Contributions to latent variable modeling in educational measurement

Zwitser, R.J.

Chapter 1

Introduction

Through all stages of education, from kindergarten to university, we use tests to quantify what students know or can do. In this thesis, I focus on tests that are designed to measure some sort of *ability*. Examples of such abilities are the ability to read, the ability to write, or the ability to interpret graphs and tables. It is generally accepted that these abilities, sometimes also more generally referred to as *constructs*, cannot directly be measured in a single observation. What can be observed is the response on a single task. Such a task does not represent the construct as a whole, but represents one aspect of the construct. Since one single task does not represent the total construct, tests usually consist of multiple, separately scored tasks, usually called *items*. One of the main questions in educational measurement is the following: how to summarize the item scores into a meaningful final score that represents the ability that is supposed to be measured. This question is prominent in this thesis. In this introduction, I will first define the term construct in more detail. Then, I will elaborate on latent variable models. Finally, I will introduce the main chapters of this thesis.

1.1 The construct

There are different views on what a construct is. The first is based on the so-called *market basket* approach (Mislevy, 1998), where the construct is defined by a (large) set of items. For instance, if one wants to measure the ability to interpret graphs at Grade 6 level, the construct *interpreting graphs* can be defined with a large collection of tasks covering all relevant aspects at the intended level. This should include tasks representing the diversity in types of graphs as well as the diversity in complexity of the figures. If the construct is

defined by a large set of items, then it makes sense to define the final score as a summary statistic on the total set of items, e.g., an estimate of the percentage of tasks that is mastered.

Another view is to consider a construct as a *latent variable* (Lord & Novick, 1968). Since the work of Spearman (1904) and the development of factor analysis, psychologists mostly think of a psychological construct (e.g., intelligence, depression, or introversion) as a trait that cannot directly be observed, but that exists as a common cause that explains the covariance between observed variables. The relationship between observed variables and the latent trait is formalized in the *item response theory* (IRT, Lord, 1980). In IRT, the latent trait is operationalized as a parameter in a *latent variable model*. These models describe the statistical relationship between observations on single tasks and the latent variable, usually denoted by θ. This latent variable approach also became popular in educational testing. The construct is then viewed as a latent variable, and scoring with respect to the construct implies statistical inference about a student's 'θ-value'.

1.2 Latent variable models

In this thesis, I mainly focus on a particular class of latent variable models: the *unidimensional monotone latent variable models*. These models share the following three assumptions. The first is *unidimensionality* (UD), which means that the model contains only one latent variable θ. The second is *local independence* (LI), which means that conditional on θ, item scores are statistically independent. The third is monotonicity (M), which means that there is a monotone, non-decreasing relationship between item scores and the latent variable θ.

Within the class of unidimensional monotone latent variable models, several distinctions can be made. Here, I only describe the distinction between parametric and nonparametric models. In parametric models, the relationship between item scores and θ is described by a parametric *item response function* (IRF). A well-known example is the Rasch Model (Rasch, 1960) for dichotomous items responses, scored with either 0 or 1. This model

is based on the following IRF:

$$P(X_i = 1|\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)},$$

in which $P(X_i = 1|\theta)$ denotes the probability of a score 1, conditional on θ, and b_i denotes a parameter related to item i. The θ parameters are also referred to as person parameters. Other well-known examples are the Two- and the Three-Parameter Logistic Model (Birnbaum, 1968), and the Normal Ogive Model (Lord & Novick, 1968). Nonparametric models put nonparametric restrictions on the IRF $P(X_i|\theta)$. Examples are the *Monotone Homogeneity Model* (MHM, Mokken, 1971), which only assumes UD, LI, and M, and the *Double Monotonicity Model* (Mokken, 1971), which additional to UD, LI, and M, also assumes *invariant item ordering* (IIO):

$$P(X_1 = 1|\theta) \leq P(X_2 = 1|\theta) \leq \cdots \leq P(X_K = 1|\theta),$$

for all θ, and for all K items. The main benefit of these nonparametric models is that they put in general less restrictions on the data, and are therefore more likely to fit the data. A counterpart, however, is that some of the well-known applications of parametric models, such as inference from incomplete data, are limited.

1.3 This thesis

In this thesis, I will describe three studies related to the use of unidimensional monotone latent variable models in educational measurement. I will briefly introduce them in the next three sections.

1.3.1 CML Inference with MST Designs

The first study is about conditional likelihood inference from multistage testing (MST) designs. In MST designs, items are administered in blocks/modules consisting of multiple items. The modules differ in difficulty. Modules are administered to students depending on their responses to earlier modules. The simplest example of a MST is a two stage test (Lord, 1971b). In the first stage, all students take the same first module. This module is often called

the *routing test*. In the second stage, students with a score lower than or equal to c on the routing test take a more easy module, whereas students with a score higher than c on the routing test take a more difficult module. MST is an example of *adaptive testing* (Van der Linden & Glas, 2010), which means that the difficulty level of the test is adapted to the ability level of the student. In order to know which items are easy and which items are difficult, items used in an adaptive tests are usually pretested in a linear, non adaptive, pretest. In such a pretest, item characteristics are determined. Thereafter, the characteristics are assumed to be the same during the adaptive administration. A consequence is that the final score also depends on this assumption about the item characteristics. Therefore, especially in high-stakes testing where test results can have important consequences for the test taker, it is important to check these assumptions after the adaptive administration. This implies that we want to estimate, or at least validate, the parameters of the model from the adaptive test data. In this chapter, I focus on the estimation of item parameters in MST designs.

It is generally known that item and person parameters cannot consistently be estimated simultaneously (Neyman & Scott, 1948). For that reason, the estimation procedure is usually performed in two steps. First, the item parameters are estimated with a conditional likelihood (Andersen, 1973a) or marginal likelihood (Bock & Aitkin, 1981) method. This step is called *calibration*. In the section step, the person parameters are estimated, conditional on the item parameters. For MST designs, it was already described how item parameters can be estimated with the marginal likelihood method (Glas, 1988; Glas, Wainer, & Bradlow, 2000). And it has been claimed that for MST designs the conditional likelihood method can not be used (Glas, 1988; Eggen & Verhelst, 2011; Kubinger, Steinfeld, Reif, & Yanagida, 2012). In this chapter, I will illustrate that also in MST designs item parameters can be estimated with the conditional likelihood method, a method that in some cases is preferable over the marginal likelihood method. This chapter is therefore not directly about the estimation of θ, but about the calibration step that precedes the final scoring. With the item parameters and the data obtained from the MST, the usual methods can be used to obtain the final θ estimates.

1.3.2 The Nonparametric Rasch Model

The second study is about ordering individuals with the sum score. As introduced above, one of main questions in educational measurement is how to summarize item scores into a final score. A criterion with which the use of a particular statistic could be justified, is as follows: if a unidimensional model fits the data and if the model contains a sufficient statistic for the parameter θ, then the sufficient statistic could be used as final score, since the sufficiency property implies that the statistic contains all statistical information about the parameter θ. Within the class of unidimensional monotone latent variable models, both the Rasch model (Rasch, 1960), as well as the One Parameter Logistic Model (Verhelst & Glas, 1995) contain a sufficient statistic for θ. However, it might be that these models do not fit the data, and then the justification argument described above does not hold. In case of a lack of model fit, a nonparametric alternative might be considered. Chapter 3 is about a nonparametric equivalent of the justification criterion described above. Nonparametric models can be used for ordinal inferences. If we want to justify the use of a statistic to order individuals, we must have a statistic that contains all statistical information about the ordering with respect to θ. For the well-known MHM, the use of the sum score has often been justified based on the *stochastic ordering of the latent trait* (SOL) property (see, e.g., Mokken, 1971, and Meijer, Sijtsma, & Smid, 1990). In this chapter, however, we argue that this property is not satisfactory as justification for using sum scores to order individual students. To arrive at a nonparametric model that contains a statistic that keeps all available statistical information about the ordering of θ, or at least does not contradict it, we first define the *ordinal sufficiency* property. Then we take the sum score as an example, and we will introduce a nonparametric model with an ordinal sufficient statistic for the parameter θ: this model is called the *nonparametric Rasch Model*.

1.3.3 DIF in International Surveys

The third study is about final scores in international surveys, especially the *Programme for International Student Assessment* (PISA). A factor that complicates the statistical modeling of surveys is the substantive amount of

differential item functioning (DIF). There is therefore no single model that fits the data in each country. However, this is exactly what PISA assumes: after data cleaning and the elimination of some bad performing items, PISA fits a generalization of the Rasch model in an international calibration (OECD, 2009a), and the person parameters are taken as final score. The last couple of years, the consequences of ignoring DIF in the model have been a topic of debate, and recently a couple of modeling approaches that take DIF into account have been proposed (Kreiner & Christensen 2007; 2013; Oliveri & Von Davier 2011; 2014). In this chapter, we explain that these approaches are not fully satisfactory, and we propose an alternative, DIF-driven modeling approach for international surveys. The core of this approach is that we define the construct as a set of items. Therefore, comparisons with respect to the construct, between different populations, are equivalent to comparisons of the responses to these items. The only aspect that complicates these comparisons is the incomplete data collection design. In this chapter, we illustrate how latent variable models (plural, because different models are used in different countries) are used to get an estimate of the complete data matrix. Since we use different models in different countries, this procedure is very flexible with respect to DIF. With the estimated complete data matrix, all kinds of comparisons between countries can be made. We will illustrate this with real PISA data.

1.4 Note about notation

The research projects that are described in the next three chapters are based on collaboration with some colleagues. Therefore, I write *we* instead of *I*. Furthermore, notation is sometimes not consistent between chapters. However, within chapters we have striven to be consistent and to introduce all notation.