PSYCHOMETRIKA—VOL. 80, NO. 1, 65–84
MARCH 2015
DOI: 10.1007/S11336-013-9369-6

CONDITIONAL STATISTICAL INFERENCE WITH MULTISTAGE TESTING DESIGNS

ROBERT J. ZWITSER

CITO INSTITUTE FOR EDUCATIONAL MEASUREMENT

GUNTER MARIS

CITO INSTITUTE FOR EDUCATIONAL MEASUREMENT AND UNIVERSITY OF AMSTERDAM

In this paper it is demonstrated how statistical inference from multistage test designs can be made based on the conditional likelihood. Special attention is given to parameter estimation, as well as the evaluation of model fit. Two reasons are provided why the fit of simple measurement models is expected to be better in adaptive designs, compared to linear designs: more parameters are available for the same number of observations; and undesirable response behavior, like slipping and guessing, might be avoided owing to a better match between item difficulty and examinee proficiency. The results are illustrated with simulated data, as well as with real data.

Key words: multistage testing, adaptive testing, item response theory, parameter estimation, conditional maximum likelihood.

For several decades, test developers have been working on the development of adaptive test designs in order to obtain more efficient measurement procedures (Cronbach & Gleser, 1965; Lord, 1971a, 1971b; Weiss, 1983; Van der Linden & Glas, 2010). It is often shown that the better match between item difficulty and the proficiency of the examinee leads to more accurate estimates of person parameters.

Apart from efficiency, there are more reasons for preferring adaptive designs over linear designs, where all items are administered to all examinees. The first reason is that a good match between difficulty and proficiency might decrease the risk of undesirable response behavior. Examples of such behavior are guessing (slipping), which is an unexpected (in)correct response given the proficiency of the examinee. The avoidance of guessing or slipping might therefore diminish the need for parameters to model this type of behavior. This implies that adaptive designs could go along with more parsimonious models, compared to linear test designs. The second reason is that model fit is expected to be better for adaptive designs. Conditional on a fixed number of items per examinee, an adaptive design contains more items compared to a linear design. This implies that, although the number of possible observations is the same in both cases, the measurement model for an adaptive tests contains more parameters than the same measurement model for a linear test. An ultimate case is the *computerized adaptive test* (CAT, Weiss, 1983; Van der Linden & Glas, 2010) with an infinitely large item pool. A CAT with N dichotomous items has 2^N different response patters. Since the corresponding probabilities sum to one, the measurement model should estimate $2^N - 1$ probabilities. Observe, however, that the number of items in such a design is also $2^N - 1$. In a later section, we will show that in this case the Rasch model (Rasch, 1960) is a saturated model.

The usual approach for the calibration of a CAT is to fit an item response theory model on pretest data obtained from a linear test. The estimated item parameters are then considered as fixed during the adaptive test administration (Glas, 2010). This approach is valid if the item

Requests for reprints should be sent to Robert J. Zwitser, Psychometric Research Center, Cito Institute for Educational Measurement, P.O. Box 1034, 6801 MG, Arnhem, The Netherlands. E-mail: robert.zwitser@cito.nl

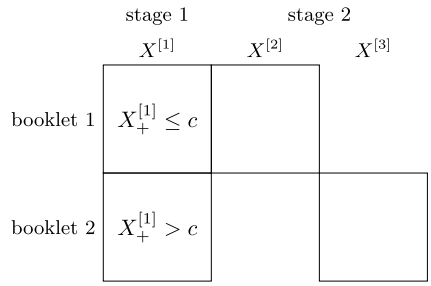


FIGURE 1.
Example of a multistage design.

parameters have the same values during the pretest administration and the actual adaptive test administration. However, factors like item exposure, motivation of the examinees, and different modes of item presentation may result in parameter value differences between the pretest stage and the test stage (Glas, 2000). This implies that, for accountability reasons, one should want to (re)calibrate the adaptive test after test administration.

In this paper, we go into the topic of statistical inference from adaptive test designs, especially multistage testing (MST) designs (Lord, 1971b; Zenisky, Hambleton, & Luecht, 2010). These designs have several practical advantages, as "the design strikes a balance among adaptability, practicality, measurement accuracy, and control over test forms" (Zenisky et al., 2010). In MST designs, items are administered in block/modules with multiple items. The modules that are administered to an examinee depends on their responses to earlier modules. An example of an MST design is given in Figure 1. In the first stage, all examinees take the first module.[1] This module is often called the *routing test*. In the second stage, examinees with a score lower than or equal to c on the routing test take module 2, whereas examinees with a score higher than c on the routing test take module 3. Every unique sequence of modules is called a booklet.

In the past, only a few studies have 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 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; Kubinger, Steinfeld, Reif, & Yanagida, 2012). The main purpose of this paper is to demonstrate that this conclusion was not correct. This will be shown in Section 1. In order to demonstrate the practical value of this technical conclusion, we elaborate on the relationship between model fit and adaptive test designs. In Section 2, we first show in more detail that the fit of the same measurement model is expected to be better for adaptive designs in comparison to linear designs. Then, second, we propose how the model fit can be evaluated. In Section 3, we give some illustrations to elucidate our results. 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.

[1] We use a superscript $[m]$ to denote random variables and parameters that relate to the mth module. Multiple modules, e.g., modules 1 and 2 are denoted by the superscript [1,2].

1. Conditional Likelihood Estimation

Throughout the paper, we use the Rasch model (Rasch, 1960) in our derivations and examples. Let $\mathbf{X}$ be a matrix with item responses of K examinees on N items. The model is defined as follows:

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

in which x_{pi} denotes the response of examinee p, $p=1,\ldots,K$, on item i, $i=1,\ldots,N$, and in which θ_p and b_i are parameters related to examinee p and item i, respectively. The θ-parameters are often called *ability* parameters, while the b-parameters are called *difficulty* parameters. The Rasch model is an exponential family distribution with the sum score

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

and

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

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 examinees 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 examinee'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.

1.1. Estimation of Item Parameters

Suppose that every examinee responds to all three modules ($\mathbf{X}^{[1]}$, $\mathbf{X}^{[2]}$, and $\mathbf{X}^{[3]}$). That is, we have complete data for every examinee. 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]

$$P_{\boldsymbol{b}}(\mathbf{x}|\theta)=P_{\boldsymbol{b}^{[1]}}\left(\mathbf{x}^{[1]}|x_+^{[1]}\right)P_{\boldsymbol{b}^{[2]}}\left(\mathbf{x}^{[2]}|x_+^{[2]}\right)P_{\boldsymbol{b}^{[3]}}\left(\mathbf{x}^{[3]}|x_+^{[3]}\right)$$
$$P_{\boldsymbol{b}}\left(x_+^{[1]},x_+^{[2]},x_+^{[3]}|x_+\right)P_{\boldsymbol{b}}(x_+|\theta)$$

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

where

$$P_{\boldsymbol{b}^{[m]}}\big(\mathbf{x}^{[m]}|x_+^{[m]}\big) = \frac{\prod_i \exp(-x_i^{[m]} b_i^{[m]})}{\gamma_{x_+^{[m]}}(\boldsymbol{b}^{[m]})}, \quad m = 1, 2, 3,$$

$$P_{\boldsymbol{b}}\big(x_+^{[1]}, x_+^{[2]}, x_+^{[3]}|x_+\big) = \frac{\gamma_{x_+^{[1]}}(\boldsymbol{b}^{[1]})\gamma_{x_+^{[2]}}(\boldsymbol{b}^{[2]})\gamma_{x_+^{[3]}}(\boldsymbol{b}^{[3]})}{\gamma_{x_+}(\boldsymbol{b})},$$

$$P_{\boldsymbol{b}}(x_+|\theta) = \frac{\gamma_{x_+}(\boldsymbol{b})\exp(x_+\theta)}{\sum_s \gamma_s(\boldsymbol{b})\exp(s\theta)},$$

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

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

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:

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

To turn a sample from $\mathbf{X}$ into a realization of data from MST, we do the following: *If* the score of an examinee 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 an examinee 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 $X_+^{[1]} \leq c$:

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

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:

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

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

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

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

where

$$\gamma_{x_+^{[1,2]}}\left(\boldsymbol{b}^{[1,2]}\right) = \sum_{j=0}^{n^{[1,2]}} \gamma_j\left(\boldsymbol{b}^{[1]}\right)\gamma_{x_+^{[1,2]}-j}\left(\boldsymbol{b}^{[2]}\right).$$

Second, we consider the probability that $X_+^{[1]}$ is lower than or equal to c conditionally on $X_+^{[1,2]}$:

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

Hence, we obtain

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

We next consider the distribution of $X_+^{[1,2]}$ 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:

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

we obtain

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

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

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

Obviously, a similar result holds for an examinee 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 probability.

1.2. 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 examinee's previous responses. This means that when an examinee receives a score lower than or equal to c on module 1, we use the probability of the observations conditionally on θ only

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

instead of the correct probability 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 3.1.2, we illustrate this phenomenon. If we compare the probability 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 probability 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 examinee.

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

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

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 $\mathbf{X}^{\text{obs}}$; and the design variable $X_+^{[1]} \leq c$. The complete probability of the observations can be written as follows:

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

From this, we immediately obtain the marginal likelihood function:

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

Since either $P(X_+^{[1]}\le c|\mathbf{x}^{[1]})=1$ and $P(X_+^{[1]}>c|\mathbf{x}^{[1]})=0$, or $P(X_+^{[1]}\le c|\mathbf{x}^{[1]})=0$ 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 if had we planned *beforehand* to which examinees 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 3.1.2, we show that misspecification of the population distribution can cause serious bias in the estimated item parameters.

1.3. Estimation of Person Parameters

In principle, it is straightforward to estimate the ability parameter θ of an examinee who was administered the second module by the maximum likelihood method from the distribution of the sufficient statistic $X_+^{[1,2]}$ conditionally on θ and the design:

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

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 an examinee 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.

2. Model Fit

We have mentioned in the Introduction that adaptive designs may be beneficial for model fit. The arguments were that adaptive designs could probably avoid different kinds of undesirable behavior, and that more parameters are available for the same number of observations. In the next paragraph, we elucidate the latter argument. Thereafter, in order to investigate the model fit, we propose two goodness of fit tests for MST designs.

2.1. Model Fit in Adaptive Testing

The Rasch model is known as a very restrictive model. Consider, for instance, the marginal model with a normal distribution for the person parameters. In a linear test design with N items, $2^N - 1$ probabilities are modeled with only $N + 1$ parameters (i.e., N item parameters, plus two parameters for the mean and standard deviation of the examinee population distribution (μ, and σ, respectively), minus one parameter that is fixed for scale identification, e.g., $\mu = 0$).

However, in the following example, we demonstrate that the Rasch model is less restrictive in cases with adaptive designs. For this example, consider a theoretically optimal CAT that selects items one-by-one from an infinitely large item pool. This implies that knots do not exist in the paths of the administration design. Consequently, a CAT of length two contains three items, a CAT of length three contains seven items, and so on: a CAT of length N contains $2^N - 1$ items.

Let us consider a CAT of length two. This design contains three items: one routing item, and two follow up items. Hence, we obtain five parameters in the model. This design has 2^2 possible outcomes. Since the probabilities of these four outcomes sum to one, the model describes $2^2 - 1$ probabilities with four parameters. This over-parameterization could be solved by fixing another parameters, for instance, by fixing σ to zero. With this fixation, we obtain the following probabilities:

$$
\begin{aligned}
P(X_1 = 0, X_2 = 0) = P_{00} &= \frac{1}{[1 + \exp(-b_1)][1 + \exp(-b_2)]};\\
P(X_1 = 0, X_2 = 1) = P_{01} &= \frac{\exp(-b_2)}{[1 + \exp(-b_1)][1 + \exp(-b_2)]};\\
P(X_1 = 1, X_3 = 0) = P_{10} &= \frac{\exp(-b_1)}{[1 + \exp(-b_1)][1 + \exp(-b_3)]};\\
P(X_1 = 1, X_3 = 1) = P_{11} &= \frac{\exp(-b_1 - b_3)}{[1 + \exp(-b_1)][1 + \exp(-b_3)]}.
\end{aligned}
$$

These equations could be transformed into the following equations for b_1, b_2, and b_3:

$$
\begin{aligned}
b_1 &= -\log\left(\frac{P_{10} + P_{11}}{P_{01} + P_{00}}\right);\\
b_2 &= -\log\left(\frac{P_{01}}{P_{00}}\right);\\
b_3 &= -\log\left(\frac{P_{11}}{P_{10}}\right).
\end{aligned}
$$

Two things are worth to be noticed. First, we can see the model is saturated. Second, since σ was fixed to zero, the model results in person parameters that are all equal, which is remarkable in a measurement context. This taken together demonstrates nicely that the Rasch model is not suitable for statistical inference from a CAT. It could easily be shown the same conclusion holds for extensions to N items.

For MST designs, we easily find that the Rasch model is less restrictive compared to linear designs. Consider, for instance, a test of four items per examinee. In a linear design, we obtain fifteen probabilities and five parameters. However, for the MST design with two stages with and two items within each stage, we have six items (seven parameters) to model fifteen observation. Since model restrictiveness is a ratio of the number of possible observations and the number of parameters we see that the same model can be more or less restrictive, depending on the administration design.

2.2. Likelihood Ratio Test

In order to evaluate model fit, we propose two tests that are based on the method that was suggested by Andersen (1973b). He showed that the item parameters $\boldsymbol{b}$ can be estimated by maximizing the conditional likelihood

$$\mathcal{L}(\boldsymbol{b}) = \frac{\exp(-\sum_{p=1}^{K}\sum_{i=1}^{N} b_i x_{pi})}{\prod_{p=1}^{K}\gamma_{x_{+p}}(\boldsymbol{b})},$$

as well as by maximizing $\mathcal{L}^{(t)}(\boldsymbol{b})$, which is the likelihood for the subset of data for which holds that $X_+ = t$. This conclusion has led to the following likelihood ratio test (LRT): In the general model, item parameters were estimated for all score groups separately, while in the special model, only one set of item parameters was estimated for all score groups together. For a complete design with N items, Andersen (1973b) considered

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

as the test statistic, in which $\hat{\boldsymbol{b}}^{(t)}$ are the estimates that are based on the subset of data with examinees that have a total score equal to t.

Let us denote K_t as the number of examinees with sum score t. It is shown that if $K_t \to \infty$ for $t = 1, \ldots, N-1$, 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 general model and the specific model.

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

$$Z = 2\sum_{g}\sum_{t=1}^{N_g-1}\log\bigl[\mathcal{L}^{(gt)}\bigl(\hat{\boldsymbol{b}}^{(gt)}\bigr)\bigr] - 2\log\bigl[\mathcal{L}(\hat{\boldsymbol{b}})\bigr], \tag{9}$$

where N_g denotes the number of items in booklet g, $\mathcal{L}^{(gt)}(\hat{\boldsymbol{b}}^{(gt)})$ denotes the likelihood corresponding to the subset of data with examinees that took booklet g and obtained a total score t, and $\hat{\boldsymbol{b}}^{(gt)}$ denotes the estimates based on this subset of data. 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 specific 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 gray in Figure 2) do not provide statistical information and therefore the number of parameters estimated in the general 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

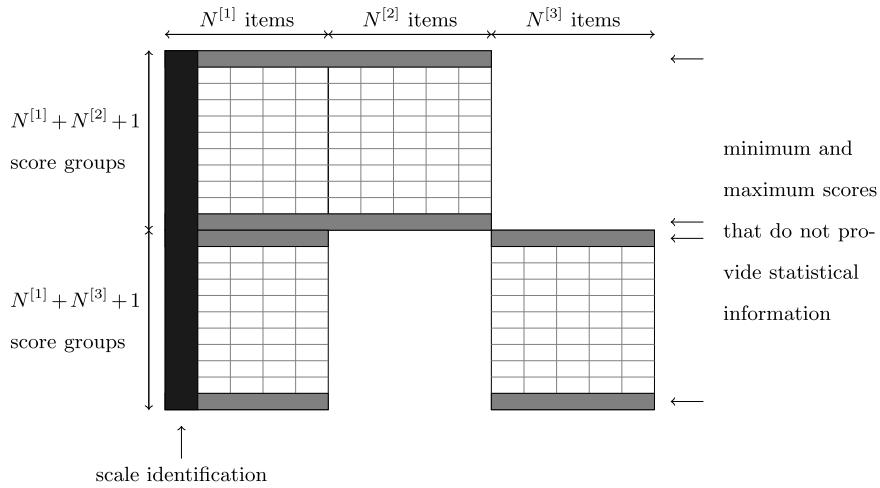


FIGURE 2.
Degrees of freedom in a general booklet design.

of degrees of freedom is

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

The number of parameters of the general 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 $\boldsymbol{b}^{[1,2]}$, 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 general model is $(c+N^{[2]})(N^{[1]}-1)+(c+N^{[2]}-1)N^{[2]}+(N^{[1]}+N^{[3]}-c-1)(N^{[1]}-1)+(N^{[1]}+N^{[3]}-c-2)N^{[3]}$. Therefore, the number of degrees of freedom is

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

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 number of parameters under the general 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 little observations. To increase the power, the procedure can also be based on W sets of sum scores instead of single values t. Then

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

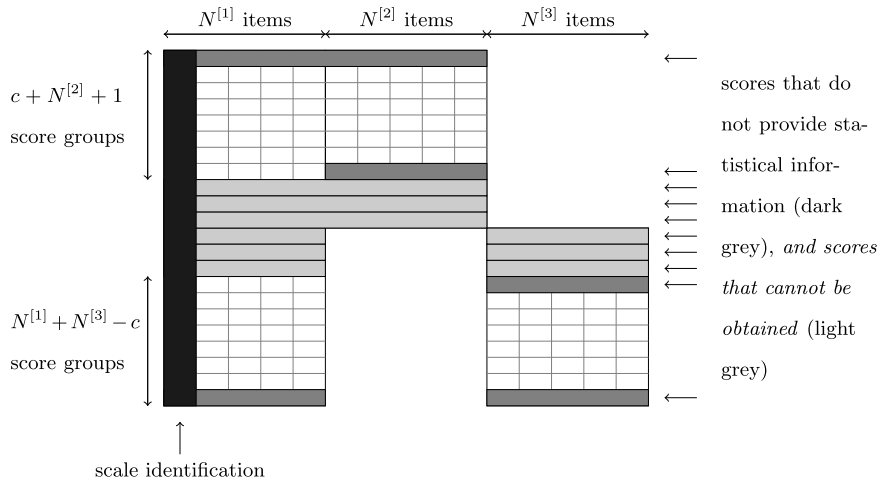


FIGURE 3.
Degrees of freedom in an MST design.

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

2.3. *Item Fit Test*

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

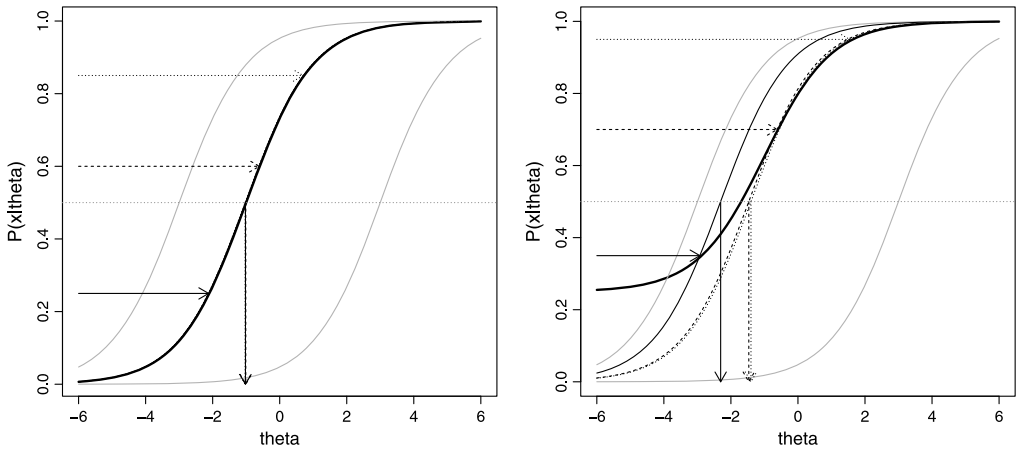
$$\hat{\boldsymbol{b}}^{(S_v)} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\boldsymbol{b}^{(S_v)}, \boldsymbol{\Sigma}^{(S_v)}\right),$$

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

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

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.



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

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

3. Examples

In this section, we demonstrate some properties of CML inference in MST. After a short description of the simulation design (Section 3.1.1), 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 3.1.2). In Sections 3.1.3 and 3.1.4, we will demonstrate, respectively, the robustness and efficiency of the MST design. Finally, in Section 3.2, we will demonstrate with real data the benefits of MST on model fit.

3.1. Simulation

3.1.1. 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 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.

3.1.2. Comparison of Methods In the first example, 10,000 examinees were sampled and the test was administered in an MST design. The item parameters were estimated according to three methods: first, according to the correct conditional likelihood as in (3); second, according to an ordinary CML method that takes into account the incomplete design, but not the multistage aspects of the design; and third, the MML method, in which the person parameters are assumed to be normally distributed. The scales of the different methods were equated by fixing the first item parameter at zero.

The average bias, standard errors (SE), and root mean squared error (RMSE) are per method and per module displayed Table 1. Both ordinary CML and MML inference lead to serious bias in the estimated parameters. The standard errors were nearly the same between the three methods.

TABLE 1.
Average bias, standard error (SE), and root mean squared error (RMSE) of the item parameters per module.

	method	module 1	module 2	module 3
BIAS($\hat{\delta}, \delta$)	MST CML	0.000	−0.001	−0.001
	Ordinary CML	0.001	−0.089	0.291
	Ordinary MML	−0.003	0.097	−0.345
SE($\hat{\delta}$)	MST CML	0.033	0.036	0.055
	Ordinary CML	0.034	0.037	0.052
	Ordinary MML	0.030	0.035	0.052
RMSE ($\hat{\delta}$)	MST CML	0.033	0.036	0.055
	Ordinary CML	0.043	0.096	0.295
	Ordinary MML	0.047	0.104	0.349

Therefore, finally, the RMSEs of the proposed CML method are much lower than the RMSEs of the ordinary CML and MML methods.

3.1.3. Goodness of Fit In a second simulation study, we demonstrated the model fit procedure that is described in Section 2. The simulation consisted of 1,000 trials. In each trial, three different cases were simulated.

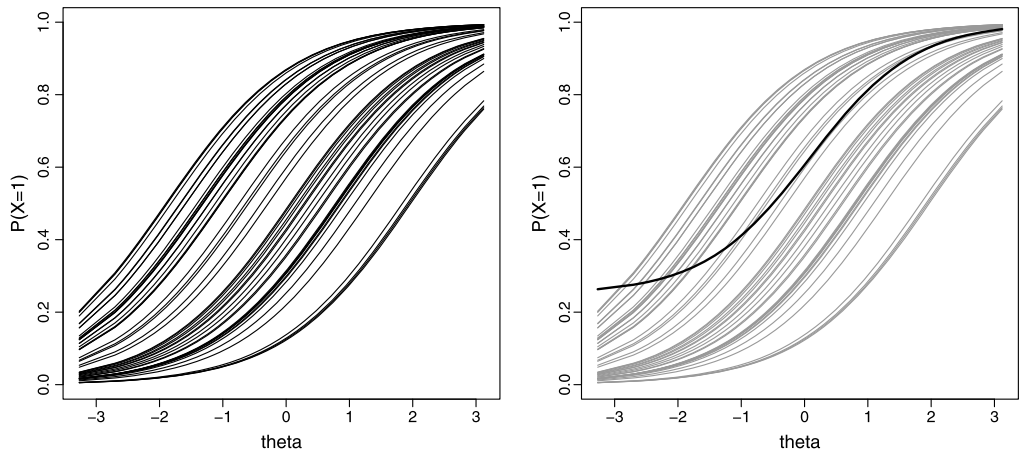
- Case 1: the MST design described above.
- Case 2: a complete design with all 50 items, except for the easiest item in module 3. The excluded item was replaced by an item according to the *3-parameter logistic model* (3PLM, Birnbaum, 1968) which is defined as follows:

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

  where, compared to the Rasch model, a_i and c_i are additional parameters for item i. This 3PLM item has the same item difficulty (i.e., the b-parameter) as the excluded item. However, instead of $a=1$ and $c=0$, which would make (11) equal to (1), we 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 examinees 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 examinees were equally distributed over the different score groups. The number of degrees of freedom in cases 1 and 3 is

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



(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.

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

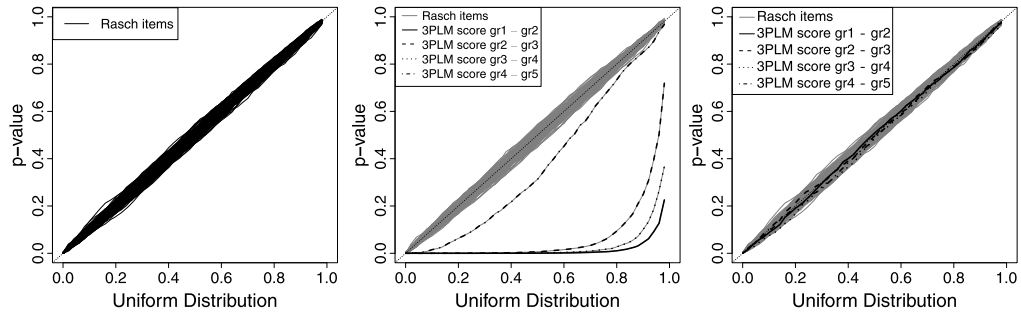
Case	D^-	p-value
Case 1	0.016	0.774
Case 2	0.968	<0.001
Case 3	0.048	0.100

and in case 2

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

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 cases 1 and 3, but not in case 2.

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 2.3, 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 groups 4 and 5 are nearly uniformly distributed.



(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.

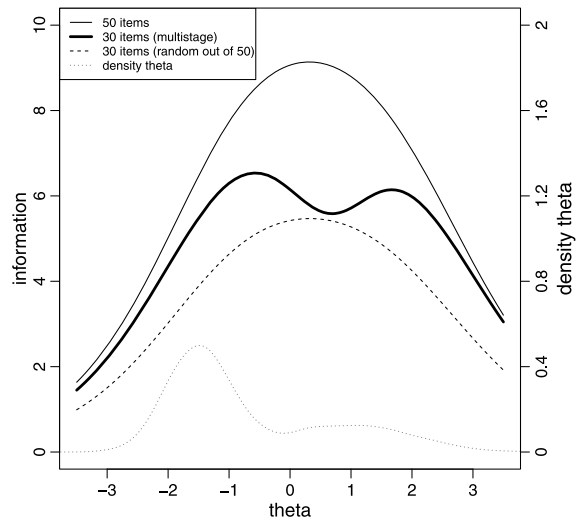


FIGURE 7.
Person information in a complete design with 50 items, an MST design with 30 items, and a complete design with 30 items.

3.1.4. 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

$$I(\theta) = I^{[1,2]}(\theta)P\left(X_+^{[1]} \leq c|\theta\right) + I^{[1,3]}(\theta)P\left(X_+^{[1]} > c|\theta\right).$$

Here, $I^{[1,2]}(\theta)$ denotes the Fisher information function for modules 1 and 2. The distribution of θ is also shown in Figure 7. It can be seen that, for most of the examinees 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 examinees, the information based on the MST is not much less than the information based on all 50 items.

3.2. Real Data

The data for the following examples were taken from the Dutch *Entrance Test* (in Dutch: *Entreetoets*), which consists of multiple parts that are administered annually to approximately 125,000 grade 5 pupils. In this example, we took the data from 2009, which consists of 127,746 examinees. 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 examinees[3] with the *One-Parameter Logistic Model* (OPLM, Verhelst & Glas, 1995; Verhelst et al., 1993). The program contains routines to estimate integer item discrimination parameters, as well as item difficulty parameters.

The examples in this section illustrate the two factors by which model fit could be improved with MST designs. First, the difference in restrictiveness of the same model in different administration designs, and second, the avoidance of guessing owing to a better match between item difficulty and examinee proficiency.

3.2.1. Better Fit Owing to More Parameters In Section 2.1, we explained that the restrictiveness of measurement models depends on the administration design. In order to demonstrate this, two small examples are given.

In the first example, nine items were randomly selected from the set of 120 math items. The items were sorted based on the proportion correct in the original data set. Then they were assigned to two designs:

- a MST design with the three most easy items in module 2, the three most difficult items in module 3, and the remaining three items in module 1;
- a linear test design with six items, namely the first two of each module.

In the MST design, module 2 will be administered to examinees with a sum score 0 or 1 on module 1, while module 3 will be administered to examinees with a sum score 2 or 3 on module 1. Observe that in both designs six items are administered to each examinee, so in both cases 64 (2^6) different response patters could occur. However, in the MST case, the Rasch model has 8 free item parameters to model 63 probabilities, while in the linear test only 5 free parameters are available. Since the number of different score patterns is limited, model fit could be evaluated by a comparison between the observed frequencies (O), and the expected frequencies according to the model (E). The difference between the two could be summarized with the *total absolute difference* (TAD):

$$TAD = \sum_{\mathbf{x}} |O_{\mathbf{x}} - E_{\mathbf{x}}|,$$

in which $O_{\mathbf{x}}$ and $E_{\mathbf{x}}$ are the observed and expected frequency of response pattern $\mathbf{x}$.

The sampling of items was repeated in 1,000 trials. In each trial, parameters of both designs were estimated and the TAD for both designs was registered. The mean TAD over the 1,000 trials was 11,317 for the linear design, while it was 9,432 for the MST design.

In the second example, nine particular items were selected. The item characteristics of these items in the original test, based on the OPLM model (Verhelst & Glas, 1995), are displayed in Table 3. The focus in this example is not on the variation in b-parameters, but on the variation in a-parameters. With these nine selected items, two administration designs are simulated:

1. a linear test with item 30, 66, 79, 85, 110, and 118
2. a MST with the following modules:

[3] A sample had to be drawn because of limitations of the OPLM software package w.r.t. the maximum number of observations.

TABLE 3.
Item characteristics of nine selected math items in Example 1

Item no.	a_i	b_i	proportion correct	Item no.	a_i	b_i	proportion correct
19	4	0.311	0.529	85	4	0.247	0.576
30	2	0.311	0.520	88	3	0.194	0.597
66	2	0.334	0.510	110	3	0.372	0.488
79	3	0.435	0.450	118	4	0.254	0.571
83	2	0.402	0.479				

- module 1 (routing test): item 79, 88, and 110 ($a_i = 3$)
- module 2: item 30, 66, and 83 ($a_i = 2$)
- module 3: item 19, 85, and 118 ($a_i = 4$)

Observe that for an individual examinee the maximum difference in a-parameters is two within the linear test, while it is only one within a booklet of the MST. We expect that the model fit is better in the second case, because we avoid that items that have large differences in a-parameters are assigned to the same examinee.

For both cases, the Rasch model was fitted on the data of the total sample of 127,746 examinees. The LRTs, based on two score groups, confirm the lack of fit of both cases, $Z(5) = 1{,}660.16$, $p < 0.001$, and $Z(12) = 139.44$, $p < 0.001$, respectively. However, the ratio Z/df indicates that the fit of the Rasch model is substantially better in the MST design compared to the linear design. This observation is confirmed by the TAD statistics. The TAD of the linear test was 15,376, while the TAD of the MST was 4,453.

3.2.2. Better Fit Owing to Avoidance of Guessing For the following example, we have selected 30 items that seem to have parallel ICCs, although the LRT, based on two score groups, indicated that the Rasch model did not perfectly fit, $Z(29) = 400.93$, $p < 0.001$. 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 examinees with a low (high) score on the routing test are removed from the third (second) module.

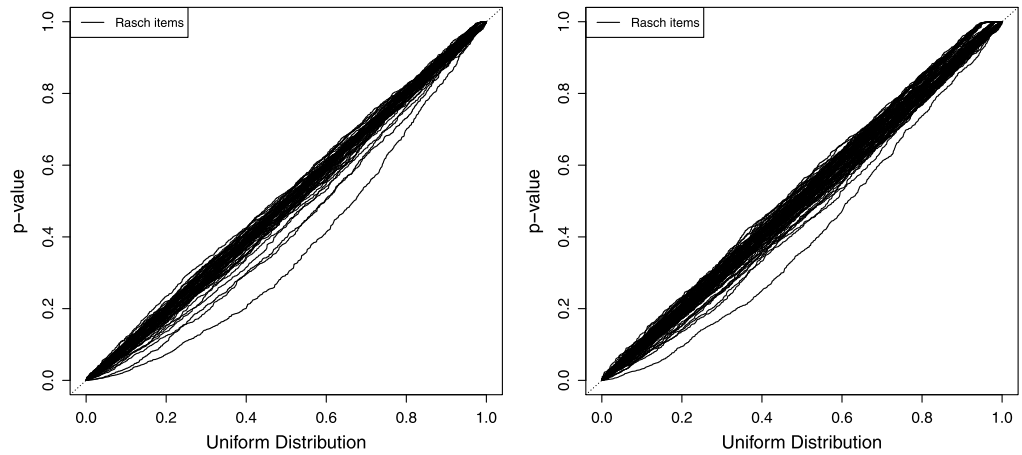
In order to demonstrate the item fit tests, we drew 1,000 samples of 1,000 examinees from the data. First, we estimated the parameters of the 30 Rasch items with a complete design and an MST design. In both cases, all items seem to fit the Rasch model reasonably well (see Figure 8a and Figure 8b).

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 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).

4. Discussion

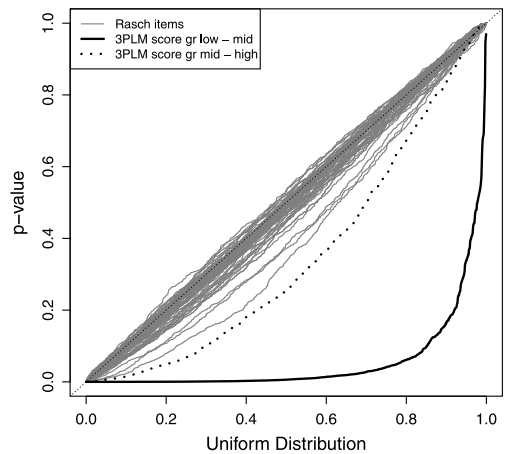
In this paper, we have shown that the CML method is applicable with data from an MST. We have demonstrated how item parameters can be estimated for the Rasch model, and how model fit can be investigated for the total test, as well as for individual items.

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

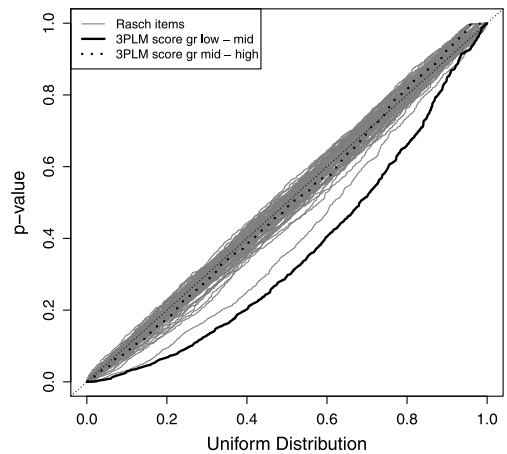


(a) Rasch - Complete design

(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.

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 can 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.

In this paper, we have used the Rasch model in our examples. Although the Rasch model is known as a restrictive model, we have emphasized that the Rasch model is less restrictive in adaptive designs compared to linear designs. However, if more complicated models are needed, then 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).

Our presumption was that adaptive designs are more robust against undesirable behavior like guessing and slipping. This has been illustrated by the simulation in Section 3.1.3. The fit for case 1 and the lack of fit for case 2 were as 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 was only administered

to examinees with a high score on the routing test, i.e., examinees with a high proficiency level. In general, it could be said that changing the measurement model into a more complicated model is not the only intervention possible in cases of misfit. Instead, the data generating design could be changed. The example with real data in Section 3.2 did show that this could also be done afterward. This means 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. However, this raises the question how to estimate person parameters in this approach. Should they be based on all item responses, or only the multistage part with which the item parameters were estimated? The answer to this question is left for future research.

The design can also be generalized to more modules and more stages, as long as the likelihood on the design contains statistical information about the item parameters. It should however be kept in mind that estimation error with respect to the person parameters can be factorized into two components: the estimation error of the person 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 examinee 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 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 person 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 person 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*. 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* (pp. 395–479). Reading: Addison-Wesley.

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

Cronbach, L.J., & Gleser, G.C. (1965). *Psychological test and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

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. (2000). Item calibration and parameter drift. In W.J. Van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: theory and practice* (pp. 183–199). Dordrecht: Kluwer Academic Publishers.

Glas, C.A.W. (2010). Item parameter estimation and item fit analysis. In W.J. Van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 269–288). Berlin: Springer.

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

Kubinger, K.D., Steinfeld, J., Reif, M., & Yanagida, T. (2012). Biased (conditional) parameter estimation of a Rasch model calibrated item pool administered according to a branched testing design. *Psychological Test and Assessment Modeling*, *52*(4), 450–460.

Lord, F.M. (1971a). The self-scoring flexilevel test. *Journal of Educational Measurement*, *8*(3), 147–151.

Lord, F.M. (1971b). A theoretical study of two-stage testing. *Psychometrika*, *36*, 227–242.

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

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.
Van der Linden, W.J. & Glas, C.A.W. (Eds.) (2010). *Elements of adaptive testing*. New York: Springer.
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* (pp. 215–238). New York: Springer.
Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1993). *OPLM: one parameter logistic model*. Arnhem: Cito. Computer program and manual.
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* (pp. 245–269). Dordrecht: Kluwer Academic Publishers.
Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.
Weiss, D.J. (Ed.) (1983). *New horizons in testing: latent trait test theory and computerized adaptive testing*. New York: Academic Press.
Zenisky, A., Hambleton, R.K., & Luecht, R. (2010). Multistage testing: issues, designs and research. In W.J. Van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). Berlin: Springer.

Manuscript Received: 28 MAR 2012
Published Online Date: 6 DEC 2013