustration for the two-dimensional case is shown in Figure 2. Although more complicated, the concept is a direct extension of the unidimensional case. Instead of thinking of the threshold as a line, it now is represented as a threshold surface. At each ordered pair of the latent abilities, θ_1, θ_2, the proportion of the normal curve lying above the threshold corresponds to the conditional probability of correct response. When taken in concert, the proportions will form the item response surface.

One advantage of using the factor analytic approach for the multidimensional model is that it allows one to estimate the correlation between factors that underlie test performance. For the two-factor solution, where θ_1 and θ_2 are the two latent abilities, $r_{(\theta_1, \theta_2)} = \phi_{12}$. Then, the underlying composite is scaled so that ψ_i can be expressed as

$$\psi_i = \sqrt{1 - \lambda_1^2 - \lambda_2^2 - 2\lambda_1\lambda_2\phi_{12}}. \quad (12)$$

Many researchers represent the two-dimensional normal ogive model presented in Equation 4 as

$$P\{U_i = 1 | \theta_1, \theta_2\} = \frac{1}{1 + e^{-1.7(a_1\theta_1 + a_2\theta_2 + d)}}, \quad (13)$$

where a_1 corresponds to λ_1, the discrimination parameter for the θ_1 trait; a_2 corresponds to λ_2, the discrimination parameter for the θ_2 trait; d corresponds to a *location* parameter; and 1.7 is the scaling factor to make the logistic and normal ogive models equivalent. [The parameter d is only equivalent to the b parameter (i.e., item difficulty) for the unidimensional case.]

Other MIRT Models

Despite the popularity of the two-parameter compensatory MIRT model, other models exist. For example, Sympson (1978) developed the two-dimensional noncompensatory or partial compensatory MIRT model. This model is described as *noncompensatory* due to the multiplicative nature of the logit. Because the individual component probabilities are multiplicative, the over-all probability of a correct response is bounded in the upper limit by the smallest component probability. Thus, in this model, being high on one ability cannot compensate for being low on the other ability. Spray and Ackerman (1986) also developed a generalized MIRT model that combines the characteristics of the compensatory and noncompensatory models. With the advent of Markov chain Monte Carlo methods, such as Gibbs sampling (Gelfand & Smith, 1990; Geman & Geman, 1984), the item and ability parameters for these models can now be estimated. However, researchers are just beginning to study and apply these models and, hence, much work remains.

Estimating and Representing MIRT Item Parameters

Overview

Although other programs are available for estimating multidimensional item parameters for dichotomously scored data (e.g., TESTFACT; Wilson, Wood, & Gibbons, 1991), NOHARM is a popular choice because it has the capability to perform confirmatory analyses and it is accessed easily. NOHARM is the acronym for the *normal ogive by harmonic analysis robust method.* The program was written by Fraser (1988) to fit the unidimensional and multidimensional normal ogive models of latent trait theory, as presented by McDonald (1967). NOHARM can be downloaded from http://www.niagarac.on.ca/~cfraser/download/. This program uses a nonlinear factor analytic approach (McDonald, 1967) to estimate item parameters in either an exploratory or confirmatory mode. If practitioners or researchers were unsure of the underlying structure, they would use the exploratory mode of NOHARM. If a particular structure is hypothesized, then the confirmatory mode should be used. In the confirmatory mode the user specifies which dimension or dimensions each

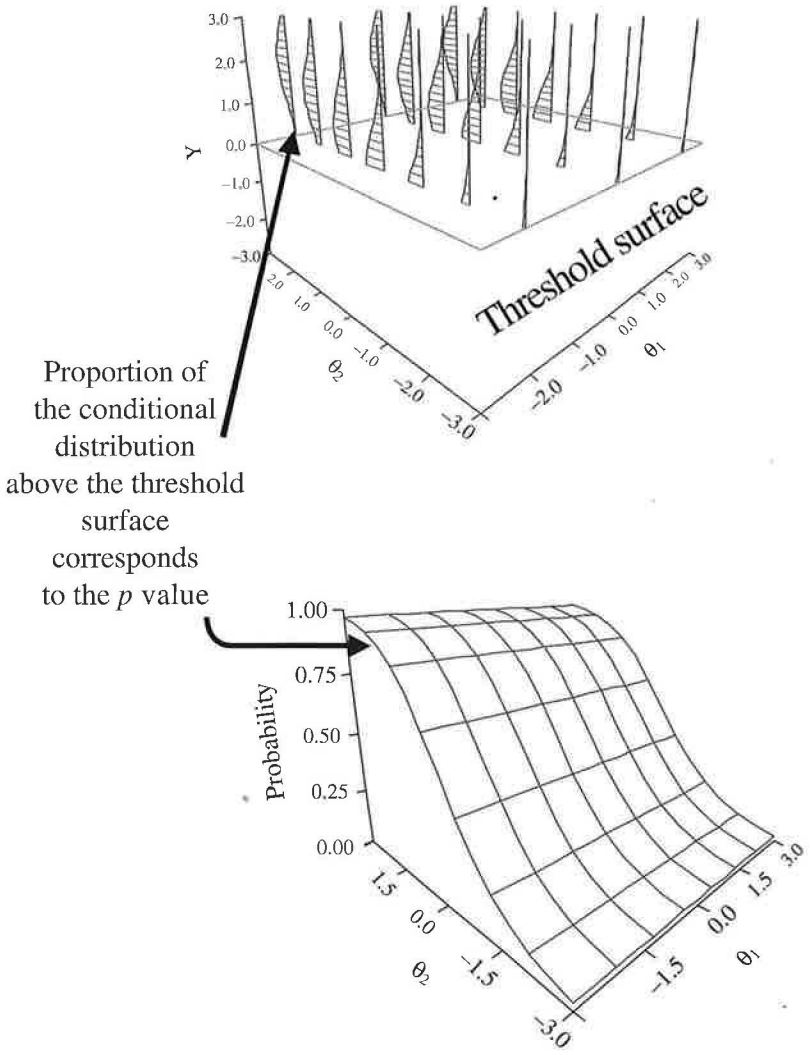


FIGURE 2. A two-dimensional (θ_1, θ_2) item response surface and the corresponding conditional distributions related to a two-dimensional biserial relationship.

item is measuring. NOHARM does not have the capability to estimate the examinees' ability levels.

Estimation of Multidimensional Item Parameters Using NOHARM

The NOHARM program estimates the threshold parameters, β_{i0}, using a closed form expression by solving the sample analog of Equation 4. Discrimination parameters, β_i, are estimated using unweighted least squares minimizing the expression

$$q = \sum_{i \neq j} \left(p_{ij} - \pi_{ij}^{r} \right)^2, \qquad (14)$$

where π_{ij}^{r} is the rth term approximation of a normalized Hermite-Tchebycheff polynomial estimation of the proportion answering items i and j correctly, using a quasi-Newton algorithm. p_{ij} is the proportion of examinees with correct scores on items i and j (McDonald, 1997).

Over the past decade NOHARM has emerged as a popular and widely used program to estimate item parameters for the compensatory multidimensional IRT model. Despite this popularity, only preliminary work by Balassiano and Ackerman (1995) and Ackerman, Kelkar, Neustel, and Simpson (2001) has been conducted to investigate the accuracy of the estimation process used by NOHARM. To our knowledge, no research has been conducted on the accuracy of item parameter estimation in confirmatory mode. As a result, more research is needed.

Using the NOHARM Program for MIRT Analyses

NOHARM Input File. The NOHARM analysis used to fit the two-dimensional model for the 35-item mathematics test is described in this section. The input file is presented in Appendix A. To begin, NOHARM requires a title on line 1. This title cannot be longer than 75 characters. On the second line, the program requires eight integers, separated by a space. These integers specify (a) the number of items to be analyzed (35 in our example), (b) the number of dimensions to be fit (2 in our example), (c) the number of examinees ($n = 6{,}000$), (d) an integer indicating the type of input data (0 for raw data[3] and 1 for a lower triangular product moment correlation matrix), (e) an integer indicating whether the analysis is exploratory or confirmatory (0 for confirmatory and 1 for exploratory), (f) an integer indicating how starting values are set (0 if the program generates the starting value; 1 if starting values are supplied by the user), (g) an integer specifying whether the raw product moment correlation should be printed in the output (0, if yes and 1, if no) and, finally, (h) an integer specifying whether the residual matrix should be printed in the output (again, 0, indicating yes and 1, indicating no).

Next, the lower asymptotes (i.e., the c or "pseudo-guessing" parameter) must be entered. NOHARM fixes the c parameter to these specified values. Even if a two-parameter model is estimated, the user must specify these parameters to be 0 for each item. Lines 3–5 in the current example specify the fixed guessing parameter for each item (0.00 is used for all items in the current example).

When conducting a confirmatory analysis, the user must supply two additional pattern matrices. The first pattern matrix represents whether the factor loadings are to be fixed at their initial value of zero (represented in the matrix by 0), estimated (represented in the matrix by 1), or constrained to be equal to other factor loadings (represented in the matrix by 2). The number of dimensions being fit determines the number of columns in the matrix. The pattern matrix representing how to estimate the factor loadings for the confirmatory analyses conducted in our example are presented on lines 6–40. The first column in the matrix represents the first dimension in our model; mathematical ability. Because all of the items were thought to measure mathematical ability, this column contains all 1's, indicating that factor loadings for this dimension should be estimated for all items. The second column in our matrix represents the second dimension in our model: spatial abil-

ity. Since only the first six items were thought to measure spatial ability, the first six entries in this column are all 1's and the remaining entries are all 0's. This coding indicates that factor loadings on this dimension should only be estimated for the first six items and that the factor loadings on this dimension for the remaining items should be fixed at 0.

The second pattern matrix represents values of the correlation matrix of abilities and is structured in a similar manner to the pattern matrix (i.e., 0 for fixed, 1 for free, and 2 for constrained). The user need only enter the lower triangular portion of this matrix. Also, note that the diagonal elements of this matrix must be fixed at their initial value of 1.0. This is indicated by the 0 elements on the diagonal. The pattern matrix is presented on lines 42 and 43. Typically either the raw data, separated by spaces, or the item correlation matrix would follow at the bottom of the input file. However, due to space constraints, this is not provided in our example.

NOHARM Output File. The NOHARM output file includes a statement of all the input variables including the title, the number of items, dimensions, and examinees, the sample correlation matrix, the fixed guessing parameters, the pattern matrices, and the initial values for item parameter estimation. Only the final multidimensional item parameter estimates are shown in Appendix B. Item location, corresponding to β_{i0} in Equation 5, is called FINAL VECTOR f 0 in the NOHARM output. Item discrimination is called FINAL MATRIX F (coefficients of theta) in the NOHARM output. Item discrimination corresponds to β_i in Equation 6. The correlation between factors is also estimated in the section entitled FINAL MATRIX P in the NOHARM output.

In addition to estimating the item parameters, NOHARM also calculates a residual matrix (i.e., the difference between the observed correlation matrix and the reproduced correlation matrix that would result using the item parameter estimates) and two summaries of the matrix, the sum of squares of residuals and the root mean square of residuals. Goodness-of-fit can then be assessed in at least three different ways. First, a model is deemed to provide good fit to the observed data when the difference between the observed and reproduced correlation matrix is small, producing very small sum of squares of residuals and root mean square of residuals. McDonald (1999) notes that a model provides a sufficiently close approximation to the data if a more complex model cannot be found that is identified (in the statistical sense) and interpretable (in the substantive or psychological sense). McDonald also contends that all decisions produced by evaluating the residual matrix rest on the user and on their substantive knowledge about the test items.

A second method of evaluating the model-data fit is to determine if the multidimensional model provides a better fit to the data than the unidimensional model by comparing the residuals for each of the models, similar to the approach taken in structural equation modeling (e.g., Bollen, 1989; Hayduk, 1987; McDonald, 1999). Typically, however, these residuals are quite small regardless of the model fit. In the example presented in this module, the root mean square of residuals differed by less than 0.01 between the unidimensional and multidimensional models.

A third approach for assessing fit is to compare χ^2 fit statistics for each model based on the residuals (de Champlain & Tang, 1997; Gessaroli & de Champlain, 1996). However, these statistics are also based on the residual matrix and, as a result, the difference in the value of these statistics is comparable to the difference in residuals (i.e., in the example presented in this module, the difference between the χ^2 fit statistics was less than 0.01). Although these three approaches are available, there is little consensus among researchers about how best to assess model-data fit. As a result, research is still needed in this area.

Graphical Representations of Multidimensional Items and Information

In the previous section, techniques used to estimate MIRT item parameters for compensatory abilities were presented. The item characteristics were represented as numeric values. Recall, however, that item response theory has many graphical techniques for representing item and test characteristics. In this section, graphical representations of items are illustrated using the NOHARM two-dimensional solution for the 35-item mathematics achievement test. The plots presented in this section are drawn largely from the work of Ackerman (1996) using the computer program DISSPLA (Computer Associates International, 1989). Some key differences between unidimensional and multidimensional item response theory are also described.

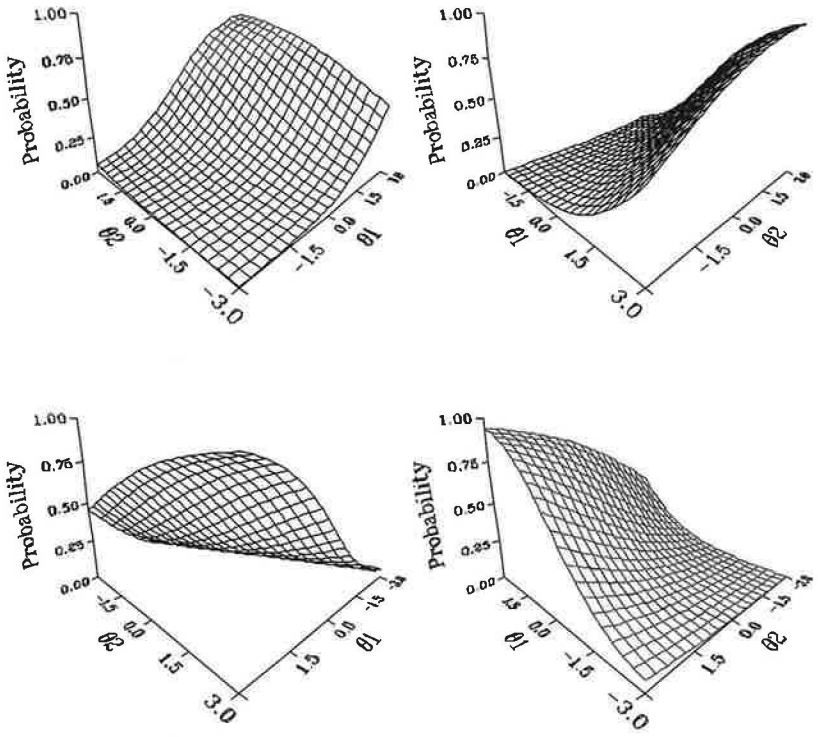


FIGURE 3. Four perspectives of the item characteristic surface for Item 3 from the mathematics subtest.

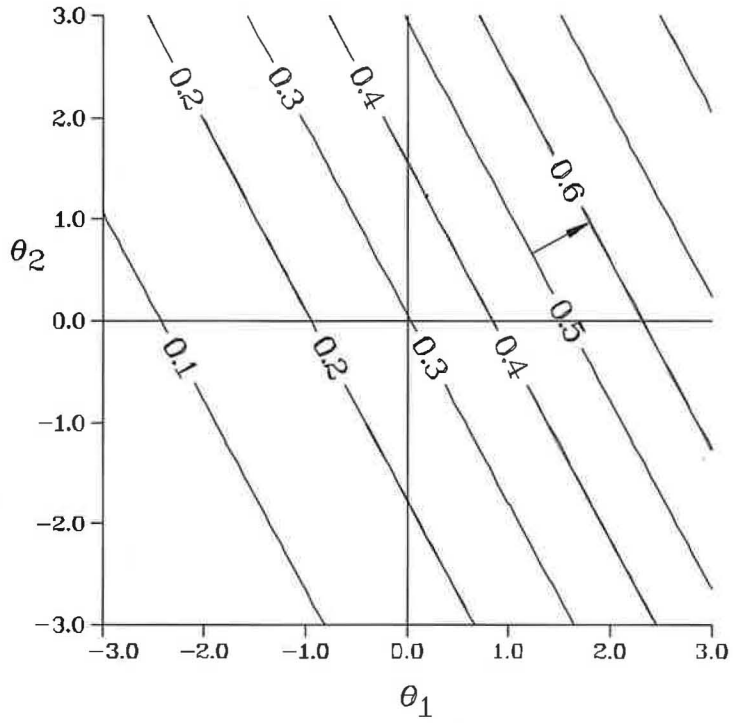


FIGURE 4. An equiprobability contour plot with Item Vector 3 from the mathematics subtest.

Item Representation. To begin, the ICC used in unidimensional IRT represents the conditional probability that an examinee with a given latent ability or θ will correctly answer an item. The item characteristic surface (ICS) used in multidimensional IRT represents the probability that an examinee with a given $\theta_1\theta_2$ composite will correctly answer an item. In other words, the item characteristic *curve* in unidimensional IRT is analogous to the two-dimensional *surface* in multidimensional IRT. The ICS for Item 3 from the 35-item mathematics test ($a_1 = 0.55$, $a_2 = 0.30$, $d = -0.87$) is presented in Figure 3 (the item parameters are shown in Table 1 and in the NOHARM output in Appendix B). The ICS for this item reveals it is quite difficult and moderately discriminating, especially for the θ_1 ability.

A representation more readily interpretable than the ICS is a plot of the item equiprobability contours, as shown in Figure 4. Each contour corresponds to the probability that an examinee with a given $\theta_1\theta_2$ composite will correctly answer an item. Examinees lying on the same contour line will have the same probability of a correct response. For the compensatory model, the contours are equally spaced and parallel across the response surface. The contour lines become closer as the slope of the response surface becomes steeper or more discriminating.

The contour plot is often combined with an item vector plot to further enhance the interpretability of each item. This plot depicts items as vectors in an orthogonal Cartesian coordinate system representing the two-dimensional latent ability space. Note that an orthogonal coordinate system is used merely for clarity so that distance measures and vectors can be easily calculated, understood, and interpreted. This representation does not imply that there is a zero correlation between the two latent abilities. In a vector plot, multidimensional items are graphically depicted based on three characteristics: discrimination, difficulty, and location. Discrimination corresponds to the length of the item response vector (Reckase & McKinley, 1991). This length represents the maximum amount of discrimination, and is referred to as MDISC. For item i, MDISC is given by

$$\text{MDISC} = \sqrt{a_{i1}^2 + a_{i2}^2}, \qquad (15)$$

where a_1 and a_2 are the discrimination parameters. The tail of the vector lies on the $p = .50$ equiprobability contour for the two-parameter compensatory MIRT model. If extended, all vectors would pass through the origin of the latent ability plane. Further, MDISC will always be positive so the item vectors will only be located in the first and third quadrants of the two-dimensional Cartesian coordinate system. MDISC is analogous to the a parameter in unidimensional IRT.

Difficulty corresponds to the location of the vector in space. The signed distance from the origin to the $p = .50$ equiprobability contour, denoted by D, is given by Reckase (1985) as

$$D = \frac{-d_i}{\text{MDISC}}, \qquad (16)$$

where d_i is the location parameter for item i. The sign of this distance indicates the relative difficulty of the item. Items with negative D are relatively easy and are in the third quadrant, whereas items with positive D are relatively hard and are in first quadrant. D is analogous to the b parameter in unidimensional IRT.

Location corresponds to the angular direction of each item relative to the positive θ_1 axis (Reckase & McKinley, 1991). The location[4] of item i is given by

$$\alpha_i = \arccos\frac{a_{i1}}{\text{MDISC}_i}. \qquad (17)$$

Vectors that lie close to the θ_1 axis represent items that primarily measure θ_1, while vectors that lie close to the θ_2 axis represent items that primarily measure θ_2. As vectors approach a location of 45° from the θ_1 and θ_2 axes, they represent

Table 1. Multidimensional Item Parameter Estimates for Six Spatial Items

Item	d	a_1	a_2	MDISC	D	α_i
1	0.45	2.01	0.63	2.11	−0.21	18
2	2.03	1.01	0.64	1.20	−1.70	33
3	−0.87	0.55	0.30	0.63	1.39	29
4	−0.32	0.42	0.16	0.45	0.71	21
5	0.07	0.44	0.15	0.46	−0.15	17
6	0.47	0.91	0.25	0.94	−0.50	15

items that measure a composite of both abilities. Vectors with a location of exactly 45° represent items that measure θ_1 and θ_2 equally well. In other words, a vector with location α_i greater than 45° is a better measure of θ_2 than θ_1, whereas a vector with a location α_i less than 45° is a better measure of θ_1.

By examining the discrimination, difficulty, and location of each item response vector, the degree of similarity in the $\theta_1\theta_2$ composite for all items on the test can be determined. The vector plot for Item 3, first presented in Figure 3, is shown in Figure 4. Notice this item vector lies on a line that passes through the origin and creates an angle α_i with the positive θ_1 axis. The vector originates at, and is graphed orthogonal to, the $p = .50$ equiprobability contour. This item is quite difficult ($d = -0.87, D = 1.39$) and moderately discriminating ($a_1 = 0.55$, $a_2 = 0.30$, MDISC = 0.63), as previously described, but from the vector plot one can see that this item is measuring a composite of θ_1 and θ_2 ($\alpha_1 = 29°$). However, since the item lies closer to the θ_1 axis, it is a better measure of the θ_1 latent ability.

Validity Sector. Vector plots can also be used to compare items and to identify the valid subtest (Shealy & Stout, 1993). The valid subtest contains items that represent the $\theta_1\theta_2$ composite which the test is designed to measure. Ackerman (1992) described the validity sector as a well defined section in the two- dimensional coordinate system containing items from the valid subtest. By specifying the width of the validity sector, practitioners and researchers can define the range of composite skills that the test *should* measure. The width of the sector can vary, depending on the purpose of the test. Items located outside the validity sector are referred to as invalid items. These items can be deleted from the test when it is desirable to produce a more homogeneous measure of the valid subtest and increase internal consistency.

Table 1 contains the item parameter estimates for the six items believed to measure the spatial ability dimension. Figure 5 contains the vector plot for the six spatial items. Due to the model fit, all other items lie directly on the θ_1 axis, and are not presented. Item 3 is the most difficult ($d = -0.87, D = 1.39$) and it is moderately discriminating ($a_1 = 0.55$, $a_2 = 0.30$, MDISC = 0.63). Item 2 is the least difficult ($d = 2.03, D = -1.70$) and it is quite discriminating ($a_1 = 1.01, a_2 = 0.64$, MDISC = 1.20). Item 1 is the most discriminating ($a_1 = 2.01$, $a_2 = 0.63$, MDISC = 2.11) and it is relatively easy ($d = 0.45, D = -0.21$). Item 4 is the least discriminating ($a_1 = 0.42$, $a_2 = 0.16$, MDISC = 0.45) and it is relatively difficult ($d = -0.32, D = 0.71$).

Figure 5 also contains a possible θ_1 validity sector for the math test. In this example, a stringent validity sector was prescribed (i.e., the valid subtest contained only items with an angular direction of 20° or less). This decision resulted in three items that did not lie within the pure θ_1 validity sector. These three items—2, 3, and 4—were associated with the spatial dimension in the hierarchical cluster and DETECT analyses. Alternatively, Items 1, 5, and 6 (which were never clearly associated with spatial ability in the dimensionality analyses) remain with the tolerable range of the ability composite and, therefore lie within the θ_1 validity sector. Clearly, however, all six items measure the $\theta_1\theta_2$ composite to varying degrees.

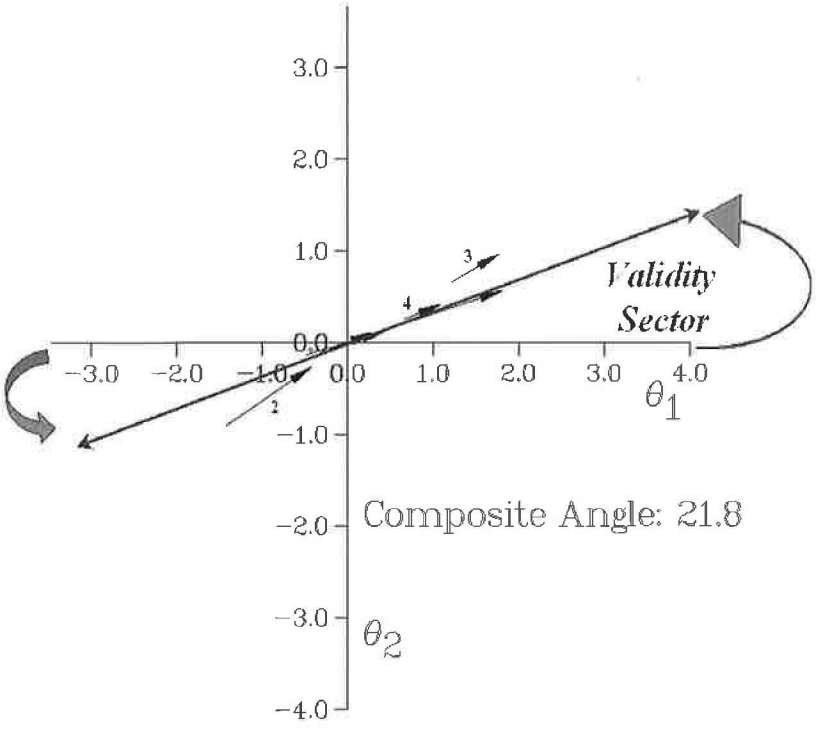


FIGURE 5. The vector plot for the six spatial items from the mathematics subtest.

Information. In item response theory, measurement precision is evaluated using information. The reciprocal of the information function is the asymptotic variance of the maximum likelihood estimate of ability. This relationship implies the larger the information function, the smaller the asymptotic variance and the more measurement precision. Multidimensional information (MINF) serves as one measure of precision. MINF is computed in a manner similar to its unidimensional IRT counterpart except the direction of the information is also considered, as shown in the formula

$$\text{MINF} = P_i(\theta)\left[1 - P_i(\theta)\right]\left(\alpha_{i1}\cos\alpha_{i1} + \alpha_{i2}\cos\alpha_{i2}\right)^2. \quad (18)$$

MINF provides a measure of information at any point on the latent ability plane (i.e., measurement precision relative to the $\theta_1\theta_2$ composite). MINF can be computed at the item level or at the test level (where the test information is the sum of the item information functions).

Reckase and McKinley (1991) developed a *clamshell* plot to represent information with MINF (the representation was said to resemble clamshells, hence the term). To create the clamshells, the amount of information is computed at 49 uniformly spaced points on a 7×7 grid in the $\theta_1\theta_2$ space. At each of the 49 points, the amount of information is computed for 10 different directions or ability composites from 0° to 90° in 10° increments and represented as the length of the 10 lines in each clamshell. Figure 6 con-

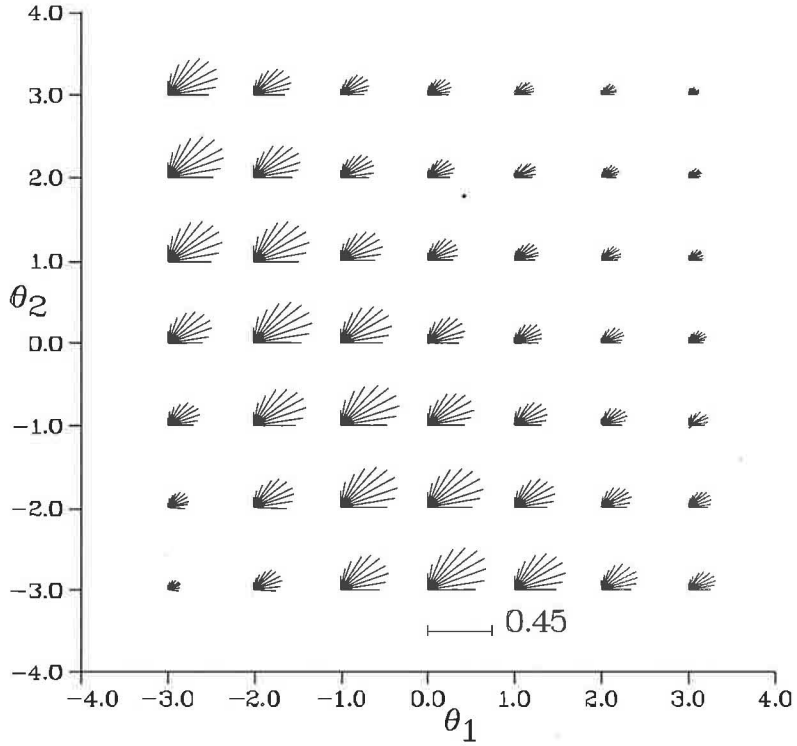


FIGURE 6. The clamshell plot for the six spatial items from the mathematics subtest.

tains the clamshell plot for the three spatial items that lie *outside* the validity sector specified in Figure 5. These three items provide the most measurement precision for the $\theta_1\theta_2$ composite from 28° to 31°, which is beyond the acceptable angular composite of the prescribed θ_1 validity sector.

Ackerman (1996) also developed a *number* plot to represent multidimensional information. To create the numbers, the amount of information is computed in 49 uniformly spaced points on a 7×7 grid in the $\theta_1\theta_2$ space. At each of the 49 points, the direction and amount of information are computed. The direction of maximum information is given as a numeric value on the grid while the amount of information is represented by the size of the font for each numeric value (the larger the font, the greater the information). Figure 7 contains the number plot for the three spatial items *outside* the θ_1 validity sector. These items provide the most measurement precision for the $\theta_1\theta_2$ composite in the angular direction between 28° and 31°.

Future Research Directions in MIRT

Overview

This module would not be complete if we stopped at demonstrating the use of MIRT without discussing some directions for future research. Although many researchers and practitioners believe that educational and psychological tests measure multiple constructs or dimensions, MIRT is still in the early stages of development (Reckase, 1997). As a result, research on MIRT theory and applications of MIRT models will become more prevalent in the future. Four specific areas of future research are immediately apparent to us: (a) test development, (b) diagnostic information based on ability estimation, (c) differential item functioning, and (d) MIRT models for polytomous data and higher dimensional space. Each of these areas is described briefly.

Test Development

Multidimensional IRT analyses can help the practitioner provide evidence that test scores are being properly used and interpreted (Ackerman, 1994, 1996). If the results of a test are reported as a single score, then it is assumed implicitly that all the items are measuring the same skill or same composite of skills. Dimensionality analyses can help establish the degree to which this is true. Response data is an interaction between examinees and items. For some examinees, these data may be unidimensional; for others, multidimensional. Thus, dimensionality analyses should be part of a standard set of analyses conducted after each test administration. Dimensionality needs to be separated into valid, replicable traits and construct irrelevant traits (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). This division can help the practitioner decide if multiple scores should be reported. In addition, if multiple scores are being reported, then vectors plots and information plots can support the constructs presented in table of specifications and provide insight into the relative composite of abilities each item is best measuring.

Diagnostic Information Based on Ability Estimation

What is the effect of using a unidimensional model to estimate student proficiency when the data are actually multidimensional? Although a great deal of research has been conducted on the effects on model misspecification for item parameter estimation, little research has been conducted on the effects of model misspecification for ability parameter estimation. One exception is a study conducted by Walker and Beretvas (in press). They demonstrated that when multidimensional data are analyzed using a unidimensional model, incorrect inferences can be made about student proficiency. Furthermore, these errors were made primarily for those examinees that differed on the secondary dimension and these examinees were more likely to be placed into different proficiency classifications based on the two different models. This outcome likely occurs because difficulty and dimensionality are confounded in the unidimensional estimate of ability, resulting in a multidimensional composite that does not remain consistent throughout the estimated unidimensional ability scale (Reckase, Carlson, Ackerman, & Spray, 1986).

At a time when proficiency classifications in curricular areas are often being used to make high-stakes decisions for students, further research into this area is important. Furthermore, modeling the data in a multidimensional manner, when it is appropriate, allows practitioners to make separate inferences about an examinee for each of the distinct dimensions. This additional information is a valuable asset to anyone who wants to learn more about *why* students are not proficient on a particular construct and

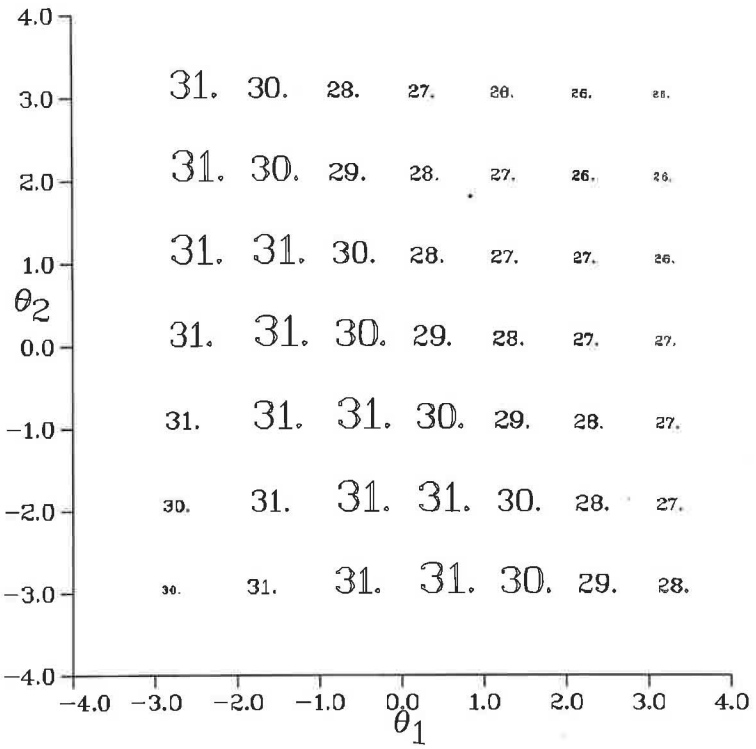


FIGURE 7. The number plot for the six spatial items from the mathematics subtest.

this information could provide teachers with more diagnostic information.

Differential Item Functioning

Differential item functioning (DIF) occurs when examinees from different groups have a different probability or likelihood of answering an item correctly, after conditioning on ability. A number of DIF methods are available, each with different theoretical strengths and well documented empirical support (Camilli & Shepard, 1994; Clauser & Mazor, 1998). Unfortunately, little progress has been made in understanding *why* DIF occurs despite the availability of these methods. To address this problem, Roussos and Stout (1996) proposed the DIF analysis framework. It is a two-stage approach intended to bridge the gap between substantive and statistical analyses. The first stage is a substantive analysis in which DIF hypotheses are generated. Substantive analyses can be guided by test specifications, content reviews, empirical analyses, or psychological analyses (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001). The second stage is a statistical analysis in which the DIF hypotheses are tested. By combining substantive and statistical analyses, researchers can begin to identify and study systematically the sources of DIF.

The Roussos and Stout (1996) framework is based on the multidimensional model for DIF (MMD) proposed by Shealy and Stout (1993). MMD is a theoretical account for how DIF occurs. It is based on the premise that DIF is produced by multidimensionality. The main construct that the test is intended to measure is the primary dimension. DIF items are believed to elicit at least one dimension in addition to the primary dimension (e.g., Ackerman, 1992; Camilli & Shepard, 1994; Kok, 1988; Lord, 1980; McDonald, 1999, 2000; Roussos & Stout, 1996; Shealy & Stout, 1993).

Few studies have been conducted with real test data to support the utility of the Rousso and Stout (1996) framework or, for that matter, of evaluating the claim that multidimensionality produces DIF. Two exceptions are noted. First, Gierl and Khaliq (2001) used the DIF analysis framework to study translation differences on ninth-grade achievement tests where the test was administered in the languages of English and French (i.e., the English test was translated into French). An 11-member committee of testing specialists conducted a substantive analysis of existing DIF items. In their review, four sources of translation DIF were identified. The two translators used these four sources to categorize a new set of DIF items from sixth- and ninth-grade Mathematics and Social Studies achievement tests. Each item was associated with a specific source of translation DIF and each item was anticipated to favor a specific group of examinees. Finally, a statistical analysis was conducted on the items in each category. Results from the statistical analyses revealed that the translators correctly predicted the group that would be favored for many of the DIF items across content areas and grade levels. Second, Walker and Beretvas (2001) showed that DIF occurred for open-ended mathematics items in favor of students who were more proficient in their ability to communicate in writing. Similar to the complex data structure explored in this module, these items were hypothesized to be multidimensional, measuring mathematical communication in addition to general mathematical ability, while the re- maining items were hypothesized to be unidimensional. Further research that explores the presence of DIF in the context of MIRT is needed.

MIRT Models for Polytomous Data and Higher Dimensional Space

Although multidimensional models for polytomously scored data exist (e.g., Kelderman & Rijkes, 1994; van der Linden & Hambleton, 1997), most research in MIRT has been conducted using dichotomously scored data in a two-dimensional space. Indeed, the focus in this module was on dichotomously scored items using two-dimensional data. However, the increased use of open-ended items on large-scale tests makes it more important than ever to explore MIRT models for polytomously scored data. How well the available multidimensional models for polytomously scored data represent the underlying structure in this response format has not yet been fully explored. In fact, research on the accuracy of NOHARM item parameter estimation with dichotomously scored data is far from complete. Furthermore, extending the graphical representations of MIRT items to the polytomous case is not yet a reality, nor is extending the graphical representations to higher dimensional space, although it is possible to model higher dimensional space. One aspect of MIRT that can be used for higher dimensions with polytomous items is related to dimensionality assessment. Specifically, DIMTEST has been modified for use with polytomous data in a program called Poly-DIMTEST (Nadakumar,

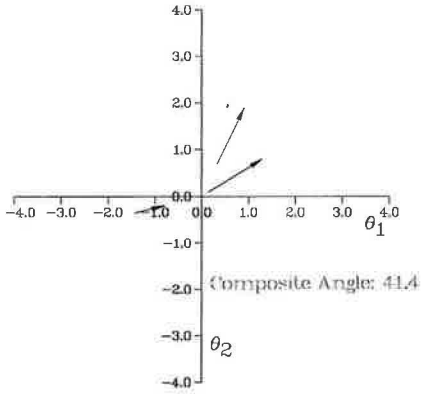
Yu, Li, & Stout, 1998). However, relatively few studies have been conducted to evaluate the accuracy of this program and, as a result, much more research is needed.

Self-Test

1. What is the most important thing to do before conducting a MIRT analysis?
2. What is the difference between complex and simple structure?
3. How can one determine if the data are multidimensional?
4. The most common two-dimensional IRT is called a "compensatory" model. Why do we use this term?
5. A common way to represent items in a two-dimensional latent space is to use a vector. How is the vector drawn, and what characteristics of the item is it representing?
6. When an item is measuring a composite of two identifiable abilities, the information for this item is actually computed for several different composites. Why is this?
7. The a_1, a_2, and d parameter estimates for three items are presented below:

Item 1:	0.9650	0.6350	0.1550
Item 2:	−0.2330	1.1240	0.6940
Item 3:	−0.9980	0.5700	1.1990

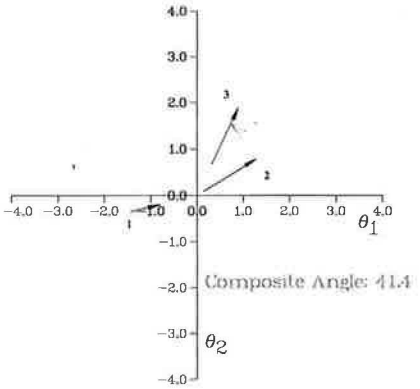
Compute MDISC, D, α_1 in radians, and α_2 in degrees for each item and then identify each item in the vector plot below.



8. What are the two methods for representing information in multidimensional item response theory? How do these methods compare and contrast?

Answers to Self-Test

1. The most important thing to do before conducting an MIRT analysis is to determine whether or not the data are multidimensional. This should include both substantive analyses, based on the item content and the cognitive skills needed to correctly answer items, as well as empirical analyses. If the data are indeed multidimensional, each of the dimensions needs to be clearly identified.
2. Simple structure occurs when all test items measure primarily only one construct, latent trait, or dimension. Complex structure occurs when some items measure a composite of constructs, latent traits, or dimensions, as opposed to primarily one dimension. In both cases, multiple dimensions are being measured by test items.
3. Substantive analyses can be conducted that make use of test specifications, as well as experts who have extensive knowledge of the content and examinees' cognitive processes needed to answer items. If subsets of items appear to be measuring different content knowledge and/or processes, then these sets of items have the potential to represent distinct dimensions underlying the test. Without substantive evidence for multidimensionality, empirical analyses can be conducted that try to find distinct clusters of items, such as HAC or DETECT. With substantive evidence for multidimensionality, DIMTEST can be used to test whether two or more groups of items are dimensionally distinct.
4. We say the model is compensatory because the terms in the exponent (logit) are additive. This means that one can achieve the same probability of a correct response by having a low ability on θ1 and high ability on θ2 or just the opposite, high ability on θ1 and low ability on θ2. Compensation is greatest when an item has equal discrimination parameters, e.g., a_1 =1.0 and a_2 = 1.0.
5. Vectors are drawn on lines which pass through the origin. They only occur in the first and third quadrants because the discrimination parameters are constrained to be greater than or equal to zero. The tail of the vector lies on the p = .5 equiprobability contour. The tail of the vector is located a signed distance (i.e., + = 1st quadrant; − = 3rd quadrant) from the origin. This distance is equal to D. The length of the vector represents the amount of discrimination, MDISC. The angle the vector makes with the θ_1-axis represents the composite of skills that the item is best measuring.
6. Unless one of the discrimination parameters is equal to zero, the item is capable of distinguishing between different levels of multiple composites of ability. When we draw an item vector, we are indicating the direction or composite of θ_1 θ_2 that is being best measured. It is this composite direction that the item is maximally discriminating or providing the most information. The "clamshell" plot illustrates how much information the item is providing for 10 different composites in equal increments from the 0° composite (information about only θ_1) to the 90° composite (information about only θ_2).
7. Item 1: MDISC = 0.6536, D = −1.4763, α_1, = 0.2394, and α_2 = 14
   Item 2: MDISC = 1.3210, D = 0.1764, α_1 = 0.5531, and α_2 = 32
   Item 3: MDISC = 1.3276, D = 0.7517, α_1 = 1.1270, and α_2 = 65



8. Reckase and McKinley (1991) developed a *clamshell* plot to represent information. To create the clamshells, the amount of information is computed at 49 uniformly spaced points on a 7 × 7 grid in the θ_1 θ_2 space. At each of the 49 points, the amount of information is computed for 10 different directions or ability composites from 0° to 90° in 10° increments and represented as the length of the 10 lines in each clamshell. Alternatively, Ackerman (1992) developed a *number* plot to represent multidimensional information. To create the numbers, the amount of information is computed in 49 uniformly spaced points on a 7 × 7 grid in the θ_1 θ_2 space. At each of the 49 points, the direction and amount of information are computed. The direction of maximum information is given as a numeric value on the grid while the amount of information is represented by the size of the font for each numeric value in which the size of the font is proportional to the amount of information.

Notes

The authors contributed equally to the preparation of this manuscript. Therefore, the authors are listed in alphabetical order.

We appreciate the comments and feedback provided by W. Todd Rogers, Richard Luecht, Steve Hunka, our graduate students, and three anonymous reviewers.

[1] Throughout this module the terms *construct, dimension, factor,* and *latent ability* are used interchangeably to refer to the concept or characteristic that a test is designed to measure. In multidimensional IRT, we study tests designed to measure two or more of these concepts or characteristics.

[2] A complete description of the five conditions required to achieve Thurstone's (1947) definition of simple structure are presented in McDonald (1999, pp. 179–180).

[3] If raw data are used, then spaces must be present between item responses or, as an alternative, the user can run the PRODMOM program that accompanies NOHARM to generate the lower triangular correlation matrix.

[4] In some computer programs, like Microsoft EXCEL, the arccosine of a number is expressed in radians (ranging from 0 to π) rather than degrees. Radians can be quickly converted to degrees by remembering that 1 radian = $180/\pi$. For example, 0.30 radians equals 17°.

References

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

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7,* 255–278.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20,* 311–330.

Ackerman, T. A., Kelkar, V., Neustel, S., & Simpson, M. (2001). *A simulation study examining NOHARM's ability to recover two-dimensional generated item parameters.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Balassiano, M., & Ackerman, T. A. (1995). *An evaluation of NOHARM estimation accuracy with a two-dimensional latent space.* Unpublished manuscript.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

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

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31–44.

Computer Associates International, Inc. (1989). *DISSPLA 10.0* [Computer software]. Garden City, NY: Authors.

de Champlain, A. F., & Tang, K. L. (1997). CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model. *Educational and Psychological Measurement 57,* 174–178.

Douglas, J., Kim, H. R., Roussos, L., Stout, W., & Zhang, J. (1999). *LSAT dimensionality analysis for the December 1991, June 1992, and October, 1992, administrations.* (Statistical Report No. 95-05). Newton, PA: Law School Admission Council.

Fraser, C. (1988). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: The University of New England.

Froelich, A. G. (2000). *Assessing the unidimensionality of test items and some asympototics of parametric item response theory.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Department of Statistics.

Froelich, A. G., & Habing, B. (2001). *Refinements of the DIMTEST methodology for testing unidimensionality and local independence.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Gallagher, A. M. (1998). Gender and antecedents of performance in mathematics testing. *Teachers College Record, 100,* 297–314.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75,* 165–190.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85,* 398–409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721–741.

Gessaroli, M. E., & De Champlain, Andre F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement, 33,* 157–179.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20,* 26–36.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38,* 164–187.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10,* 287–302.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8,* 35–41.

Hartz, S., Roussos, L., & Stout, W. (2000). *DIMTEST and HCA/CCPROX: Two conditional covariance based dimensionality assessment tools.* Handout provided at an instructional workshop conducted at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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

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

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20,* 1–14.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL.* Baltimore, MD: The Johns Hopkins University Press.

Kelderman, H., & Rijkes, C. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59,* 149–176.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Department of Statistics.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 263–274). New York: Plenum Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs, No. 15.*

McDonald, R. P. (1997). Normal-ogive multidimensional model. In V. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257–269). New York: Springer.

McDonald, R. P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Erlbaum.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3–31.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28,* 99–117.

Nandakumar, R., & Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18,* 41–68.

Nandakumar, R., Yu, F., Li, H., & Stout, W. (1998). Assessing unidimensionality of polytomous data. *Applied Psychological Measurement, 22,* 99–115.

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

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

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data.* Paper presented at the annual meeting of the Psychometric Society, Toronto, ON, Canada.

Reckase, M. D., & McKinley, R. L. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement, 14,* 361–373.

Roussos, L. (1992). *Hierarchical agglomerative clustering computer programs manual.* Unpublished manuscript, University of Illinois at Urbana-Champaign, Department of Statistics.

Roussos, L., Reese, L., & Harris, V. (1997). *Formulation of DETECT's conditional convariance parameter and evaluation of estimator bias.* Unpublished manuscript, Law School Admission Council, Newton, PA.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355–371.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Spray, J. A., & Ackerman, T. A. (1986, June). *The effect of item response dependency on trait or ability dimensionality.* Paper presented at the annual meeting of the Psychometric Society, Toronto, Canada.

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

Stout, W., Froelich, A. G., & Gao, F. (2001). Using resampling to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, T. A. B. Snijders (Eds.), *Essays on Item Response Theory* (pp. 357–376). New York: Springer-Verlag.

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 D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82–98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago, IL: University of Chicago Press.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38,* 147–163.

Walker, C. M., & Beretvas, S. N. (in press). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement.*

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Chicago, IL: Scientific Software International.

Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 231–249.

Appendix A

Confirmatory Analysis Using Two-Dimensional Mathematics Data

```
35  2  6000  1  0  0  0  0
0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
0.00  0.00  0.00  0.00  0.00  0.00  0.00
1     1
1     1
1     1
1     1
1     1
1     1
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0
1     0

0
1     0
```

RAW CORRELATION MATRIX FROM PRODMOM PROGRAM IS INSERTED HERE

Appendix B

Results

Final vector f0
0.448
2.031
−0.868
−0.321
0.074
0.470
0.322
0.346
1.335
0.225
−1.318
−5.326
1.140
0.758
0.301
−0.545
0.110
0.965
0.734
−1.945
0.878
−0.211
1.673
−0.199
−0.170
−0.453
−0.174
−1.489
0.523
−0.093
−0.278
0.972
−0.481
0.344
−0.344

FINAL MATRIX F (coefficients of theta)

	1	2
1	2.006	0.630
2	1.007	0.639
3	0.549	0.295
4	0.417	0.159
5	0.444	0.150
6	0.905	0.254
7	0.426	0.000
8	1.148	0.000
9	0.566	0.000
10	0.816	0.000
11	2.807	0.000
12	3.633	0.000
13	1.084	0.000
14	0.965	0.000
15	0.548	0.000
16	1.072	0.000
17	1.058	0.000
18	0.957	0.000
19	0.953	0.000
20	2.775	0.000
21	0.516	0.000
22	0.412	0.000
23	0.628	0.000
24	0.787	0.000
25	0.545	0.000
26	0.573	0.000
27	0.689	0.000
28	2.842	0.000
29	0.731	0.000
30	0.631	0.000
31	0.730	0.000
32	1.032	0.000
33	1.178	0.000
34	1.796	0.000
35	1.672	0.000

FINAL MATRIX P (covariances [correlations] of theta)

	1	2
1	1.000	
2	0.841	1.000

RESIDUAL CORRELATION MATRIX IS INSERTED HERE

Sum of squares of residuals (lower off-diagonals) = 2.903305438E-0002
Root mean square of residuals (lower off-diagonals) = 6.985345330E-0003