Measurement and Research Department Reports 2012-1

Conditional Statistical Inference with Multistage Testing Designs

Robert J. Zwitser
Gunter Maris

Measurement and Research Department Reports 2012-1

Conditional Statistical Inference with Multistage Testing Designs

Robert J. Zwitser

Gunter Maris

Cito
Arnhem, 2012

Abstract

In this paper it is demonstrated how statistical inference from multistage test designs, where students are administered different modules of items depending on their responses to earlier modules, can be made with the method of conditional maximum likelihood. It is shown how the match between item difficulty and student proficiency may result in a better fit of simple measurement models, owing to the avoidance of undesirable response behavior like slipping and guessing. Attention is given to the assessment of model fit. The results were illustrated with simulated data as well as with real data.

1 Introduction

For several decades, test developers have been working on the development of adaptive test designs in order to obtain more efficient and robust measurement procedures. The general idea is that the better match between item difficulty and the proficiency of the students leads to more efficient parameter estimates and lowers the risk of undesirable response behavior, like guessing and slipping. Consequently, adaptive designs can go along with simpler models, e.g., the necessity of guessing parameters may diminish, and these simpler models may still fit the data well.

Two well-known examples of adaptive designs are *computerized adaptive testing* (CAT) and *multistage testing* (MST). Although both include a variety of designs, the general difference between CAT and MST is that, in CAT, items are selected individually, while, in MST, items are selected in blocks/modules. An example of an MST design is given in Figure 1. In the first stage, all students take the first module[1]. 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 module 2, whereas students with a score higher than c on the routing test take module 3. Every unique sequence of modules is called a booklet.

CAT versus MST Research on MST has primarily focused on the structure of the design (e.g., the number of modules or the number of items per module), automated test assembly, and efficiency comparisons with ordinary paper-and-pencil testing and CAT (see Zenisky, Hambleton, & Luecht, 2010, for an overview). CAT designs are more efficient, while some benefits of MST designs are that it offers test developers the opportunity to have control over the content of the total test form and more control over item

[1]We use a superscript (m) to denote random variables and parameters that relate to the m-th module.

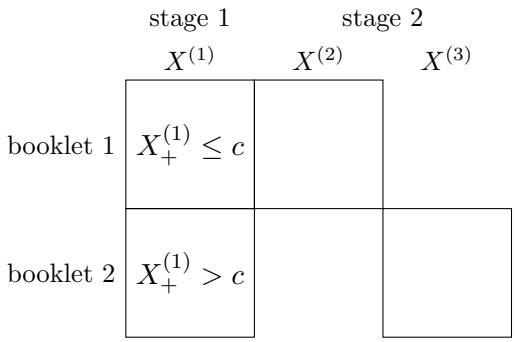


Figure 1: Example of a multistage design.

exposure. In addition, test takers may skip an item or change a response within the module (Luecht & Nungester, 1998; Luecht, Brumfield, & Breithaupt, 2006; Mead, 2006; Hendrickson, 2007; Zenisky et al., 2010).

However, these applications of MST designs, as well as CAT designs, are based a calibrated item bank. But, and this is another benefit of MST, an item bank is not required. Items could be assigned to modules by content specialists, or based on a minor pretest, and then calibrated after test administration. The only requirement is a suitable measurement model and a method to obtain estimates for the parameters. In other words, in MST the benefits of adaptive testing can be accomplished without an item bank.

In the past, only a few studies focused on the calibration of items in an MST design. Those were based on Bayesian inference (Wainer, Bradlow, & Du, 2000) or *marginal maximum likelihood* (MML) inference (Glas, 1988; Glas, Wainer, & Bradlow, 2000). In this paper, we consider statistical inference from the *conditional maximum likelihood* (CML) perspective (Andersen, 1973a). A benefit of this method is that, in contrast to Bayesian inference or MML, no assumptions are needed about the distribution of ability in the population, and it is not necessary to draw a random sample from the population. However, it has been suggested that the CML method cannot be applied with MST (Glas, 1988; Eggen & Verhelst, 2011). We will show

in Section 1 that this conclusion is not correct, and we consider parameter estimation by the CML method with MST feasible. In Section 2, we will propose how the model fit can be evaluated. Some illustrations are given to elucidate our results (Section 3). Throughout the paper, we use the MST design in Figure 1 for illustrative purposes. The extent to which our results for this simple MST design generalize to more complex designs is discussed in Section 4.

2 Conditional Likelihood Estimation

Throughout the paper, we will use the Rasch model (Rasch, 1960) in our derivations and examples. The model is defined as follows:

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}, \boldsymbol{b}) = \prod_{p=1}^{M} \prod_{i=1}^{N} \frac{\exp[(\theta_p - b_i)x_{pi}]}{1 + \exp(\theta_p - b_i)}, \tag{1}$$

in which M and N denote the number of persons and items, respectively. The Rasch model is an exponential family distribution with the sum score

$$X_{p+} = \sum_i X_{pi} \text{ sufficient for } \theta_p$$

and

$$X_{+i} = \sum_p X_{pi} \text{ sufficient for } b_i.$$

Statistical inference about $\mathbf{X}$ is hampered by the fact that the person parameters θ_p are incidental. That is, their number increases with the sample size. It is known that, in the presence of an increasing number of incidental parameters, it is, in general, not possible to estimate the (structural) item parameters consistently (Neyman & Scott, 1948). This problem can be overcome in one of two ways. The first is MML inference (Bock & Aitkin, 1981): if the students can be conceived of as a random sample from a well-defined population characterized by an ability distribution G, inferences can be based on the marginal distribution of the data. That is, we integrate the

incidental parameters out of the model. Rather than estimating each student's ability, only the parameters of the ability *distribution* need to be estimated. The second is CML inference: since the Rasch model is an exponential family model, we can base our inferences on the distribution of the data $\mathbf{X}$ conditionally on the sufficient statistics for the incidental parameters. Obviously, this conditional distribution no longer depends on the incidental parameters. Under suitable regularity conditions, both methods can be shown to lead to consistent estimates of the item difficulty parameters.

2.1 Estimation of Item Parameters

Suppose that every student responds to all three modules ($\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$). That is, we have complete data for every student. We now consider how the (distribution of the) complete data relate(s) to the (distribution of the) data from MST and derive the conditional likelihood upon which statistical inferences can be based.

The complete data likelihood can be factored as follows[2]:

$$\begin{aligned} P_{\boldsymbol{b}}(\mathbf{x}|\theta) =& P_{\boldsymbol{b}^{(1)}}(\mathbf{x}^{(1)}|x_+^{(1)}) P_{\boldsymbol{b}^{(2)}}(\mathbf{x}^{(2)}|x_+^{(2)}) P_{\boldsymbol{b}^{(3)}}(\mathbf{x}^{(3)}|x_+^{(3)}) \\ & P_{\boldsymbol{b}}(x_+^{(1)}, x_+^{(2)}, x_+^{(3)}|x_+) P_{\boldsymbol{b}}(x_+|\theta) \end{aligned}$$

where

$$\begin{aligned} P_{\boldsymbol{b}^{(m)}}(\mathbf{x}^{(m)}|x_+^{(m)}) &= \frac{\prod_i \exp(-x_i^{(m)} b_i^{(m)})}{\gamma_{x_+^{(m)}}(\boldsymbol{b}^{(m)})}, \quad m = 1, 2, 3, \\ P_{\boldsymbol{b}}(x_+^{(1)}, x_+^{(2)}, x_+^{(3)}|x_+) &= \frac{\gamma_{x_+^{(1)}}(\boldsymbol{b}^{(1)}) \gamma_{x_+^{(2)}}(\boldsymbol{b}^{(2)}) \gamma_{x_+^{(3)}}(\boldsymbol{b}^{(3)})}{\gamma_{x_+}(\boldsymbol{b})}, \\ P_{\boldsymbol{b}}(x_+|\theta) &= \frac{\gamma_{x_+}(\boldsymbol{b}) \exp(x_+\theta)}{\sum_s \gamma_s(\boldsymbol{b}) \exp(s\theta)}, \end{aligned}$$

[2]Whenever possible without introducing ambiguity, we ignore the distinction between random variables and their realizations in our formulae.

and $\gamma_s(\boldsymbol{b}^{(m)})$ is the *elementary symmetric function* of order s:

$$\gamma_s(\boldsymbol{b}^{(m)}) = \sum_{\mathbf{x}:x_+=s} \prod_i \exp(-x_i b_i^{(m)}),$$

which equals zero if s is smaller than zero or larger than the number of elements in $\boldsymbol{b}^{(m)}$.

The various elementary symmetric functions are related to each other in the following way:

$$\gamma_{x_+}(\boldsymbol{b}) = \sum_{i+j+k=x_+} \gamma_i(\boldsymbol{b}^{(1)})\gamma_j(\boldsymbol{b}^{(2)})\gamma_k(\boldsymbol{b}^{(3)}).$$

To turn a sample from $\mathbf{X}$ into a realization of data from MST, we do the following: *If* the score of a student on module 1 is lower than or equal to c, we *delete* the responses on module 3, *otherwise*, we *delete* the responses on module 2. We now consider this procedure from a formal point of view. Formally, considering a student with score module 1 lower than or equal to c and deleting the responses on module 3 means that we consider the distribution of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ conditionally on θ *and* the event $X_+^{(1)} \leq c$:

$$P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\theta, X_+^{(1)} \leq c) = \frac{P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\theta)}{P_{\boldsymbol{b}^{(12)}}(X_+^{(1)} \leq c|\theta)}, \quad \text{if } x_+^{(1)} \leq c. \qquad (2)$$

That is, the *if* refers to conditioning and *deleting* to integrating out. In the following, it is to be implicitly understood that conditional distributions are equal to zero if the conditioning event does not occur in the realization of the random variable.
We now show that the conditional distribution in (2) factors as follows:

$$\begin{aligned} &P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\theta, X_+^{(1)} \leq c) \\ =&P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|x_+^{(12)}, X_+^{(1)} \leq c) P_{\boldsymbol{b}^{(12)}}(x_+^{(12)}|\theta, X_+^{(1)} \leq c). \end{aligned}$$

That is, the score $X_+^{(12)}$ is sufficient for θ and hence the conditional likelihood $P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|x_+^{(12)}, X_+^{(1)} \leq c)$ can be used for making inferences about

$\boldsymbol{b}^{(12)}$.

First, we consider the distribution of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ conditionally on $X_+^{(12)}$, which is known to be independent of θ:

$$P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(12)}|x_+^{(12)}) = \frac{\prod_i \exp(-x_i b_i^{(1)}) \prod_j \exp(-x_j b_j^{(2)})}{\gamma_{x_+^{(12)}}(\boldsymbol{b}^{(12)})}$$

where

$$\gamma_{x_+^{(12)}}(\boldsymbol{b}^{(12)}) = \sum_{j=0}^{n} \gamma_j(\boldsymbol{b}^{(1)}) \gamma_{x_+^{(12)}-j}(\boldsymbol{b}^{(2)}).$$

Second, we consider the probability that $X_+^{(1)}$ is lower than or equal to c conditionally on $X_+^{(12)}$:

$$P_{\boldsymbol{b}^{(12)}}(X_+^{(1)} \leq c|x_+^{(12)}) = \frac{\sum_{j=0}^{c} \gamma_j(\boldsymbol{b}^{(1)}) \gamma_{x_+^{(12)}-j}(\boldsymbol{b}^{(2)})}{\sum_{j=0}^{n} \gamma_j(\boldsymbol{b}^{(1)}) \gamma_{x_+^{(12)}-j}(\boldsymbol{b}^{(2)})}.$$

Hence, we obtain

$$P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|X_+^{(1)} \leq c, x_+^{(12)}) = \frac{\prod_i \exp(-x_i b_i^{(1)}) \prod_j \exp(-x_j b_j^{(2)})}{\sum_{j=0}^{c} \gamma_j(\boldsymbol{b}^{(1)}) \gamma_{x_+^{(12)}-j}(\boldsymbol{b}^{(2)})}. \quad (3)$$

We next consider the distribution of $X_+^{(12)}$ conditionally on θ and $X_+^{(1)} \leq c$. Since the joint distribution of $X_+^{(1)}$ and $X_+^{(2)}$ conditionally on θ has the following form:

$$P_{\boldsymbol{b}^{(12)}}(x_+^{(1)}, x_+^{(2)}|\theta) = \frac{\gamma_{x_+^{(1)}}(\boldsymbol{b}^{(1)}) \gamma_{x_+^{(2)}}(\boldsymbol{b}^{(2)}) \exp([x_+^{(1)} + x_+^{(2)}]\theta)}{\sum_{0 \leq j+k \leq n} \gamma_j(\boldsymbol{b}^{(1)}) \gamma_k(\boldsymbol{b}^{(2)}) \exp([j+k]\theta)},$$

we obtain

$$\begin{aligned} P_{\boldsymbol{b}^{(12)}}(x_+^{(12)}|\theta, X_+^{(1)} \leq c) &= \frac{P_{\boldsymbol{b}^{(12)}}(x_+^{(12)}, X_+^{(1)} \leq c|\theta)}{P_{\boldsymbol{b}^{(12)}}(X_+^{(1)} \leq c|\theta)} \\ &= \frac{\sum_{j \leq c} \gamma_j(\boldsymbol{b}^{(1)}) \gamma_{x_+^{(12)}-j}(\boldsymbol{b}^{(2)}) \exp(x_+^{(12)}\theta)}{\sum_{\substack{0 \leq j+k \leq n \\ j \leq c}} \gamma_j(\boldsymbol{b}^{(1)}) \gamma_k(\boldsymbol{b}^{(2)}) \exp([j+k]\theta)}. \end{aligned}$$

Finally, we can write the likelihood for a single student in MST who receives a score lower than or equal to c on module 1:

$$\begin{aligned}
&P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\theta, X_+^{(1)} \leq c) \\
=&P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|x_+^{(12)}, X_+^{(1)} \leq c) P_{\boldsymbol{b}^{(12)}}(x_+^{(12)}|\theta, X_+^{(1)} \leq c) \\
=&\frac{\prod_i \exp(-x_i^{(1)} b_i^{(1)}) \prod_j \exp(-x_j^{(2)} b_j^{(2)}) \exp(x_+^{(12)}\theta)}{\sum_{\substack{0 \leq j+k \leq n \\ j \leq c}} \gamma_j(\boldsymbol{b}^{(1)}) \gamma_k(\boldsymbol{b}^{(2)}) \exp([j+k]\theta)}
\end{aligned} \tag{4}$$

Obviously, a similar result holds for a student who receives a score higher than c on module 1 and hence takes module 3.

With the results from this section, we can safely use CML inference, using (3) as the conditional likelihood.

2.2 Comparison with Alternative Estimation Procedures

The first way to deal with an MST design is to ignore the fact that the assignment of items depends on the student's previous responses. This means that when a student receives a score lower than or equal to c on module 1, we use the likelihood of the observations conditionally on θ only

$$P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\theta) = \frac{\prod_i \exp(-x_i^{(1)} b_i^{(1)}) \prod_j \exp(-x_j^{(2)} b_j^{(2)}) \exp(x_+^{(12)}\theta)}{\sum_{0 \leq j+k \leq n} \gamma_j(\boldsymbol{b}^{(1)}) \gamma_k(\boldsymbol{b}^{(2)}) \exp([j+k]\theta)} \tag{5}$$

instead of the correct likelihood in (4) as the basis for statistical inferences. It has been observed that if we use the conditional likelihood corresponding to the distribution in (5) as the basis for estimating the item parameters, we get bias in the estimators (Eggen & Verhelst, 2011). In Section 4.1.2, we illustrate this phenomenon. If we compare the likelihood in (4) with that in (5), we see that the only difference is in the range of the sum in the denominators. This reflects that in (4) we take into account that values of $X_+^{(1)}$ larger than c cannot occur, whereas in (5) this is not taken into account.

The second way to deal with an MST design is to separately estimate the parameters in each step of the design (Glas, 1989). This means that inferences

with respect to $\mathbf{X}^{(m)}$ are based on the likelihood of $\mathbf{X}^{(m)}$ conditionally on $X_+^{(m)} = x_+^{(m)}$. This procedure leads to unbiased estimates. However, since the parameters are not identifiable, we need to impose a separate restriction for each stage in the design (e.g., $b_1^{(1)} = 0$ and $b_1^{(2)} = 0$). As a consequence, it is not possible to place the items from different stages in the design on the same scale. More important, it is not possible to use all available information to obtain a unique estimate of the ability of the candidate.

Third, we consider the use of MML inference. In the previous section, we derived the likelihood function of the data conditionally on the design. For MML inference, we could use the corresponding marginal (w.r.t. θ) likelihood conditionally on the design ($X_+^{(1)} \leq c$):

$$\begin{aligned} & P_{\boldsymbol{b}^{(12)},\boldsymbol{\lambda}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)} | X_+^{(1)} \leq c) \\ = & \int_{\mathcal{R}_\theta} P_{\boldsymbol{b}^{(12)}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)} | \theta, X_+^{(1)} \leq c) f_{\boldsymbol{b}^{(1)},\boldsymbol{\lambda}}(\theta | X_+^{(1)} \leq c) d\theta, \end{aligned}$$

in which $\boldsymbol{\lambda}$ are the parameters of the distribution of θ.

If we use this likelihood, we disregard any information about the parameters that is contained in the (marginal distribution of the) design variable: $P_{\boldsymbol{b}^{(1)},\boldsymbol{\lambda}}(X_+^{(1)} \leq c)$.

We now consider how we can base our inferences on *all* available information: the responses on the routing test $\mathbf{X}^{(1)}$; the responses on the other modules that were *administered*, which we denote by $\mathbf{X}^{obs}$; and the design variable $X_+^{(1)} \leq c$. The complete likelihood of the observations can be written as follows

$$\begin{aligned} P_{\boldsymbol{b}^{123}}(\mathbf{X}^{(1)} = \mathbf{x}^{(1)}, \mathbf{X}^{obs} = \mathbf{x}^{obs} | \theta) = & P_{\boldsymbol{b}^{(2)}}(\mathbf{X}^{(2)} = \mathbf{x}^{obs} | \theta) P_{\boldsymbol{b}^{(1)}}(\mathbf{X}^{(1)} = \mathbf{x}^{(1)} | \theta) \\ & P_{\boldsymbol{b}^{(1)}}(X_+^{(1)} \leq c | \mathbf{X}^{(1)} = \mathbf{x}^{(1)}) + \\ & P_{\boldsymbol{b}^{(3)}}(\mathbf{X}^{(3)} = \mathbf{x}^{obs} | \theta) P_{\boldsymbol{b}^{(1)}}(\mathbf{X}^{(1)} = \mathbf{x}^{(1)} | \theta) \\ & P_{\boldsymbol{b}^{(1)}}(X_+^{(1)} > c | \mathbf{X}^{(1)} = \mathbf{x}^{(1)}). \end{aligned} \tag{6}$$

From this we immediately obtain the marginal likelihood function:

$$
\begin{aligned}
& P_{\boldsymbol{b}^{123}}(\mathbf{X}^{(1)} = \mathbf{x}^{(1)}, \mathbf{X}^{obs} = \mathbf{x}^{obs}) \\
&= \int_{\mathcal{R}_\theta} P_{\boldsymbol{b}^{123}}(\mathbf{X}^{(1)} = \mathbf{x}^{(1)}, \mathbf{X}^{obs} = \mathbf{x}^{obs}|\theta) f_{\boldsymbol{\lambda}}(\theta) d\theta \qquad (7) \\
&= \left[\int_{\mathcal{R}_\theta} P_{\boldsymbol{b}^{(2)}}(\mathbf{X}^{(2)} = \mathbf{x}^{obs}|\theta) P_{\boldsymbol{b}^{(1)}}(\mathbf{X}^{(1)} = \mathbf{x}^{(1)}|\theta) f_{\boldsymbol{\lambda}}(\theta) d\theta\right] P(X_+^{(1)} \leq c|\mathbf{x}^{(1)}) + \\
&\quad \left[\int_{\mathcal{R}_\theta} P_{\boldsymbol{b}^{(3)}}(\mathbf{X}^{(3)} = \mathbf{x}^{obs}|\theta) P_{\boldsymbol{b}^{(1)}}(\mathbf{X}^{(1)} = \mathbf{x}^{(1)}|\theta) f_{\boldsymbol{\lambda}}(\theta) d\theta\right] P(X_+^{(1)} > c|\mathbf{x}^{(1)}).
\end{aligned}
$$

Since either $P(X_+^{(1)} \leq c|\mathbf{x}^{(1)}) = 1$ and $P(X_+^{(1)} > c|\mathbf{x}^{(1)}) = 0$, or $P(X_+^{(1)} \leq c|\mathbf{x}^{(1)}) = 0$ and $P(X_+^{(1)} > c|\mathbf{x}^{(1)}) = 1$, the marginal likelihood function we obtain is equal to the marginal likelihood function we would have obtained had we planned *beforehand* to which candidates we would administer which modules. This means that we may safely *ignore* the design and use a computer program that allows for incomplete data (e.g., the OPLM program, Verhelst, Glas, & Verstralen, 1993) to estimate the item and population parameters. This is an instance of a situation where the *ignorability* principle applies (Rubin, 1976).

As already mentioned, a drawback of the marginal likelihood approach is that a random sample from a well-defined population is needed and that additional assumptions about the distribution of ability in this population need to be added to the model. In Section 4.1.2, we show that misspecification of the population distribution can cause serious bias in the estimated item parameters.

2.3 Estimation of Person Parameters

In principle, it is straightforward to estimate the ability parameter θ of a student who was administered the second module by the maximum likelihood method from the distribution of the sufficient statistic $X_+^{(12)}$ conditionally

on θ and the design:

$$P_{\boldsymbol{b}^{(12)}}(x_+^{(12)}|\theta, X_+^{(1)} \leq c) = \frac{\sum_{j\leq c} \gamma_j(\boldsymbol{b}^{(1)})\gamma_{x_+^{(12)}-j}(\boldsymbol{b}^{(2)}) \exp(x_+^{(12)}\theta)}{\sum_{\substack{0\leq j+k\leq n \\ j\leq c}} \gamma_j(\boldsymbol{b}^{(1)})\gamma_k(\boldsymbol{b}^{(2)}) \exp([j+k]\theta)}.$$

As usual, we consider the item parameters as known when we estimate ability. However, as is the case for a single-stage design, the ability is estimated at plus (minus) infinity for a student with a perfect (zero) score and can be shown to be biased. For that reason, we propose a weighted maximum likelihood (WML) estimator as Warm (1989) did for single-stage designs. The weighted likelihood is the likelihood multiplied by the square root of the information.

3 Model Fit

We have mentioned in the introduction that adaptive designs may be beneficial for model fit. In order to investigate this presumption, we propose two goodness of fit tests for MST designs. These tests are based on the method that was proposed by Andersen (1973b).

3.1 Likelihood Ratio Test

Andersen (1973b) showed that the item parameters $\boldsymbol{b}$ can be estimated by maximizing the conditional likelihood $\mathcal{L}(\boldsymbol{b})$ as well as by maximizing $\mathcal{L}^{(t)}(\boldsymbol{b})$, which is

$$P_{\boldsymbol{b}}(\mathbf{x}|X_+ = t).$$

For a complete design with N items, he considered

$$Z = 2\sum_{t=1}^{N-1} \log[\mathcal{L}^{(t)}(\hat{\boldsymbol{b}}^{(t)})] - 2\log[\mathcal{L}(\hat{\boldsymbol{b}})] \tag{8}$$

as the test statistic. Let us denote M_t as the number of persons with sum score t. It is shown that if $M_t \to \infty$ for $t = 1, \cdots, N-1$, then Z tends to

a limiting χ^2-distribution with $(N-1)(N-2)$ degrees of freedom, i.e., the difference between the number of parameters in the alternative model and the null model.

This likelihood ratio test (LRT) can also be applied with incomplete designs. Then (8) generalizes to

$$Z = 2\sum_{b}\sum_{t=1}^{N_b-1} \log[\mathcal{L}^{(bt)}(\hat{\boldsymbol{b}}^{(bt)})] - 2\log[\mathcal{L}(\hat{\boldsymbol{b}})], \tag{9}$$

where N_b denotes the number of items in booklet b. This statistic can also be applied with an MST design. In that case, the sum over t has to be adjusted for the scores that can be obtained. We will illustrate this for the design in Figure 1.

Let $N^{(m)}$ be the number of items in module m. Then the number of parameters estimated in the null model is

$$\sum_{m} N^{(m)} - 1.$$

One parameter cannot be estimated owing to scale identification. In a general booklet structure without dependencies between modules, we estimate $N^{(1)}+N^{(2)}-1$ parameters in each score group in booklet 1 and $N^{(1)}+N^{(3)}-1$ parameters in booklet 2 (see Figure 2). In booklet 1, there are $N^{(1)}+N^{(2)}+1$ score groups; in booklet 2, there are $N^{(1)}+N^{(3)}+1$ score groups. However, the minimum and the maximum score groups (dark grey in Figure 2) do not provide statistical information and therefore the number of parameters estimated in the alternative model is $(N^{(1)}+N^{(2)}-1)(N^{(1)}+N^{(2)}-1)+(N^{(1)}+N^{(3)}-1)(N^{(1)}+N^{(3)}-1)$. Finally, the number of degrees of freedom is

$$\begin{aligned}&(N^{(1)}+N^{(2)}-1)(N^{(1)}+N^{(2)}-1)+\\&(N^{(1)}+N^{(3)}-1)(N^{(1)}+N^{(3)}-1)-\\&(N^{(1)}+N^{(2)}+N^{(3)}-1).\end{aligned}$$

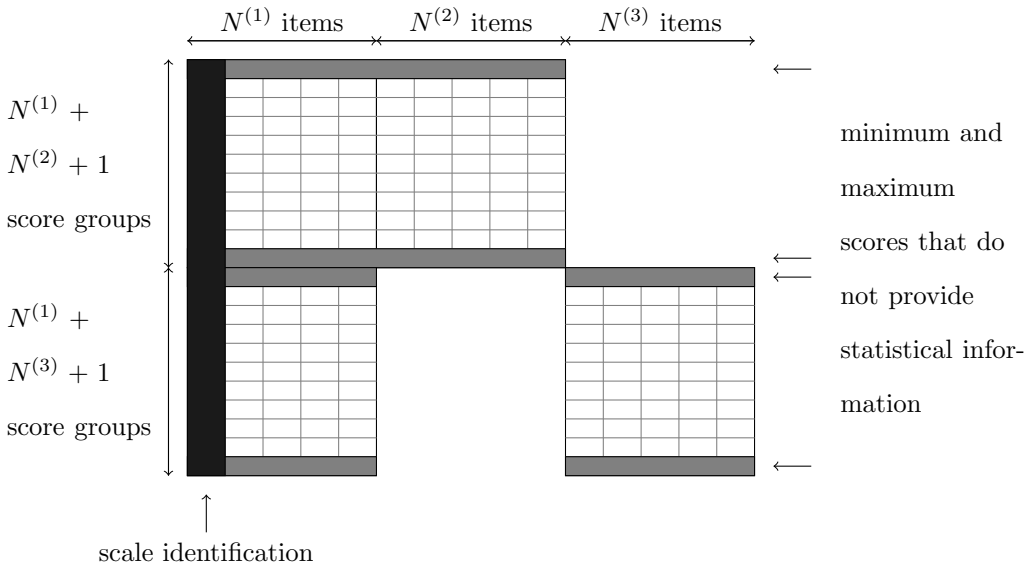


Figure 2: Degrees of freedom in a general booklet design.

The number of parameters of the alternative model in an MST design is slightly different, owing to the fact that some scores cannot be obtained. This can be illustrated by Figure 3. In booklet 1, there are $c + N^{(2)} + 1$ score groups. The score group $t = 0$ does not contain statistical information about $\boldsymbol{b}^{(12)}$, as well as the score group $t = c + N^{(2)}$ about $\boldsymbol{b}^{(2)}$. In the latter case, all items in $X^{(2)}$ must have been answered correctly. The same kind of reasoning holds for booklet 2. The number of parameters estimated in the alternative model is $(c + N^{(2)})(N^{(1)} - 1) + (c + N^{(2)} - 1)(N^{(2)}) + (N^{(1)} + N^{(2)} - c - 1)(N^{(1)} - 1) + (N^{(1)} + N^{(3)} - c - 2)N^{(3)}$. Therefore, the number of degrees of freedom is

$$
\begin{aligned}
&(c + N^{(2)})(N^{(1)} - 1) + (c + N^{(2)} - 1)(N^{(2)}) + \\
&(N^{(1)} + N^{(2)} - c - 1)(N^{(1)} - 1) + (N^{(1)} + N^{(3)} - c - 2)N^{(3)} - \\
&(N^{(1)} + N^{(2)} + N^{(3)} - 1).
\end{aligned}
$$

Score Groups In (8) and (9) the estimation of $\boldsymbol{b}^{(t)}$ is based on the data with sum score t. Here t is a single value. In cases with many items, the

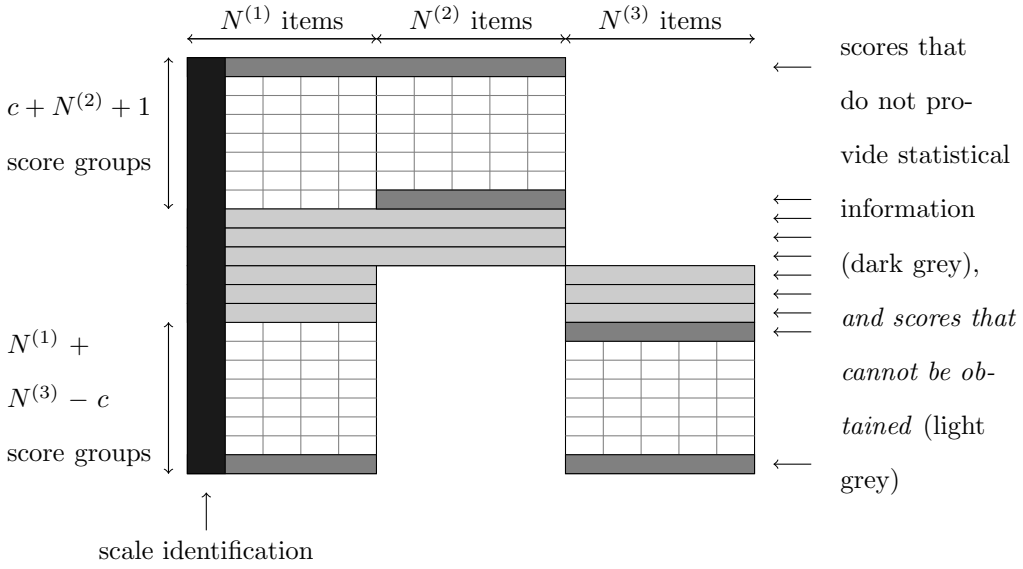


Figure 3: Degrees of freedom in an MST design.

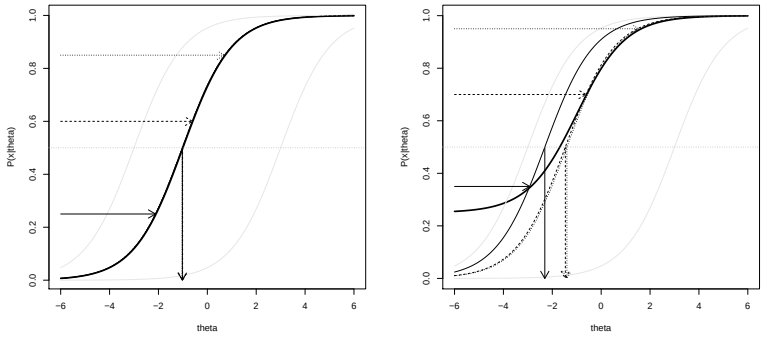
number of parameters under the alternative model becomes huge. Consequently, in some score groups, there may be little statistical information available about some parameters, e.g., information about easy items in the highest score groups. The LRT may then become conservative, since the convergence to the χ^2-distribution is not reached with many parameters and too few observations. To increase the power, the procedure can also be based on W sets of sum scores instead of single values t. Then

$$Z = 2\sum_{v=1}^{W} \log[\mathcal{L}^{(S_v)}(\hat{\boldsymbol{b}}^{(S_v)})] - 2\log[\mathcal{L}(\hat{\boldsymbol{b}})],$$

in which T is the set of possible sum scores t, v denotes the v-th score group, and $S_v \subset T$ such that $\{S_1, S_2, \cdots, S_v, \cdots, S_W\} = T$.

3.2 Item Fit Test

In the LRT defined above, the null hypothesis is tested against the alternative hypothesis that the model does not fit. The result does not provide any information about the type of model violation on the item level. Instead of



(a) A fitting item (b) A non-fitting item

Figure 4: Parameter estimates under the Rasch model in three score groups.

a general LRT, item fit tests can also be used to gain insight into the type of misfit.

What is known about the maximum likelihood estimates is that

$$\hat{\boldsymbol{b}}^{(S_v)} \sim \mathcal{N}(\boldsymbol{b}^{(S_v)}, \boldsymbol{\Sigma}^{(S_v)}),$$

and, under the null hypothesis that the Rasch model holds,

$$\forall v \; \boldsymbol{b}^{(S_v)} = \boldsymbol{b}. \tag{10}$$

Since the Rasch model is a member of the exponential family, the variance-covariance matrix, $\boldsymbol{\Sigma}$, can be estimated by minus the inverse of the second derivative of the log-likelihood function.

If the Rasch model does not fit, the estimates $\boldsymbol{b}^{(S_v)}$ can provide useful information about the type of violation, for instance, if the *item characteristic curve* (ICC) has a lower asymptote. In this case, the difference between the parameters of the score groups will have a certain pattern. This is illustrated by Figure 4. Figure 4a symbolizes a case where the Rasch model fits. Here all ICCs are parallel. The estimate of the item parameter (i.e. the scale value that corresponds to a probability of 0.5 of giving a correct response to that item) in the lower scoring group (solid arrow) is expected to be the

same as in the middle (dashed arrow) and the higher score group (dotted arrow). However, if an item has an ICC with a lower asymptote (see Figure 4b), then the estimates of the lower and the middle score groups will be different, while the estimates of the middle and the high score groups are expected to be almost the same.

4 Examples

In this section we demonstrate some properties of CML inference in MST. After a short description of the simulation design (Section 4.1.1), we compare the inference from the correct conditional likelihood with the incorrect inference from ordinary CML and from MML, in which the population distribution is misspecified (Section 4.1.2). In Section 4.1.3 and 4.1.4 we will demonstrate, respectively, the robustness and efficiency of the MST design. Finally, in Section 4.2 we will also demonstrate how data that were obtained from a ordinary linear test can be transformed into data from a MST in order to increase the model fit.

4.1 Simulation

4.1.1 Test and Population Characteristics

The first three examples are based on simulated data. We considered a test of 50 items that was divided into three modules. The first module (i.e., the routing test) consisted of 10 items with difficulty parameters drawn from a uniform distribution over the interval from -1 to 1. The second and third module both consisted of 20 items with difficulty parameters drawn from a uniform distribution over the interval from -2 to -1 and the interval from 0 to 2, respectively. The person parameters were drawn from a mixture of two normal distributions: with probability 2/3, they were drawn from a normal distribution with expectation -1.5 and standard deviation equal to

Table 1: Mean difference true and estimated item parameters.

	MST CML	Ordinary CML	MML
module 1	0.0057	0.0299	-0.0886
module 2	-0.0121	0.1601	-0.2100
module 3	0.0092	-0.2003	0.2543

0.5; with probability 1/3 they were drawn from a normal distribution with expectation 1 and standard deviation equal to 1.
When the test was administered in an MST design, the cut-off score, c, for the routing test was 5.

4.1.2 Bias Reduction

In the first example, 10,000 students were sampled and the test was administered in an MST design. In order to show the bias of the estimated item parameters according to ordinary CML and MML, these estimates, as well as the estimates based on the conditional likelihood in (3), were compared with the true parameter values. The numbers in Table 1 show, for each module, the mean difference between the true parameter values and their estimates according to each of the three estimation methods. Both ordinary CML and MML inference lead to serious bias in the estimated parameters. The MML analysis was based on the (incorrect) assumption of a normal distribution of the ability parameters with both the expectation and standard deviation as parameters. For ordinary CML inference, this bias is due to the fact that the MST design is not taken into account.
To get an impression of the behavior of CML inference based on the conditional likelihood in (3), we repeated the analysis with 100,000 students. This gave -0.0002, -0.0023, and 0.0024, respectively, as entries in Table 1. These results support the conclusion that CML inference based on the conditional likelihood in (3) is a viable and robust method of parameter estimation with

MST designs.

4.1.3 Goodness of Fit

In a second simulation study, we demonstrated the model fit procedure that is described in Section 3. The simulation consisted of 1,000 trials. In each trial, three different cases were simulated.

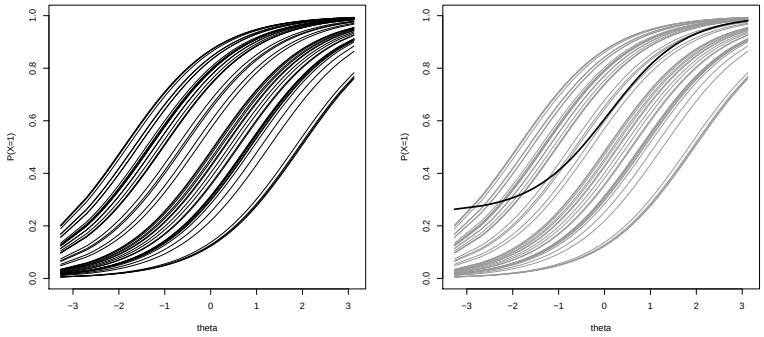
- Case 1: the MST design described above.
- Case 2: a complete design with all 50 items, except for the easiest item in module 3. The excluded item was replaced by an item according to the *3-parameter logistic model* (3PLM, Birnbaum, 1968) which is defined as follows:

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \prod_{p=1}^{M} \prod_{i=1}^{N} \left(c_i + (1 - c_i) \frac{\exp[a_i(\theta_p - b_i)x_{pi}]}{1 + \exp[a_i(\theta_p - b_i)]} \right). \tag{11}$$

  This 3PLM item has the same item difficulty as the excluded item. However, instead of $a = 1$ and $c = 0$, we have now have for this item $a = 1.2$ and $c = 0.25$. The slope (i.e., the a-parameter) was slightly changed, so that the ICC is more parallel to the other ICCs.

- Case 3: an MST with the items of case 2.

The ICCs of case 1 to 3 are displayed in Figure 5. Data were generated for a sample of 10,000 students and the item parameters of the Rasch model were estimated for each case. For the three cases above, an LRT as well as item fit tests were performed in each trial based on five score groups in each booklet. The score groups were constructed such that within each booklet the persons were equally distributed over the different score groups. The



(a) Case 1 (b) Case 2 and 3

Figure 5: (a) The ICCs of the 50 Rasch items for case 1. (b) The ICCs of the 49 Rasch items (gray), and the ICC of the 3PLM item (bold black) for case 2 and 3.

number of degrees of freedom in case 1 and 3 is

$$
\begin{aligned}
&2 \text{ (number of booklet)} \times \\
&5 \text{ (number of score groups per booklet)} \times \\
&29 \text{ (number of estimated parameters per score group)} - \\
&49 \text{ (number of estimated parameters in the null model)} \\
&= 241,
\end{aligned}
$$

and in case 2

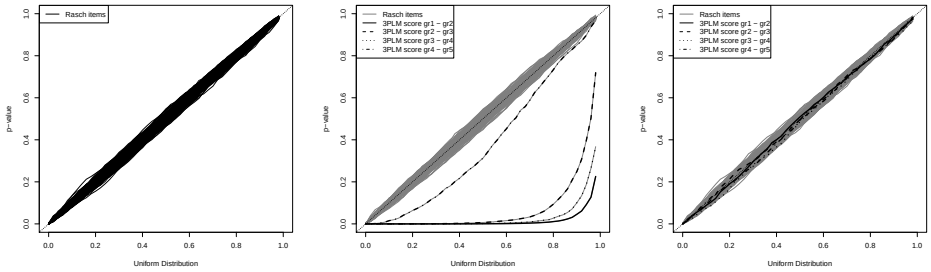
$$
\begin{aligned}
&5 \text{ (number of score groups per booklet)} \times \\
&49 \text{ (number of estimated parameters per score group)} - \\
&49 \text{ (number of estimated parameters in the null model)} \\
&= 196.
\end{aligned}
$$

Table 2: Results Kolmogorov-Smirnov test for testing the p-values of the LRTs against a uniform distribution.

Case	D^-	p-value
Case 1	0.016	0.774
Case 2	0.968	<0.001
Case 3	0.048	0.100

Likelihood Ratio Test If the model fits, then the p-values of the LRTs and the item fit tests are expected to be uniformly distributed over replications of the simulation. This hypothesis was checked for each case with a Kolmogorov-Smirnov test. The results are shown in Table 2. It can be seen that the Rasch model fits in case 1 and 3, but not in case 2. The fit for case 1 and the lack of fit for case 2 was expected. However, notice that the Rasch model also fits for case 3. In that case, one of the items is a 3PLM item, but this item is relatively easy and was only administered to students with a high score on the routing test, i.e., students with a high proficiency level. This indicates that guessing was avoided owing to this match of student proficiency and item difficulty.

Item Fit Test The distribution of the p-values of the item fit statistics is displayed graphically by QQ-plots in Figure 6. The item fit tests clearly mark the misfitting item in case 2. Notice that, as explained in Section 3.2, the item fit test in case 2 shows an effect between the lower score groups (i.e, between group 1 and 2, between group 2 and 3, and between group 3 and 4), while the p-values of the item fit tests between score group 4 and 5 are nearly uniformly distributed. The graph for case 3 in Figure 6 shows, as we already noted in the previous paragraph, that the 3PLM item does not cause misfit in the MST administration.



(a) Case 1 (b) Case 2 (c) Case 3

Figure 6: QQ-plots of the p-values of the item fit tests against the quantiles of a uniform distribution.

4.1.4 Efficiency

The relative efficiency of an MST design is demonstrated graphically by the information functions in Figure 7. Here the information of three different cases is given: all 50 items administered in a complete design, the average information over 100 random samples of 30 of the 50 items administered in a complete design, and the MST design described before. In the MST design, the total test information is

$$I(\theta) = I^{(12)}(\theta)P(X_+^{(1)} \leq c|\theta) + I^{(13)}(\theta)P(X_+^{(1)} > c|\theta).$$

Here $I^{(12)}(\theta)$ denotes the Fisher information function for module 1 and 2. The distribution of θ is also shown in Figure 7. It can be seen that, for most of the students in this population, the MST with 30 items is much more efficient than the linear test with 30 randomly selected items. In addition, for many students, the information based on the MST is not much less than the information based on all 50 items.

4.2 Real Data

The data for the examples based on real data was taken from the Dutch *Entrance Test* (in Dutch: *Entreetoets*), which consists of multiple parts that

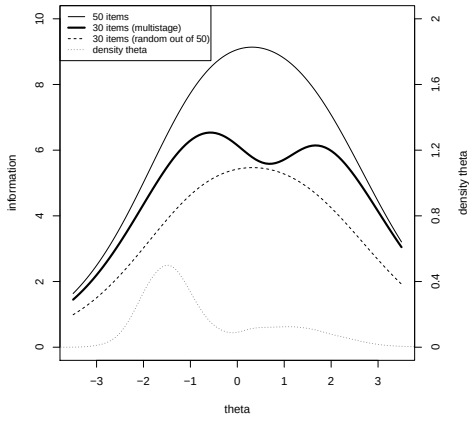
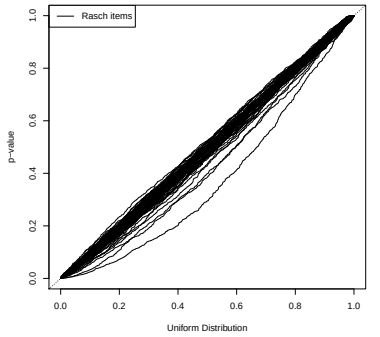


Figure 7: Person information in a complete design with 50 items, an MST design with 30 items, and a complete design with 30 items.

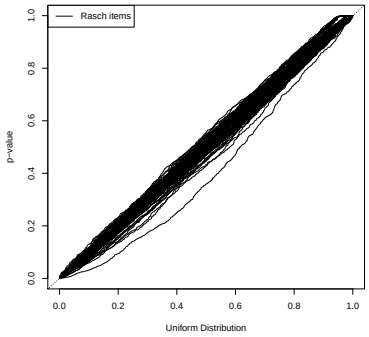
are administered annually to 125,000 grade 5 pupils. One of the parts is a test with 120 math items. To gain insight into the item characteristics, we first analyzed a sample of 30,000 students with the *One-Parameter Logistic Model* (OPLM, Verhelst & Glas, 1995; Verhelst et al., 1993). For illustrative purposes, we have selected 30 items that seem to have parallel ICCs, although the Rasch model did not fit perfectly, $R_{1c} = 420.66$, df $= 87$, $p < 0.001$. In addition to these 30 items, also one 3PLM item was selected. We can consider this example as an MST by allocating the items to three modules, after which the data of the students with a low (high) score on the routing test are removed from the third (second) module.

In order to demonstrate the item fit tests, we drew 1,000 samples of 1,000 students from the data. First, we estimated the parameters of the 30 Rasch items with a complete design and an MST design. In both cases, all items seem to fit the Rasch model reasonably well (see Figure 8a and Figure 8b).

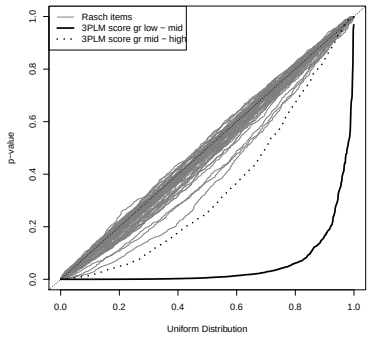
Then we added the 3PLM item to the Rasch items and again analyzed the complete design and the MST design. It can be seen from Figure 8c that



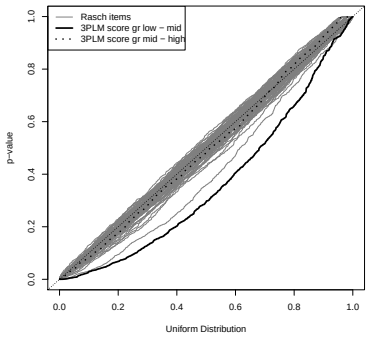
(a) Rasch - Complete design



(b) Rasch - MST design



(c) Rasch & 3PLM - Complete design



(d) Rasch & 3PLM - MST design

Figure 8: QQ-plots of the p-values of the item fit tests from the *Entrance Test* example against the quantiles of a uniform distribution.

the 3PLM item shows typical misfit in the complete design. The item fit test was based on three score groups. There is a substantial difference between the parameter estimates of the lower and the middle score group, while there seems to be a little difference between the estimates of the middle and the higher score groups. If the 3PLM item is administered in the third module of an MST design, the fit improves substantially (see Figure 8d).

5 Discussion

In this paper we have shown that the CML method is applicable with data from an MST. We have demonstrated how unbiased item parameters can be obtained for the Rasch model and how model fit can be investigated for the total test, as well as for individual items.

It is known that CML estimators are less efficient than MML estimators. When the requirements of the MML method are fulfilled, then the MML method may be preferable above the CML method. However, in practice, for instance in education, the distribution of person parameters may be skewed or multi-modal owing to all kinds of selection procedures. It was shown in an example that, when the population distribution is misspecified, the item parameters become seriously biased. For that reason, in cases where not so much is known about the population distribution, the use of the CML method may be preferable.

Our presumption was that adaptive designs are more robust against undesirable behavior like guessing and slipping. We have shown in the examples that even when some items may lead to guessing when administered to the wrong students, then simple models without guessing parameters may still fit the data well. The example with real data did also show that a distinction could be made between multistage *administration* and multistage *analysis*. Data obtained from a linear test design can be turned into an MST design for the purpose of calibration.

In this paper, we have used the Rasch model in our examples. It should be clear that the method can easily be generalized to other exponential family models, e.g., the OPLM (Verhelst & Glas, 1995) and the *partial credit model* for polytomous items (Masters, 1982).

The design can also be generalized to more modules and more stages. It should however be kept in mind that estimation error with respect to the student parameters can be factorized into two components: the estimation

error of the student parameters conditional on the fixed item parameters, and the estimation error of the item parameters. The latter part is mostly ignored, which is defensible when it is very small compared to the former part. However, when stages are added, while keeping the total number of items per student fixed, more information about the item parameters is kept in the design, and therefore less information is left for item parameter estimation. A consequence is that the estimation error with respect to the the item parameters will increase. When many stages are added, it is even possible that the increase of estimation error of the item parameters is larger than the decrease of estimation error of the student parameters conditional on the fixed item parameters. An ultimate case is a CAT, in which all information about the item parameters is kept in the design and where no statistical information is left for the estimation of item parameters. This implies that adding more and more stages does not necessarily lead to more efficiency. Instead, there exists an optimal design with respect to the efficiency of the estimation of the student parameters. Finding the solution with respect to this optimum is left open for further research.

References

Andersen, E. B. (1973a). *Conditional inference and models for measuring.* Unpublished doctoral dissertation, Mentalhygiejnisk Forskningsinstitut.

Andersen, E. B. (1973b). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123-140.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 395-479). Reading: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. *Psychometrika*, *46*, 443-460.

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete designs. *Psychológica*, *32*, 107-132.

Glas, C. A. W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics*, *13*, 45-52.

Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models.* Unpublished doctoral dissertation, Arnhem: Cito.

Glas, C. A. W., Wainer, H., & Bradlow, E. (2000). MML and EAP estimation in testlet-based adaptive testing. In C. Van der Linden W.J. & Glas (Ed.), *Computerized adaptive testing: Theory and practice* (p. 271-287). Kluwer Academic Publishers.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44-52.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*, 189-202.

Luecht, R., & Nungester, R. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of educational measurement*, *35*,

229-249.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Mead, A. (2006). An Introduction to Multistage Testing. *Applied Measurement in Education*, *19*, 185-187.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1-32.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago, The University of Chicago Press)

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model: OPLM. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (p. 215-238). New York: Springer Verlag.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1993). *OPLM: One parameter logistic model.* Computer program and manual. Arnhem: Cito.

Wainer, H., Bradlow, E., & Du, Z. (2000). Testlet response theory: An analog for the 3pl model useful in testlet-based adaptive testing. In W. Van der Linden & C. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (p. 245-269). Kluwer Academic Publishers.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427-450.

Zenisky, A., Hambleton, R. K., & Luecht, R. (2010). Multistage testing: Issues, desings and research. In W. Van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (p. 355-372). Springer.