UNIVERSITY OF AMSTERDAM

Contributions to Latent Variable Modeling in Educational Measurement

R.J. Zwitser

Summary of 'Contributions to Latent Variable Modeling in Educational Measurement'

One of the prominent questions in educational measurement is how to summarise scored item responses into a final score, such that the final score reflects the construct that is supposed to be measured. In answering this question, latent variable models play an important role. This thesis considers a couple of questions regarding the use of latent variable models in scoring item responses.

Chapter 1 is an introduction to the rest of the thesis. It is explained that there are different views on what a construct is. Furthermore, this chapter introduces some general terms and theory, after which a broad overview of the thesis is given.

The core of the thesis consists of Chapters 2 to 4, where three distinct research projects are described. The first project, described in Chapter 2, is about conditional likelihood inference from multistage testing designs. In adaptive testing, the scoring of individual test takers is usually done via estimates of person parameters. To obtain unbiased estimates, it is required that the item parameters are also unbiased. This chapter shows how to obtain item parameter estimates from multistage testing designs based on the conditional likelihood method. Besides this technical result, some more general issues related to adaptive testing, item parameter estimation, and model fit are discussed. It is explained that simple measurement models are more likely to fit the data obtained from adaptive designs compared to data obtained from linear designs. This is illustrated with simulated data, as well

as with real data taken from the Dutch *Entreetoets*.

In Chapter 3, the item response theory (IRT)-based justification of the use of the sum score is considered. Two IRT-models are well-known for their relationship between the sum score and the person parameter: the parametric Rasch Model (RM), in which the sum score is a sufficient statistic for the person parameter, and the nonparametric Monotone Homogeneity Model (MHM), in which the latent trait is stochastically ordered by the sum score. It is illustrated that there is a theoretical gap between the two: the RM enables scoring individuals by means of the sum score, while the MHM enables ordering groups by means of the sum score. To fill the gap, the concept of ordinal sufficiency is defined, and the nonparametric Rasch Model is introduced as a less restrictive nonparametric alternative that enables ordering individuals by means of the sum score.

The final project, in Chapter 4, is about differential item functioning (DIF) in international surveys. Usually, DIF is considered as a threat to validity, and as a phenomenon that hinders the comparison of performances between countries. However, in the approach described in Chapter 4, DIF is not considered as a threat, but as an interesting survey outcome reflecting qualitative differences between countries. To obtain comparable scores in a context with DIF, it is proposed not the take the person parameter estimates as a basis for comparison. Instead, it is proposed to define the construct as a market basket of items, and to take (a summary statistic of) the item responses as the basis for comparisons. Since survey data are usually incomplete, the latent variable models - probably different models in different countries - are used to describe the distribution of the item responses in the market basket. This approach is illustrated with data from the PISA cycle of 2006.

Chapter 5 is a general discussion. Three issues related to Chapters 2 to 4 are raised. The first is about an optimal adaptive test design for high-stakes testing. It is argued that this is not a computerized adaptive test (CAT) with an infinitely large and calibrated item bank. Instead, a multistage test can lead to more efficient results. The second is about what to do when the sum score is not ordinal sufficient for the person parameter. It is argued that, especially in high-stakes testing, one should look for a coarser statistic that is ordinal sufficient. The third issue is an elaboration on the positive aspects of DIF.