Contributions to latent variable modeling in educational measurement

Zwitser, R.J.

Citation for published version (APA):
Zwitser, R. J. (2015). *Contributions to latent variable modeling in educational measurement*.

Chapter 5

Discussion

In the previous three chapters, I have described the results of three research projects related to the use of latent variable models in educational testing. Each chapter ended with some topics for discussion. In this section, I will briefly discuss the topics of this thesis from a more personal perspective. I will do this chapter by chapter.

5.1 The optimal CAT for high-stakes testing

Following from the research about item parameter estimation in multistage testing designs, I would like to discuss an issue that was already introduced in the discussion of Chapter 2, and that is the relationship between the estimation error of both item and person parameters. The main question is: what is an optimal computerized adaptive test (CAT) for high-stakes testing?

In high-stakes testing, test results can have important consequences for the test taker. These consequences force us to be careful with assumptions that can have major impact on the inferences from the test data. Since differences in testing stakes may result in different item characteristics (Brennan, 2006), I think that it is preferable to use item parameter estimates that are based on the actual test data and not on pretest data that are obtained under other testing conditions. Furthermore, all sources of random error should be taken into account in the final θ estimate. As I already brought up in the discussion in Chapter 2, this latter requirement has interesting implications for the optimality of adaptive testing designs.

Let us consider first a traditional CAT that is based on a calibrated item bank, and where items are selected based on the *Fisher information* criterion (Fisher, 1922; Van der Linden & Pashley, 2010). A well-known property of such

CATs is that, compared to linear tests, person parameters can be estimated more efficiently. The increase of efficiency depends on a couple of factors. One is the availability of items which provide high information at the level of the current θ estimate. It is straightforward that the availability increases if the size of the item bank increases. And in the limiting case, an optimal CAT is based on an *infinitely large* calibrated item bank.

But are CATs based on large item banks really optimal? This way of optimizing the accuracy of the θ estimates rests on the assumption that the item parameters are known. In fact, the traditional (optimal) CAT aims to administer a subset S to an individual student such as to minimize the posterior variance of ability:

$$\text{VAR}(\Theta|\mathbf{x}_S, \boldsymbol{\lambda}), \tag{5.1}$$

where $\boldsymbol{\lambda}$ denotes a vector of fixed item parameters.

If we take seriously the requirement that item parameters should be estimated from the actual test administration, we run into trouble. We may still use the parameters $\boldsymbol{\lambda}$ from the item bank to identify the set of items S to administer to an individual student. After test administration, however, we need to recompute the item parameters from the actual test administration data.

The total uncertainty about θ now depends both on the set of items S and on the uncertainty regarding $\boldsymbol{\lambda}$. The correct posterior variance, taking into account uncertainty regarding item parameters, is the following:

$$\text{E}[\text{VAR}(\Theta|\mathbf{x}_S, \boldsymbol{\Lambda})|\mathbf{x}_S] + \text{Var}[\text{E}(\Theta|\mathbf{x}_S, \boldsymbol{\Lambda})|\mathbf{x}_S] \tag{5.2}$$

With a typical linear test (i.e., the same S for all students) administered to a large enough number of students, the second term in (5.2) is negligible, at the expense of the first term being potentially inflated. In contrast, in CAT the first term is minimized, at the expense of the second term being inflated.

This brings me to the following position: optimal inferences from high-stakes adaptive testing data implies a balance between the estimation error of person and item parameters. In that sense, MST could be more efficient compared CAT. I consider the minimization of (5.2) one of the main outstanding questions for CAT research.

5.2 To order, or not to order: that is the question

In Chapter 3, I have discussed the use of the sum score for ordering individual test takers. This was quite a formal discussion with the purpose to distinguish between the stochastic ordering of the latent trait (SOL) property (Hemker et al., 1997) that enables inferences about groups of students, and the sufficiency property that enables inferences about individual students. The representation of sufficiency in terms of the stochastic ordering of posterior distributions of θ provided the baseline for the nonparametric alternative for the Rasch model (npRM): if the purpose is to order individuals, then we have to investigate whether the ordering of sum scores corresponds with the stochastic ordering of the posterior distributions of the person parameters. The only difference between this npRM and the parametric Rasch model (RM, Rasch, 1960) is that it is not required that equal sum scores imply stochastically equal posterior distributions.

This representation of sufficiency, brings me to the following issue in high-stakes testing: based on which evidence can we classify students into ordered groups, and when do we have to decide that we cannot make a distinction between students?

In cases where the RM fits we feel comfortable to order based on the sum score, because all available statistical informations is kept by the sum score. If the RM does not fit, one can consider the less restrictive npRM, however, the difference between both models is only about cases with equal sum scores. In case of unequal sum scores, both models are equally restrictive. If the RM is seen as restrictive, then the same could be said about the npRM. But on which argument can we classify individual students if the npRM also does not fit? I will mention three options, and discuss them in the context of high-stakes testing.

The first is to leave the requirement of ordinal sufficiency of the sum score. Then we arrive at the monotone homogeneity model (Mokken, 1971). If we do this, and we use the sum score as final statistic, we decide to classify students into groups based on a summary statistic, while other available information does not support this classification. I think that this is questionable. If we classify in high-stakes conditions, then I think that this classification should not be contradicted by other available test data. The second option is to use a parametric model with more parameters, e.g., the Two-Parameter Logistic model (Birnbaum, 1968). However, a disadvantage of these models is that the relation between observed responses and the final score (i.e., estimated ability) is less than transparent, and strictly not in agreement with the sum score. The third option is the one that I would like to emphasize in this discussion: find another ordinal sufficient statistic, coarser than the sum score. One trivial option is already provided in Chapter 3: assign the score 0 if all responses are incorrect, assign the score 1 if some responses are correct and some responses are incorrect, and assign the score 2 if all responses are correct. This scoring rule is, of course, rather silly, yet it shows that we may be able to form sum score groups, which have the ordinal sufficiency property. And what if we cannot find such groups? In that case, we may ask ourselves whether we have enough evidence to make ordinal judgements.

5.3 We want DIF!

In Chapter 4, I have considered the methodology for educational surveys from the perspective of differential item functioning (DIF). DIF is usually seen as a threat to validity, and therefore as something that should be avoided. But for educational surveys, I have proposed another view on DIF: DIF could be an interesting survey outcome, reflecting the diversity among countries and the dynamics over time.

I think that the discussion about DIF needs more nuance. DIF is only a threat if it appears where it is not expected. For instance, if a test is designed to measure a unidimensional construct, and if the theory about the construct is such that the construct can be represented as a latent variable, and the relationship between item scores and latent variables is assumed to be the same in each subpopulation, then DIF contradicts the assumptions, and might cause bias in the latent variable estimation.

So, the question is: what do we expect with respect to educational surveys? In the most recent PISA survey, 65 economies were involved (OECD, 2014). The participating countries are spread over six continents, so differences with respect to cultural background, curriculum, and school system are evident. Because of the diversity among countries, I expect that item scores and background variables are dependent, also after conditioning on the total score. This implies that I expect DIF. Moreover, it provides new possibilities to use a survey instrument as a monitoring system. For instance, it could be that students in some country stay behind on a particular type of math exercises. If item analyses provide such information, then the country's content experts and policy makers could think of explanations and interventions. They could use this information to improve the curriculum, and if they succeed, then this interaction effect is expected to be different in the next cycle of the survey. These interventions might contribute to the quality of the educational system, but at the same time create DIF over time. However, successful and instrumental DIF that we should not want to avoid.