The Past and Future of Multidimensional Item Response Theory

Mark D. Reckase, ACT

Multidimensional item response theory (MIRT) is a relatively new methodology for modeling the relationships in a matrix of responses to a set of test items. MIRT has been used to help understand the skills required to successfully respond to test items, the extraneous examinee characteristics that affect the probability of response to items (DIF), and the complexities behind equating test forms, among other applications. This paper provides a short introduction to the historical antecedents of MIRT, the initial development of MIRT procedures, the similarities of MIRT procedures to other analysis techniques, and potential future directions for MIRT.
Index terms: dimensionality, factor analysis, item analysis, item response theory, multidimensional item response theory, Rasch model.

Depending on a psychometrician's background, multidimensional item response theory (MIRT) can be considered either as a special case of factor analysis or structural equation modeling, or as an extension of unidimensional item response theory (IRT). These different perspectives on MIRT affect how the methodology is applied and the interpretations of its results. In this paper, the historic underpinnings of MIRT are described from both perspectives, the differences in the perspectives are highlighted, a brief summary is given of the current MIRT technology that is being applied, and trends in the development of MIRT methodology are projected into the future to identify areas for new research.

The summary of current methodology is brief because the other papers in the Special Issue focus on the details of the procedures. Similarly, the section on future trends provides only brief coverage of areas for future research. Some of the authors of the Special Issue are already involved in some of the problems identified in the section on future trends.

Psychological processes have consistently been found to be more complex than they first appear. MIRT is a methodology that shows promise for dealing with such complexity for one class of psychological processes—those used in psychological assessments. MIRT provides a means for analyzing psychological assessment data in such a way that the underlying relationships in the data are made evident. MIRT is simple, yet it can deal with the complex. Hopefully, this paper will help the reader discover the elegant simplicity of the MIRT models and their usefulness for helping to understand psychological processes.

For the purposes of this paper, MIRT will be considered to consist of a general class of models that describe the interaction between a person and a dichotomously scored test item when the characteristics of the person are described using a vector of hypothetical constructs. Further, the characteristics of the test items are described using a set of item parameters and a functional form that relates location in the space defined by the vector of person parameters to the probability of correct response to each item. This conception of MIRT includes not only the linear logistic models proposed by Rasch (1960) and Reckase (1972), but also the multiplicative models proposed by Sympson (1978) and Whitely (1980), and other variations as well.

Applied Psychological Measurement, Vol. 21 No. 1, March 1997, 25–36

Historical Antecedents

Two psychometric research areas provide the foundations for MIRT. The one with the longer history is factor analysis (FA), although the areas of FA that directly relate to MIRT appeared well after the foundational work of Spearman (1927) and Thurstone (1947). The second area of psychometric development that had a clear influence on MIRT is unidimensional IRT. As with FA, the historical antecedents of unidimensional IRT were not the earliest contributions [such as Lazarsfeld (1950)] nor are they more recent contributions.

Factor Analysis

Anyone who reviews the mathematical procedures used for FA and MIRT will notice numerous similarities in methodology. Both methods attempt to define hypothetical scales that can be used to reproduce the data that are the focus of the analysis. Both define scales that have an arbitrary origin and unit of measurement. MIRT differs from most FA representations in that the varying characteristics of the input variables (the items) are considered to be of importance and worthy of study; FA considers the differences in the characteristics of the input variables [such as differences in means, standard deviations (SDs), and reliabilities] as nuisances to be removed from the analysis through statistical standardization. Because of the lack of interest in the characteristics of the variables, most FA texts begin with analysis of a correlation matrix, a data source that ignores all differences in means and SDs in the variables. For example, Harman's (1976) classic work on FA describes the purpose of the analysis as follows:

> The principal concern of factor analysis is the resolution of a set of variables linearly in terms of (usually) a small number of categories or "factors." This resolution can be accomplished by the analysis of the correlation among the variables. (p. 4)

The correlation matrix was considered the main source of data for analysis, and the characteristics of the input variables, such as difficulty and discrimination, were not considered. Similarly, guessing was not considered an estimable parameter related to an item, but rather was considered as another nuisance variable that should be removed through a correction to the correlation (Carroll, 1945).

As with any scientific endeavor, all experts in a field do not agree on approach and basic philosophy. Several FA researchers addressed the problem of identifying hypothetical variables that would reproduce the data from a nontraditional perspective. Five contributors can be identified as having a special place in the factor analytic history of MIRT. These individuals should be considered representatives of a set of contributors rather than being the sole proponents of a point of view. Many others, no doubt, had similar perspectives on FA. These five were selected because they were the most visible, and have continued to be active in advancing their perspective.

Horst. One of the early contributors to the field of FA who foreshadowed the development of MIRT is Paul Horst. In his work on FA, as summarized in *Factor Analysis of Data Matrices* (Horst, 1965), he consistently recommended reproducing the full data matrix, rather than the correlation matrix, from the set of hypothetical variables. He stated:

> It should be observed at the outset that most treatments of factor analysis do not, however, begin with a consideration of the x matrix [the matrix of observed scores] as such and the determination of the u matrix [the matrix of true scores]. These treatments usually begin with correlation matrices derived from the x matrix. This approach has led to much misunderstanding because the analyses applied to the correlation matrix sometimes imply that there is more information in the correlation matrix than in the data matrix x. This can never be the case. For this reason, as well as others, it is much better to focus attention first on the data matrix, as such, in considering the problems and techniques of factor analysis. (p. 96)

By working from the observed score data matrix, Horst had to confront issues related to the characteristics of the variables. He included in his book extended discussions of the issues of origin and unit of measurement and the effects of scaling transformations on FA results. More importantly for this discussion of MIRT, in Horst's work on the FA of binary matrices, he argued against trying to standardize the binary variables. Rather, he suggested partialing out the effects of variation in item difficulty, or as he called it, "dispersion of item preferences" (p. 514). This procedure, which he called "partialing out the simplex" (p. 517), is conceptually similar to estimating the difficulty parameters of the items and using those estimates to model the data.

Although there are many similarities between Horst's work and current conceptions of MIRT, he did not take the step of actually estimating item parameters other than factor loadings, or of modeling probabilities of correct response rather than the actual responses. The emphasis was still on the factors, rather than on item and person characteristics.

Christoffersson and Muthén. Christoffersson (1975) and Muthén (1978) came much closer than Horst to producing a probabilistic model of the relationship between item responses and a vector of person parameters. Both used a normal ogive model to obtain estimates of threshold parameters for items that are essentially the same as the difficulty parameters of a MIRT model. The threshold parameters were the normal deviates that specified the area beneath the normal curve equal to the proportion of incorrect responses to the items.

Christoffersson (1975) stated " ... we see that the model $[P = P^* + \varepsilon]$ expresses the observed proportions P in terms of the threshold levels $h_i, i = 1, 2, \ldots, n$, the loadings Λ, the factor correlation Φ, and a random component ε" (p. 8). P^* is given by

$$P_i^* = P(y_i^* = 1) = \int_{h_i}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left(-x^2/2\right) dx\,, \tag{1}$$

where y_i^* is the response to item i, and h_i is the threshold parameter for item i. The article (Christoffersson, 1975) presented the factor loadings and threshold estimates much as they are in MIRT analyses. The major differences between Christoffersson's presentation and MIRT is that (1) he focused on modeling the hypothetical continuous item trait rather than the probability of correct response; and (2) the probability of a correct response was not presented as a conditional function of item parameters and vectors of person parameters, but rather the probabilities were modeled as population statistics.

Muthén (1978) came closer to a direct representation of MIRT than Christoffersson (1975) by presenting a model for the m-dimensional vector $\mathbf{p}$ of the observed proportions correct. The model was specified as $\mathbf{p} = \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\theta}$ was partitioned into two parts—$(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$, where $\boldsymbol{\theta}_1'$ is an m-dimensional vector of thresholds, and $\boldsymbol{\theta}_2'$ is the vector of elements below the diagonal of the matrix of population tetrachoric correlations. Thus, the probabilities of correct response were modeled, but not conditional on the vector of person parameters.

Both Christoffersson (1975) and Muthén (1978) came very close to presenting the current conception of MIRT. The feature that is missing in their formulations is the conditional probability of correct response to each item as a function of a person's location in the θ space.

McDonald. The nonlinear FA methodology proposed by McDonald (1967) is probably the presentation of FA that is most similar to current conceptions of MIRT. McDonald addressed the problem of factor analyzing variables that are scored as 0 or 1. He indicated that the concern about difficulty factors in the analysis of dichotomous data can be dealt with easily if the regression of the observed data on the factors is allowed to be nonlinear.

In his solution to the problem of factor analyzing matrices of dichotomous data, McDonald clearly specified the important concept of local independence as the basis for the analysis of test items. "The general form of the principle of local independence is that

$$h(\mathbf{y}|\boldsymbol{\Psi}) = \prod_{i=1}^{n} h_i(y_i|\boldsymbol{\Psi}) \text{ ," (p. 16)} \qquad [(2)]$$

where

y is the vector of observed responses,

$\boldsymbol{\Psi}$ is the vector of latent "characterizations," and

$h(\mathbf{y}|\boldsymbol{\Psi})$ is the conditional density of the observed response vector given the latent "characterizations."

McDonald also presented the basic form of the item response surface as the regression of the item response on the trait space and indicated the relationship of the regression function to the conditional probability of correct response

$$p_i|\boldsymbol{\Psi} = \mathrm{E}(y_i|\boldsymbol{\Psi}) \ , \qquad (3)$$

where p_i is the probability of correct response to item i.

The distinctions between McDonald (1967) and current MIRT is that he used a polynomial model that could result in values for the dichotomous variable beyond the 0 to 1 range, and he did not provide any interpretation for the characteristics of the item variables. The focus was solely on estimating the factors, rather than understanding the characteristics of the variables or the interaction of the two. More recent work by McDonald (1985) made the relationship between FA and MIRT very clear. In fact, McDonald indicated that FA is a special case of IRT. "The view taken here is that common factor analysis is really a special case of latent trait theory, based on the principle of local independence, but one in which for convenience only the weak, zero-partial-correlation version of the principle is typically tested" (p. 203).

Bock and Aitkin. For all practical purposes, the methodology presented in Bock & Aitkin (1981) provided the convergence of ideas between IRT and FA and resulted in MIRT. They defined a normal ogive model for a multidimensional trait space that included characterizations of the items in terms of both FA and IRT representations. They even presented the results of an analysis of a section of the Law School Aptitude Test in terms of item difficulty and discrimination, labeled as intercepts and slopes. The only thing lacking from a full MIRT conceptualization was the interpretation of the item parameters as descriptive measures of the interaction between persons and items. The item parameters were still interpreted in the FA sense as a means for labeling the factors.

Bock & Aitkin (1981) presented a two-dimensional extension of the two-parameter normal ogive model along with a MIRT-type parameterization of the item characteristics. The expression for the model is given by

$$P(x_{ij} = 1|\theta_{i1}, \theta_{i2}) = \frac{1}{2\pi^{1/2}} \int_{-z_j(\theta_i)}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt \ , \qquad (4)$$

where

$$z_j(\theta_i) = d_i + a_{j1}\theta_{i1} + a_{j2}\theta_{i2} \ . \qquad (5)$$

The a, θ, and c parameters are equivalent to the parameterization used in current MIRT representations, where $\mathbf{a}$ is a vector of discrimination parameters, $\boldsymbol{\theta}$ is a vector of traits, and d is the difficulty parameter.

Although the model presented by Bock & Aitkin (1981) was essentially a MIRT model, its major use was as a FA model as shown in Bock, Gibbons, & Muraki (1988). In that article, the emphasis was still on defining factors rather than investigating the interaction of persons and items.

Item Response Theory

The focus of IRT is quite different than that of FA. Rather than trying to determine the minimum number of factors that will reproduce the data in an item response matrix, IRT models the relationship between persons and test items. Lord (1980) expressed the goal of IRT this way:

> We need to describe the items by item parameters and the examinees by examinee parameters in such a way that we can predict probabilistically the response of any examinee to any item, even if similar examinees have never taken similar items before. (p. 11)

Early work in IRT was based on the assumption that the parameter that described the examinees varied on only one dimension (Lord & Novick, 1968; Rasch, 1960), but it quickly became apparent that this assumption was often violated, and extensive research continues to determine the consequences of violating the assumption (e.g., Camilli, Wang, & Fesq, 1995). Some early attempts to consider multidimensionality in IRT are presented here, followed by a direct comparison of the IRT perspective with the FA perspective.

Rasch. The early work of Rasch (1960) dealt only with unidimensional IRT models. However, in 1962 he presented a generalization of his early model that included the possibility that the trait of the examinee could be represented by a vector rather than by a scalar (Rasch, 1962). The general model is given by

$$P\left(x|\boldsymbol{\theta}_j,\boldsymbol{\sigma}_i\right) = \frac{1}{\gamma\left(\boldsymbol{\theta}_j,\boldsymbol{\sigma}_i\right)} \exp\left[\boldsymbol{\phi}(x)'\boldsymbol{\theta}_j + \boldsymbol{\psi}(x)'\boldsymbol{\sigma}_i + \boldsymbol{\theta}_j'\boldsymbol{\chi}(x)\boldsymbol{\sigma}_i + \rho(x)\right] , \tag{6}$$

where $\boldsymbol{\phi}$, $\boldsymbol{\psi}$, $\boldsymbol{\chi}$, and ρ are functions of the score on the item, x, so that observable sufficient statistics will be available for the person and item parameters, and $\boldsymbol{\sigma}$ is the vector of item parameters. The function, $\gamma(\ )$, is a normalizing function that insures that the values of the IRT function range between 0 and 1.

Although this general Rasch model allows the examinee's trait levels to be specified by a vector of parameters, obtaining an estimate of the elements of the vector while still maintaining the properties of the Rasch model has been challenging. To maintain the sufficient statistics characteristic of the Rasch model that allow the person parameters and item parameters to be estimated independently, the elements of the $\boldsymbol{\phi}$, $\boldsymbol{\psi}$, and $\boldsymbol{\chi}$ scoring vectors must be known and not estimated from the item response data. For the unidimensional case, this scoring function yields either a 0 or a 1 regardless of the characteristics of the items. As a result, the number-correct score is a sufficient statistic for θ.

If there are two distinct θ dimensions for a set of item response data, a score of u_1 could be assigned for a correct response to an item for Dimension 1 and u_2 for Dimension 2. Then the sufficient statistic for θ_1 would be the number of correct responses (n) times the score, or nu_1. Similarly, the sufficient statistic for θ_2 would be nu_2. However, because u_1 and u_2 are constants across items, the scores on each dimension are functions of n only. Therefore, the two θ estimates will be perfectly correlated. That is, the model is still unidimensional.

To circumvent this problem, researchers have either created items with more than two score categories (1) by treating sets of dichotomous items as a single polytomous item (Reckase, 1972) or through analysis of polytomous items (Kelderman, 1994), thus allowing more complex scoring functions; or (2) by providing a different scoring function for each item prior to the Rasch analysis based on a logical analysis of item characteristics (Glas, 1992). Minimal use has been made of the multidimensional version of the Rasch model because of the complexity of the procedures and concerns about the accuracy with which the scoring functions could be specified.

Lord and Novick. The basic requirements for a MIRT model were presented in chapter 16 of Lord & Novick (1968), although a complete MIRT model was not presented. The chapter included the definition of the complete latent space and the assumption of local independence.

> Local independence means that within any group of examinees all characterized by the same values $\theta_1, \theta_2, \ldots, \theta_k$, the (conditional) distribution of the item scores are all independent of each other (p. 361)

where k is the number of dimensions. The vector $\boldsymbol{\theta}$, therefore, defines the complete latent space.

Lord & Novick (1968) also presented the relationship between the unidimensional normal ogive IRT model and the common factor model. It is notable, however, that the major part of chapter 16 was a discussion of the meaning of the item parameters and their use for solving practical testing problems. The focus was not on the meaning of the factors resulting from the common factor model.

Samejima. Another early presentation of MIRT was given by Samejima (1974). Because IRT formulations typically assume that dichotomous or polytomous responses are the result of subdividing a continuous response variable, Samejima developed a MIRT model for items that have a continuous response, z_i. Samejima's model is given by

$$P_{z_i}(\theta) = 2\pi^{-1/2} \int_{-\infty}^{a_i(\theta - b_i)} \exp\left(-\frac{u^2}{2}\right) du \,. \tag{7}$$

Although the Samejima (1974) model was clearly one of the first formal presentations of a MIRT model, with the exception of work by Bejar (1977) it seems that the model has not been applied. This is probably because item responses that can be considered continuous are rare in psychological and educational tests. Perhaps with the increased use of performance assessment, Samejima's model will find wider applications.

Comparison of the FA and IRT Approaches

The statistical formulation of FA and MIRT approaches to the analysis of matrices of dichotomous item responses are virtually identical, as shown by a comparison of the models presented by Bock & Aitkin (1981), Samejima (1974), and McDonald (1967). In fact, the software for the full-information FA methodology as presented in Bock et al. (1988) can be used for either FA or MIRT. Further, McKinley (1989) developed a program for MIRT analysis that has many of the features of confirmatory FA.

If the statistical procedures are virtually identical, what are the differences in the two types of methodologies? First, exploratory FA has historically been used to find a relatively small number of hypothetical variables that explain the relationships in a large number of empirical variables. That is, exploratory FA is basically a data reduction technique. McDonald (1985) stated it this way: "In exploratory work, then, we follow Thurstone and regard the best number of common factors as the smallest number that will account for the correlations" (pp. 51–52). MIRT, however, focuses on accurately modeling the interaction between persons and items. Although a simple representation of this interaction is valued, the goal is to gain an understanding of the person and item characteristics that influence the form of this interaction. Work by Reckase & Hirsch (1991) indicated that understanding the interaction will not be degraded by using too many dimensions, but it might be by using too few. Therefore, MIRT is generally not a data reduction technique, but rather is a technique for determining stable features of both persons and items that influence responses to test items.

Second, FA typically focuses on some of the characteristics of the input variables, mainly the correlation or variance/covariance matrices, but ignores other characteristics, such as the mean and the form of the relationship between variables. When the correlation matrix is analyzed, it is implied that the mean and SD of the variables are of little or no consequence. MIRT, however, does not use standardized variables, and actively uses the mean and SD of the item responses as represented by the difficulty and discrimination parameters. The estimation and interpretation of these parameters are a critical component of MIRT analyses.

A closely related issue to the data that are used for the analysis is the metric used to describe the solution. FA has traditionally assumed the *z*-score metric so that derivations would be simplified. The use of the correlation matrix as the basis for analysis is consistent with linearly transforming observed variables to *z* scores. If covariances are used, the linearly transformed number-correct score metric is still used, but it is transformed to have a mean of 0.0. MIRT defines the metric of analysis through the selection of the mathematical form of the item response function. This mathematical form defines the characteristics of the θ space.

Another difference is in the way that goodness of fit to the underlying hypothetical model is considered. MIRT procedures model interactions between persons and test items. Ultimately, the goal is to accurately reproduce the probability of correct response to an item for individuals at a particular point in the θ space. There is concern if a single item is not modeled well, or if there is a discrepancy in the predicted probabilities for a particular range of abilities. Conditional measures of fit for single test items has been an active area of research for some time under the labels of person fit (e.g., Liou & Chang, 1992) and appropriateness measurement (e.g., Drasgow, Levine, & McLaughlin, 1991).

FA views fit as a global measure of the ability of the hypothesized model to reproduce a variance/covariance matrix for a group of individuals, rather than for single variables and selected subgroups of individuals. The focus is on group summary measures rather than on conditional measures of fit. A good example of the global approach is presented by Reise, Widaman, & Pugh (1993). Even when using IRT methods, they developed a global measure of fit following the FA philosophy. They stated that, for IRT:

> Typically, fit is assessed at the item level by a statistic that tests the congruence between the proportion of item responses in a particular category predicted from an [item response function] and the proportion in a particular category observed in the data. We did not use any of these item-fit statistics. Rather, we adopted a model-testing approach (see Thissen, Steinberg, & Gerrard, 1986) to maintain consistency between the IRT and [confirmatory factor analysis] sections of the research. (p. 558)

Inadvertently, they emphasized the differences in philosophy between MIRT and FA.

Finally, MIRT analyses actively seek solutions that use the same latent space across tests and samples. The goal is to keep a common scale for all analyses so that items can be banked and used for fixed-form test construction and adaptive tests. Although procedures such as Procrustes rotations and coefficients of congruence have been developed for FA to match factor spaces, these receive much less emphasis than in the MIRT literature. These points will be emphasized below when applications of MIRT are discussed.

Early MIRT Development

In the late 1970s and early 1980s, a number of researchers were actively working to develop practical MIRT models. In addition to the work by Reckase (1972) on the multidimensional Rasch model, Mulaik (1972), Sympson (1978), and Whitely (1980) presented interesting models for the interaction of persons and items.

Mulaik's (1972) model is

$$P\left(x_{ij}|\theta_j,\sigma_i\right)=\frac{\sum_{k=1}^{m}\exp\left(\theta_{jk}+\sigma_{ik}\right)x_{ij}}{1+\sum_{k=1}^{m}\exp\left(\theta_{jk}+\sigma_{ik}\right)}\,, \tag{8}$$

where $x_{ij}=0,1$. This model has the interesting property that, for fixed values of the exponents, the probability of correct response increases as the number of dimensions increases. If all of the exponents are 0, the probability of correct response is $m/(m+1)$. If the parameters are to have constant interpretation, this property implies that the item parameters would have to be rescaled if the number of dimensions used to model the item response data is changed.

A model with the same property, but with the relationship in the opposite direction, was proposed by Sympson (1978) and Whitely (1980). For this model, for fixed values of the exponents the probability of correct response decreases with an increase in the number of dimensions. This model is given by

$$P\left(x_{ij}=1|\boldsymbol{\theta}_j,\mathbf{a}_i,\mathbf{b}_i,c_i\right)=c_i+(1-c_i)\prod_{k=1}^{m}\frac{\exp\left[a_{ik}\left(\theta_{jk}-b_{ik}\right)\right]}{1+\exp\left[a_{ik}\left(\theta_{jk}-b_{ik}\right)\right]}\,, \tag{9}$$

where $\mathbf{a}_i$ is a vector of discrimination parameters, and $\mathbf{b}_i$ is a vector of difficulty parameters.

When the exponent for all terms is 0, the probability of correct response is $c_i+[(1-c_i)(.5)^m]$. As m increases, this expression converges to c_i. Models of this form are sometimes called *partially compensatory* or *noncompensatory* because an increase in θ on one dimension cannot overcome a deficit on another dimension. The upper limit on the probability of correct response is determined by the smallest of the product terms.

McKinley & Reckase (1982) considered many of the variations of the general Rasch model and eventually settled on the current form of the linear logistic model as the most practical. Their model was originally presented as

$$P\left(x_{ij}=1|\boldsymbol{\theta}_j,\mathbf{a}_i,d_i\right)=\frac{\exp\left(\sum_{k=1}^{m}a_{ik}\theta_{jk}+d_i\right)}{1+\exp\left(\sum_{k=1}^{m}a_{ik}\theta_{jk}+d_i\right)}\,. \tag{10}$$

This model was labeled as a multivariate extension of the two-parameter logistic model. In contrast to the partially compensatory model presented in Equation 9, this model has been labeled a *compensatory* model because a low θ value on any dimension can be compensated for by a high θ value on another dimension. A probability of correct response of 1.0 can be obtained even with very low θ values on some dimensions by having high θ values on other dimensions. Spray, Davey, Reckase, Ackerman, & Carlson (1990) developed a generalized model that includes both the compensatory and partially compensatory models as special cases, but this generalized model has never been used to analyze real item response data.

Estimation Procedures

The best of models will receive little attention unless a sound procedure exists to estimate the parameters of the model. The Mulaik (1972) and Spray et al. (1990) models have slipped into obscurity because practical estimation procedures were never developed for them. Developing estimation procedures for MIRT models is challenging because observable sufficient statistics do not typically exist for the parameters, and the person and item parameters cannot be estimated independently. Because parameter estimation in FA and structural equations modeling has similar problems, the estimation methodologies that are now being used have borrowed heavily from those approaches.

The programs that are used most commonly for MIRT estimation are NOHARM (Fraser, 1988) and TESTFACT (Wilson, Wood, & Gibbons, 1987). Although the statistical underpinnings of these two programs are somewhat different (marginal maximum likelihood versus least squares), they yield similar results (Miller, 1991). With sample sizes over 1,000 and fairly long tests, these programs have been found to give stable parameter estimates that can be used for a number of applications.

Applications of MIRT

Since the early developments of MIRT, descriptive statistics have been developed to assist in the interpretation of MIRT analyses (Reckase, 1985; Reckase & McKinley, 1991), and numerous applications have been made of the methodology. A few examples will be described here to indicate the span and type of applications.

With the focus on the interaction of the characteristics of the test items and the persons, MIRT has been used very successfully for investigating the detailed structure of skills needed to respond to test items. Miller & Hirsch (1992), for example, were able to identify small but replicable clusters of items with clear substantive meaning. Although the percent of variance accounted for by these items was less than that typically considered as important in FA research, performance on these clusters of items was important for subsamples of the examinee population, such as scholarship candidates.

The MIRT conception of dimensions of item sensitivity and locations of subgroups in a multidimensional latent space has provided a natural means for describing differential item functioning (DIF) of test items. For example, Ackerman (1992) provided a clear representation of DIF from a multidimensional perspective.

MIRT procedures have also been used to support item selection for a test so that a unidimensional IRT model can be used to describe the item response matrix. Reckase, Ackerman, & Carlson (1988) showed how items that were sensitive to differences in abilities on multiple dimensions could be selected for a test and the result could still be modeled using a unidimensional IRT model.

As a final example, Davey & Oshima (1994) considered the difficult problem of linking together multidimensional calibrations. This work is the forerunner of equating methodology for use on tests that assess multiple skills.

Future Directions in Test and Item Analysis

Relative to other areas of psychometrics, MIRT is still in its infancy. Numerous problems still need to be addressed and methodology needs to be developed to help address the problems.

MIRT has made it clear that items and tests are much more complex than initial psychometric procedures indicated. The simple unidimensional models may not be sufficient for describing the interaction between persons and items. More complex models than those currently being used may be needed.

Some initial work has been done to understand the effects on total test score of including items measuring multiple dimensions on a test (Reckase, 1989), but more work is needed. In particular, procedures are needed to identify the skills that are assessed at various points on a score scale, and other procedures are needed to construct tests that assess skills in the same way. That is, test parallelism should be considered a multidimensional concept rather than a unidimensional one.

With the increased use of polytomous items, a multidimensional version of polytomous IRT models is needed. Muraki & Bejar (1995) have done some work in this area, but much more is needed. For example, when graded response items are used, do the different points on the item score scale represent different combinations of skills? For example, do the lower points on the scoring guidelines for writing assessments focus on basic literacy and the upper points focus on logic, organization, and style? If so, how can these changes in focus be modeled?

Following the work of Davey & Oshima (1994), methods need to be developed for linking multidimensional calibrations so item pools can be developed and multidimensional test equating can be performed. This is an area in great need of research because most of the educational tests currently being used probably assess multiple traits.

A closely related area in need of additional research is the requirements for estimating the parameters of the MIRT model. Although good procedures currently exist (e.g., TESTFACT, NOHARM), little is known about the data requirements needed to support defining high-dimensional spaces. How many items need to tap a dimension before it can be identified? What is the relationship between sample size, the heterogeneity of the examinee population, and the number of dimensions that can be identified? What does it mean to say that two dimensions are highly correlated but distinct? Estimation is a very rich area for further research.

Finally, reporting the results of tests as points in a multidimensional space is a challenge that has yet to be addressed. It seems useful to consider examinees as having vectors of traits that change with instruction and maturation. If those changes are followed over time, what types of patterns of growth result? Growth curves are typically presented based on one dimension. No doubt that is a gross simplification. To better understand the educational process, much more detailed ways of describing examinee performance are needed. MIRT should provide the technology for these necessary descriptions.

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.

Bejar, I. I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement, 1,* 509–521.

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

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

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 32,* 79–96.

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

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

Davey, T. C., & Oshima, T. C. (1994, April). *Linking multidimensional item calibrations.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171–191.

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

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1, pp. 236–258). Norwood NJ: Ablex.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed., revised). Chicago: University of Chicago Press.

Horst, P. (1965). *Factor analysis of data matrices.*

New York: Holt, Rinehart and Winston.

Kelderman, H. (1994). Objective measurement with multidimensional polytomous latent trait models. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 235–243). Norwood NJ: Ablex.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer (Ed.), *Measurement and prediction* (pp. 362–412). Princeton NJ: Princeton University Press.

Liou, M., & Chang, C. (1992). Constructing the exact significance level for a person fit statistic. *Psychometrika, 57,* 169–181.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs,* No. 15, 1–167.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale NJ: Erlbaum.

McKinley, R. L. (1989). *Confirmatory analysis of test structure using multidimensional item response theory* (RR-99-31). Princeton NJ: Educational Testing Service.

McKinley, R. L., & Reckase, M. D. (1982). *The use of the general Rasch model with multidimensional item response data* (Research Report ONR 82-1). Iowa City IA: American College Testing.

Miller, T. R. (1991). *Empirical estimation of standard errors of compensatory MIRT model parameters obtained from the NOHARM estimation program* (Research Report ONR 91-2). Iowa City IA: American College Testing.

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

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

Muraki, E., & Bejar, I. I. (1995, July). *Full information factor analysis of polytomous item responses of NCARB exams.* Paper presented at the European meeting of the Psychometric Society, Leiden, The Netherlands.

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

Rasch, G. (1960/80). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980), with foreword and afterword by B. D. Wright. Chicago, The University of Chicago Press.

Rasch, G. (1962). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4,* 321–334.

Reckase, M. D. (1972). *Development and application of a multivariate logistic latent trait model.* Unpublished doctoral dissertation, Syracuse University, Syracuse NY.

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

Reckase, M. D. (1989, August). *Controlling the psychometric snake: Or, how I learned to love multidimensionality.* Invited address at the meeting of the American Psychological Association, New Orleans.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25,* 193–204.

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

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

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552- 566.

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

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Spray, J. A., Davey, T. C., Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1990). *Comparison of two logistic multidimensional item response theory models* (Research Report ONR 90-8). Iowa City IA: American College Testing.

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: University of Chicago Press.

Whitely, S. E. (1980). *Measuring aptitude processes with multicomponent latent trait models* (Technical Report No. NIE-80-5). Lawrence: University of Kansas.

Wilson, D., Wood, R., & Gibbons, R. D. (1987). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Mooresville IN: Scientific Software.

Author's Address

Send requests for reprints or further information to Mark Reckase, ACT, 2201 North Dodge Street, Iowa City IA 52243, U.S.A.

Editor's Note

This article is part of the Special Issue on Multidimensional Item Response Theory published as Volume 20 Number 4, December 1996, edited by Terry Ackerman. The article is published in this issue due to space limitations in the December issue.