Studies in Educational Evaluation 35 (2009) 57–63

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Multidimensional IRT models for the assessment of competencies

Johannes Hartig [a,*], Jana Höhler [b]

[a] University of Erfurt, Faculty of Education, Department of Educational Research Methodology, P.O.B. 900221, D-99105 Erfurt, Germany
[b] German Institute for International Educational Research, Department of Educational Quality and Evaluation, Germany

ARTICLE INFO

Keywords:
Educational assessment
Educational measurement
Item response theory
Multidimensional item response theory
Psychometrics

ABSTRACT

Multidimensional item response theory (MIRT) provides an ideal foundation for modeling performance in complex domains, taking into account multiple basic abilities simultaneously, and representing different mixtures of the abilities required for different test items. This article provides a brief overview of different MIRT models, and the substantive implications of their differences for educational assessment. To illustrate the flexibility and benefits of MIRT, three application scenarios are described: to account for unintended multidimensionality when measuring a unidimensional construct, to model latent covariance structures between ability dimensions, and to model interactions of multiple abilities required for solving specific test items. All of these scenarios are illustrated by empirical examples. Finally, the implications of using MIRT models on educational processes are discussed.

Item response theory (IRT) provides a methodological basis for model-based measurement. Item responses are modeled as a function of individual trait levels and item properties. Item difficulties and individual trait levels can be described on a common scale. Today, IRT constitutes a mainstream basis for psychological measurement (Embretson & Reise, 2000). The potential use of multidimensional IRT (MIRT) for educational assessment has been recognized for about two decades now (e.g., Embretson, 1984; Reckase, 1990). MIRT holds the potential to adequately model performance in complex domains, since multiple abilities can be taken into account simultaneously, even with mixtures of abilities required for individual test items. For these reasons, MIRT is a highly interesting methodology for assessing competencies within educational contexts. If competencies are regarded as the disposition to perform successfully in a given real-life context (Koeppen, Hartig, Klieme, & Leutner, 2008), it may be necessary and desirable to model these dispositions as multidimensional constructs. This definition of competencies is applied here, where we understand a *competence* as a relatively broad construct (e.g., foreign language competence) incorporating multiple specific *abilities* (e.g., listening and reading comprehension). *Performance* is used as a more general term referring to observable results from test taking.

This article aims to provide a brief overview of existing MIRT models that can be applied within different typical contexts of educational assessment. These models can deliver useful tools for gaining detailed information that is potentially more fruitful, especially within educational contexts, than the information gained from more traditional, classical measurement models. In the next section, the central features distinguishing different models are illustrated, and their substantive implications for assessment studies are characterized. Subsequently, three prototypical scenarios of MIRT applications in educational assessment are described. For each of these scenarios, empirical examples are given.

1. Central features of MIRT models

MIRT models can be regarded as generalizations of unidimensional IRT models such as the Rasch model, the two-parameter logistic model, and the normal-ogive model (e.g., McDonald, 2000; Reckase, 1997). For illustration purposes, we will restrict most of the examples within this article to models for dichotomous responses with logistic item response functions, with the logit function defined as

$$\text{logit}(y) = \frac{\exp(y)}{1+\exp(y)}. \quad (1)$$

For all models presented in the following sections, there also exist generalizations for polytomous responses and similar models with other mathematical link functions (e.g., normal-ogive). For an overview of IRT models for dichotomous and polytomous responses see, for instance, Embretson and Reise (2000).

While in unidimensional models the probability of successfully answering a test item depends on one underlying ability dimension, in MIRT this probability of success is modeled as a function of multiple ability dimensions. For example, in the unidimensional Rasch model for dichotomous responses, the

* Corresponding author. Tel.: +49 361 7372041; fax: +49 361 7372019.
E-mail address: johannes.hartig@uni-erfurt.de (J. Hartig).

0191-491X/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.stueduc.2009.10.002

probability of person ν to respond correctly to item i is modeled as a function of a single individual ability θ_ν and the item difficulty b_i:

$$\Pr(x = 1|\theta_\nu, b_i) = \text{logit}(\theta_\nu - b_i). \quad (2)$$

In the multidimensional extension, the single ability θ_ν is replaced by a vector θ_ν of multiple abilities. The MIRT model contains an item-specific difficulty parameter b_i and an item-specific loading vector λ_i that defines the relations of item i to the ability dimensions in the model:

$$\Pr(x = 1|\boldsymbol{\theta}_\nu, b_i) = \text{logit}(\boldsymbol{\lambda}_i'\boldsymbol{\theta}_\nu - b_i). \quad (3)$$

While the unidimensional model in Eq. (2) contains one single ability θ, in Eq. (3) performance is modeled as a function of an *ability profile*. Correspondingly, the MIRT measurement model will provide individual ability profiles as test results rather than single scores.

Different MIRT models assume different statistical relations between the ability dimensions and successful performance. Additionally, the pattern of relations between dimensions and items can be defined by a loading matrix with a simple structure (*between-item multidimensionality*) or by a complex loading structure (*within-item multidimensionality*), and thus varies in its complexity. Firstly, we concentrate on this complexity of the relation between latent dimensions and test items. Secondly, compensatory and non-compensatory interactions of multiple dimensions affecting performance within the same item are described. As an additional feature, we focus on the number of latent dimensions within the model. All of these features are considered with regard to their substantive implications for educational assessment.

1.1. Between- and within-item multidimensionality

One important distinction between different MIRT models that is closely related to the design of an assessment is whether the probability of success in every item is affected only by one of the dimensions in the model, or whether responses to one item can be modeled as depending on multiple ability dimensions simultaneously. The former case is denoted as between-item multidimensionality, the latter as within-item multidimensionality (Adams, Wilson, & Wang, 1997). In models with between-item multidimensionality, separate disjunctive clusters of items are used to measure each dimension in the model (independent-cluster structures; cf. McDonald, 2000). In factor analytic terms, these models are characterized by a simple structure of loadings; they can be regarded as a combination of several unidimensional measurement models into one common model. The combination allows relations between the latent ability dimensions to be modeled. In models with within-item multidimensionality, mixtures of multiple abilities are modeled as underlying responses to single items. Fig. 1 gives a schematic illustration of models with between- and within-item multidimensionality.

It is important to note that the choice of the loading structure strongly depends on the intended interpretation of the latent dimensions. Thus, a suggestion as to which model should be preferred against another can only be made with regard to the specific research question especially because different models may be equivalent in terms of their fit to empirical data (e.g., Hartig & Höhler, 2008).

Models that incorporate within-item multidimensionality are suitable for modeling interactions between different abilities and task demands. Here, the probability of solving an item can be modeled as a function of a combination of different dimensions of abilities. Hence, within-item multidimensional models imply explicit assumptions about the abilities required for the different items, which necessitates strong theoretical assumptions. These assumptions also require statements as to whether the combination of required abilities is compensatory or not (see next section for a discussion of this question). Consequently, models with within-item multidimensionality are particularly interesting for modeling performance in complex tasks that cannot be explained by a single ability dimension for each task. These models may be adequate if researchers are interested in the particular abilities contributing to the overall competence for solving specific test items. They allow models of the interaction between test-takers' abilities and test items to be tested, thus fulfilling a key requirement of psychometric models of competence (see Koeppen et al., 2008).

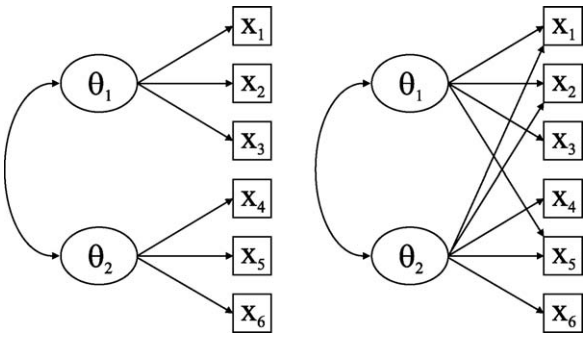

Fig. 1. Schematic illustration of two MIRT models with two latent dimensions θ_1 and θ_2 and six items x_1 to x_6. The model on the left incorporates between-item multidimensionality; each item measures only one of the two dimensions. The model to the right includes within-item multidimensionality; several items (x_1, x_2, and x_5) are affected by both dimensions.

An advantage of models with between-item multidimensionality is that they are less complex than models with within-item multidimensionality, and the latent variables can be easily interpreted. Within these models, estimated scores for the latent variables provide straightforward measures of performance in a specific set of test items. In many cases, these measures will be highly correlated because items draw on the same set of abilities to some extent. However, there is no need to consider the specific interplay or weighting of the different abilities required for solving more complex items. The latent dimensions in the between-item multidimensional model represent the necessary combination of all the abilities required to solve the respective items, regardless of how these abilities need to be integrated. Any overlap is represented in the latent correlations. Hence, if the main research interest is to gain descriptive measures of performance in certain content areas, the between-item model is more suitable than more complex models with within-item multidimensionality.

1.2. Compensatory versus non-compensatory interaction of multiple dimensions

In models with within-item multidimensionality, the multiple dimensions that are required for individual items can be integrated in different ways. The most fundamental feature is whether this integration follows a *compensatory* or *non-compensatory* function. Most MIRT models are compensatory models, meaning a low ability in one dimension can be compensated by a high ability in a second dimension, and vice versa. In a non-compensatory model, the probability of success will only approach one if all abilities required for a particular item are high. The difference between both integrations can be illustrated with two simple models. The item response function (IRF) for a two-dimensional Rasch model with an item loading on both dimensions and an item difficulty of zero can be written as

$$\Pr(x|\theta_1, \theta_2) = \text{logit}(\theta_1 + \theta_2). \quad (4)$$