Measurement and Research Department Reports 2000–5

A SELECTION PROCEDURE FOR POLYTOMOUS ITEMS IN COMPUTERIZED ADAPTIVE TESTING

P.W. van Rijn
T.J.H.M. Eggen
B.T. Hemker
P.F. Sanders

Cito

Measurement and Research Department Reports 2000–5

A Selection Procedure for Polytomous Items in Computerized Adaptive Testing

P.W. van Rijn
T.J.H.M. Eggen
B.T. Hemker
P.F. Sanders

Cito
Arnhem, september 2000

Abstract

In the present study, a procedure which was developed to select dichotomous items in computerized adaptive testing was applied to polytomous items. The aim of this procedure is to select the item with maximum weighted information. In a simulation study, the item information function was integrated over a fixed interval of ability values and the item with the maximum area was selected. This maximum interval information item selection procedure was compared to a maximum point information item selection procedure. No substantial differences between the two item selection procedures were found when computerized adaptive tests were evaluated on bias and root mean square of the ability estimate.

Index terms: CAT, GPCM, polytomous items, item selection, Fisher information function, decomposition.

1 Introduction

Computerized adaptive testing is one of the major developments of item response theory (IRT). Its main advantages over traditional paper and pencil tests are that the same precision can be achieved with shorter tests and that each examinee is given a test adapted to his or her ability (Weiss, 1982). Until recently, most of the research and applications in computerized adaptive testing concerned dichotomous items and not polytomous items (Dodd, De Ayala & Koch, 1995). While item selection, which is an essential component of computerized adaptive testing, has received considerable attention in the last few years (Eggen, 1999; Van der Linden, 1998; Berger & Veerkamp, 1997; Tang, 1996; Chang & Ying, 1996), only a few studies have dealt with the selection of polytomous items in computerized adaptive testing.

There are two main approaches to item selection. The first is item selection based on item information, where the most informative item at a certain estimated ability level is selected. Fisher information is the most commonly used form of information used in item selection (Berger & Veerkamp, 1997), though Kullback-Leibler information has also been used (Eggen, 1999; Chang & Ying, 1996). The second approach to select an item is Bayesian item selection in which a prior or posterior distribution of ability is used in combination with a Bayesian variant of information (see Van der Linden, 1998). This study focuses on item selection based on Fisher information with no specific prior distribution and its purpose is to use Fisher information more optimally to estimate ability. Some research has already been done on Fisher information functions of polytomous items (Muraki, 1993; Donoghue, 1994; Akkermans & Muraki, 1997), the results of which have been used in the present study.

There are many IRT models for polytomous items (e.g., Hemker, Sijtsma, Molenaar & Junker, 1997; Dodd et al., 1995; Mellenbergh, 1995; Thissen & Steinberg, 1986). One of the most commonly used models is the generalized partial credit model (GPCM; Muraki, 1992) which is an extension of Masters' (1982) partial credit model (PCM). The PCM, in contrast to the GPCM, has been investigated in some computerized adaptive testing applications (Dodd et al., 1995; Baek, 1997).

In the present study, two item selection procedures for the GPCM and their effects on accuracy and precision of the ability estimates were investigated. One item selection procedure was the procedure that selects the most informative item at the current estimate of the examinee's

ability. The other item selection procedure which was investigated is an application of Berger and Veerkamp's (1997) general weighted information criterion to polytomous items. This procedure amounts to selecting the item with the maximum of a weighted average of information function values, e.g., the maximum of an area under the information function. The latter seems an obvious choice since GPCM item information functions do not necessarily have to be single-peaked (Muraki, 1993). Selecting items on the basis of a single point of the information function may therefore lead to the selection of non-optimal items. Another advantage of using an area under the item information function is that the uncertainty of the ability estimate can be taken into account. Especially in the beginning of a CAT, this uncertainty can be quite considerable (Berger & Veerkamp, 1997; Chang & Ying, 1996).

The description of the GPCM, followed by information functions of GPCM items, is presented in Section 2 of this paper. Subsequently, an important feature of the GPCM, i.e., decomposition of polytomous items and its consequences on item information is discussed. Some item selection procedures for dichotomous items are then reviewed, followed by the presentation of selection procedures for polytomous items. Finally, a simulation study on the different item selection procedures and its results are presented and discussed.

2 Generalized Partial Credit Model

A score $X_i = 0, 1, \ldots, m_i$ can be obtained on item $i = 1, \ldots, B$ from an item bank with B items. A higher score indicates a better performance and m_i indicates the maximum score on item i. The probability of obtaining a score k on item i, given the value of the ability θ, is denoted by where $k = 0, 1, \ldots, m_i$. In the GPCM (Muraki, 1992), it is assumed that the probability of obtaining

$$P_{ik}(\theta) = \Pr(X_i = k \mid \theta),$$

$$P_{ik|k-1,k}(\theta) \equiv \frac{P_{ik}(\theta)}{P_{i,k-1}(\theta) + P_{ik}(\theta)} = \frac{\exp[a_i(\theta - b_{ik})]}{1 + \exp[a_i(\theta - b_{ik})]}, \tag{2.1}$$

a score k on item i, given that the score is k or k-1 and given the value of the ability θ, is governed by the logistic function, that is,

where a_i is a slope parameter and b_{ik}, $k = 0,\ldots, m_i$, are the item category parameters. The probability of obtaining a score k is

$$P_{ik}(\theta) = \frac{\exp[\sum_{v=0}^{k} a_i(\theta - b_{iv})]}{\sum_{c=0}^{m_i} \exp[\sum_{v=0}^{c} a_i(\theta - b_{iv})]}, \tag{2.2}$$

for $k = 0, 1,\ldots, m_i$, and $b_{i0} \equiv 0$. Equation 2.2 is called the item category response function (ICRF). If $m_i = 1$, the model reduces to Birnbaum's (1968) two parameter logistic (2-PL) model. If all a_i are equal, the GPCM reduces to the PCM. These two restrictions combined yield the Rasch model.

The first derivative of the ICRF with respect to θ is given by

$$\frac{\partial P_{ik}(\theta)}{\partial \theta} = a_i P_{ik}(\theta)[k - \sum_{c=0}^{m_i} c P_{ic}(\theta)], \tag{2.3}$$

and the likelihood of the responses on n GPCM items is

$$L = \prod_{i=1}^{n} P_{ik_i}(\theta). \tag{2.4}$$

Here, k_i is the score on item i with maximum score m_i. Both Equation 2.3 and 2.4 are needed to obtain the information function in the next section.

3 Information Function

The test information function (TIF) for n polytomous items is defined as (Samejima, 1969) For a single item i, $I_i(\theta)$ is the item information function which equals (Muraki, 1993)

$$I(\theta) = -E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right) = \sum_{i=1}^{n}\left[-E\left(\frac{\partial^2 \ln P_{ik}(\theta)}{\partial \theta^2}\right)\right] = \sum_{i=1}^{n} I_i(\theta). \tag{3.1}$$

$$I_i(\theta) = \sum_{k=0}^{m_i} P_{ik}(\theta)\left[-\frac{\partial^2}{\partial \theta^2}\ln P_{ik}(\theta)\right] = \sum_{k=0}^{m} P_{ik}(\theta)\left\{\left[\frac{\frac{\partial}{\partial \theta}P_{ik}(\theta)}{P_{ik}(\theta)}\right]^2 - \frac{\frac{\partial^2}{\partial \theta^2}P_{ik}(\theta)}{P_{ik}(\theta)}\right\}. \tag{3.2}$$

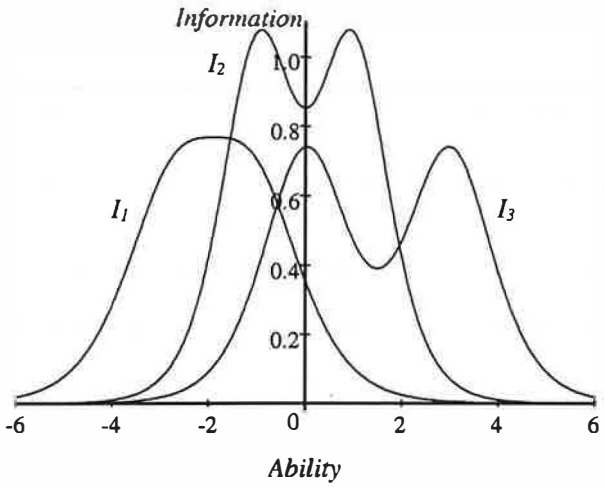
It can be derived for the GPCM that this can be written as (Donoghue, 1994)

$$I_i(\theta) = a_i^2\left[\sum_{k=0}^{m_i} k^2 P_{ik}(\theta) - \left(\sum_{k=0}^{m_i} k P_{ik}(\theta)\right)^2\right]. \tag{3.3}$$

Muraki (1993) showed that the IIF of a GPCM item does not necessarily have to be single-peaked, but can have as many peaks as there are estimated item category parameters (b_{ik}). Multi-modality of IIFs (i.e., with more than one peak) can only take place if item category parameters are sequentially ordered. Akkermans and Muraki (1997) showed for trinary GPCM items that if the second and first item category parameter differ by more than $4\ln 2/a_i$, i.e., $a_i(b_{i2} - b_{i1}) \geq 4\ln 2$, the IIF becomes bimodal. Both the slope parameter a_i and the b_{ik}'s determine the locations of one or more of the peaks. In Figure 1, information is plotted as a function of the ability θ of three trinary GPCM items. The item parameters are: $a_1 = 1.5$, $b_{11} = -2.8$, and $b_{12} = -1.0$; $a_2 = 2.0$, $b_{21} = -1.0$, and $b_{22} = 1.0$; $a_3 = 1.7$, $b_{31} = 0.0$, and $b_{32} = 3.0$. Unimodal as well as bimodal item information functions are plotted.

Figure 1

Information functions of three trinary GPCM items.



The maximum value of the IIF is often used in item selection based on Fisher information. For trinary GPCM items, the maximum of the IIF decreases in $b_{i2} - b_{i1}$. Some maxima for a trinary GPCM item i with different relations between the item category parameters are given below (adapted from Akkermans & Muraki, 1997):

$$\begin{aligned}
&\max I_i(\theta) \approx a_i^2 && \text{for } a_i(b_{i2} - b_{i1}) \to -\infty, && \text{unimodal,}\\
&\max I_i(\theta) = \frac{2}{3}a_i^2 && \text{for } a_i(b_{i2} - b_{i1}) = 0, && \text{unimodal,}\\
&\max I_i(\theta) = \frac{1}{2}a_i^2 && \text{for } a_i(b_{i2} - b_{i1}) = 2\ln 2, && \text{unimodal,}\\
&\max I_i(\theta) = \frac{1}{3}a_i^2 && \text{for } a_i(b_{i2} - b_{i1}) = 4\ln 2, && \text{bimodal.}
\end{aligned} \tag{3.4}$$

The value of the slope parameter a_i is of particular importance because it has a linear effect on the difference between the item category parameters and a quadratic effect on the value of the maximum of the IIF.

4 Decomposition of Polytomous Items

Decomposition of a polytomous item is the splitting up of the IRF of an item score X_i into H (with $2 \leq H \leq m_i$) items $i_1, i_2, \ldots i_H$, where the sum of the separate IRFs yields the original IRF of these H items. The total maximum score of the original item and the sum of the item scores in the decomposition remains the same (i.e. , $m_i = \sum_{h=1}^{H} m_{i_h}$, $m_i \geq H$ and $m_i > m_{i_h}$ for all h). Huynh (1994) showed that a trinary PCM item i can be decomposed into two locally independent dichotomous (Rasch) items (i.e., $m_{i_1} = m_{i_2} = 1$) if the second and first item category parameter differ by more than 2ln2, i.e., $b_{i2} - b_{i1} \geq 2\ln 2$. This is a sufficient condition for decomposition. The likelihood functions for estimating the ability from the original item and those from the items in the decomposition are the same (see Equation 2.4). In a subsequent article, Huynh (1996, Definition 2) defined a trinary PCM item to be indecomposable if $b_{i2} - b_{i1} < 2\ln 2$. Furthermore, Huynh (1996) showed that a PCM item with any number of categories can always be decomposed into independent Rasch items and indecomposable trinary items (i.e., $m_{i_h} \in \{1, 2\}$).

Huynh (1996) noted that it is not known whether this decomposition can be generalized to more complex models for partial credit items, like the GPCM, where there is no simple sufficient statistic as there is in the case of the PCM. However, it should be noted that the sufficient statistic is not used in the proofs presented by Huynh (1994, 1996) and decomposition concerns one item at a time. Since the slope parameter a_i may be considered a constant for a single item, a_i seems to pose no problems for the derivations leading to (in)decomposability as long as a_i remains the same for the items in the decomposition. The difference between the second and first item category parameter (b_{i2} and b_{i1}) of a trinary GPCM item has to become more than $2\ln 2 / a_i$ to bring about decomposability.

Because of the decomposition property, the maximum information given by a PCM item with more than three categories cannot exceed that of the sum of its decomposition (Akkermans & Muraki, 1997). This also holds for the GPCM (see above). The information functions of the original item and of its decomposition are the same because the likelihood functions for estimating ability are equal. The amount of information yielded by an indecomposable trinary GPCM item, however, can exceed the information obtainable with any two independent dichotomous items,

given equal a_i's. The maximum information to be obtained with any two dichotomous items is ½ a_i^2 (Hambleton & Swaminathan, 1985, p. 107) and the maximum information of an indecomposable trinary item is at least ½ a_i^2 (see Equation 3.4).

These points about the information of trinary items and items with any number of categories are important when considering computerized adaptive testing and more specific, polytomous item selection, since items with more categories generally provide more information and thus are selected more often. If an item with more than three categories cannot give more information than its decomposition, it is questionable whether it should be used in a CAT. The use of binary and trinary items can result in CATs with the same TIF as the use of items with various (high) m_i's, so it might be more appropriate to administer binary and trinary items in a CAT. The administration of indecomposable trinary items in a CAT can give more information than any two independent dichotomous items and may therefore be more efficient.

5 Item Selection

First, a short overview of Fisher information-based item selection procedures for dichotomous items is given. Then, the item selection procedures for polytomous items are presented and discussed.

Selection Procedures for Dichotomous Items

Note that the notation which is used here has been simplified because only one test with fixed length is considered. Let the subscript $u = 1,\ldots, n$ indicate the order of items in the test where n is the number of items in a test. Let $S_{u-1} = \{i_1,\ldots, i_{u-1}\}$ be the set of items selected, and $R_u = \{1,\ldots, B\} \setminus S_{u-1}$ be the remaining set of items.

The most common item selection procedure, the maximum point (Fisher) information item selection criterion (MPI) selects as the next item i_u the most informative item j at the current ability estimate from R_u. This can be represented as

$$i_u = \max_j I_j(\hat{\theta}),\ j \in R_u, \tag{5.1}$$

where $\hat{\theta}$ is the current ability estimate. An important reason for using this selection criterion is that the reciprocal of the TIF evaluated at the estimation of ability θ is an estimate of the asymptotic variance of the maximum likelihood estimator of θ (Hambleton & Swaminathan, 1985, p. 89).

Berger and Veerkamp (1997) proposed a general weighted information criterion, which is defined as

$$i_u = \max_j \int_{-\infty}^{\infty} W_{u-1}(\mathbf{x}_{u-1};\theta) I_j(\hat{\theta}) d\theta, \quad j \in R_u, \tag{5.2}$$

where vector $\mathbf{x}_{u-1}$ denotes the responses to previously administered items. W_{u-1} ($\mathbf{x}_{u-1}$; θ) can be any kind of weight function in which all previous responses are used. Several item selection criteria can be defined by using different kinds of weight functions. For example, a step function or a more complicated function like the likelihood function of θ can be used. MPI is a special case of the general weighted information criterion, i.e., weighing the current ability estimate with one and all other ability values with zero. One problem in using the integral of the IIF is that the total area under the information function of a dichotomous 2-PL item is equal to a_i (Birnbaum, 1968, p. 460). When no weight function is used, item selection depends only on a_i, so the test is no longer adaptive. In the simulation study performed by Berger and Veerkamp (1997), the use of the general weighted information criterion resulted in CATs with more test information and less mean squared error of the ability estimates than the use of the MPI criterion when the likelihood function of the ability was used as a weight function. The differences between the criteria, however, were small. Because of the specific (non-unimodal) form of the IIFs and the (in)decomposability issue, the weighted information criterion might be more useful in the polytomous case.

Selection Procedures for Polytomous Items

Two selection procedures for polytomous items were investigated using Berger and Veerkamp's (1997) general weighted information criterion. These are the maximum point Fisher information criterion (MPI) in Equation 5.1 and the maximum interval Fisher information criterion (MII) which can be represented as

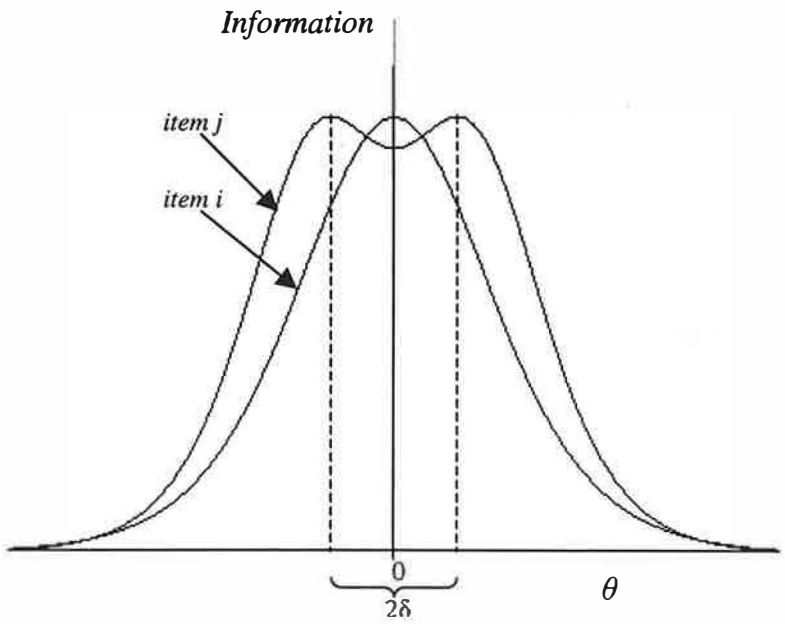
$$i_u = \max_j \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} I_j(\theta)d\theta,\ j \in R_u, \tag{5.3}$$

where δ determines the width of the interval of integration. The value of this integral is the product of the average of the information function values in the interval and the interval width. Since information functions of GPCM items do not have to be single-peaked (Akkermans & Muraki, 1997) and the uncertainty of the ability estimate at the beginning of a CAT can be quite considerable (Chang & Ying, 1996), this criterion seems more appropriate than MPI.

If an information function has more than one peak and the ability estimate is not stable at the beginning of a CAT administration, a non-optimal item could be selected by MPI. See Figure 2 for a visualization of this situation where the information functions of two items are plotted. This figure shows that if the ability estimate is zero and not very certain, item *j* is a better choice than item *i*. MII with a certain δ would select item *j* and MPI would select item *i*. It is expected that this situation occurs frequently enough to influence the quality of CAT in terms of the accuracy and precision of the ability estimate and that MII results in better ability estimates than MPI.

Figure 2

Case where MII selects a different item than MPI.



The value of δ should not be too large because the area under the information function of GPCM items equals $m_i a_i$ if $\delta \rightarrow \infty$ (see Appendix A), which means item selection becomes dependent only on the number of categories and the slope parameter a_i. As a consequence, the test is no longer adaptive because the ability of the subject has no influence on the selection of the items.

In this study, the value of δ is fixed. However, it can be used as a parameter that gets smaller as more items are administered in a CAT. In the beginning of a CAT, when the uncertainty of the ability estimate is relatively high, the item information of surrounding abilities is taken into account with a wider interval. When this uncertainty becomes smaller, the focus can be increasingly switched to the ability estimate alone by narrowing the interval. Since the influence of the width on a CAT can be determined better with a fixed interval, the interval was not narrowed here.

The quality of a CAT in terms of accuracy and precision of the ability estimate, given an infinite item bank, can be considered as a function of the interval width used in MII. It was expected that MII with δ equal to 0.5 or 1.0 would perform better than MPI when the ability distribution is standard normal because MII uses more information than MPI. If the interval becomes too large, however, the quality of the CAT will decline. Therefore, it was expected that MII with δ equal to 0.5 results in better CATs than MII with δ equal to 1.0.

Other factors in the administration of a CAT were included in this study in order to investigate possible interactions with one of the item selection procedures. The number of items in the item bank, the number of categories of the items in the item bank, and the distribution of the item category parameters (b_{ik}'s) were investigated.

6 Simulation Study

Design

CATs were simulated to compare maximum point information (MPI) with maximum interval information (MII) using polytomous items in a realistic setting. Random item selection was used as a benchmark. Only simulated item banks were used and the items were generated from the GPCM.

The independent variables were number of items in the item bank (2 levels), item category parameter distribution (3 levels), number of categories of items (3 levels), item selection procedure (4 levels), and stopping rules in combination with ability distribution (2 levels, see below). These variables generated a design with 90 cells, which is shown in Table 1. Random item selection was omitted when a maximum standard error was used as stopping rule to prevent exhaustion of the item banks. Each simulated CAT was evaluated on bias and root mean square of the ability estimate.

Table 1

Number of CAT simulations in the design of the simulation study.

Number of CAT simulations	Item selection procedure				
	MPI	MII-1.0	MII-0.5	Random	*Total*
Stopping rule – Fixed test length; Ability distr. – N(0, 1)					
- Number of items in item bank (150 and 500)					
- Number of categories (3, 4 and a mix of 2, 3, and 4)					
- Item category parameter distribution					
(b_{ik} – N(-1, 1), N(0, 1) and N(1,1))	18	18	18	18	72
Stopping rule – Max. std. err. of 0.20; Ability values = {-3, -2, -1, 0, 1, 2, 3}					
- Number of items in item bank (150 and 500)					
- Number of categories (3, 4 and a mix of 2, 3, and 4)					
- Item category parameter distribution					
(b_{ik} – N(0, 1))	6	6	6	0	18
Total	24	24	24	18	90

Item selection

Three experimental item selection procedures were used and random item selection was used as a benchmark. The experimental item selection procedures were maximum point information (MPI), maximum interval information with a fixed interval width of 1.0 (MII-1.0), and maximum interval information with a fixed interval width 0.5 (MII-0.5). The widths of 1.0 and 0.5 were chosen on the basis of arguments given in the section on item selection. The integrals were approximated using Simpson's rule with 100 sample points.

Item banks and item parameter distributions

The simulated item banks consisted of 150 and 500 items with parameters pseudo randomly drawn from normal distributions with $\log a_i \sim N(0, 0.5)$ and $b_i \sim N(-1, 1)$, $N(0, 1)$ and $N(1, 1)$, as can be seen in Table 1. Item category parameters were randomly drawn from their distributions after which they were ordered and assigned to the respective categories. Items with ordered item category parameters were used only because the emphasis in this study was on items with multi-modal IIFs. Different numbers of item categories were used in the item banks: 3, 4 and a mixture of 2, 3, and 4. The latter is called a mixed item bank.

Simulees and stopping rules

Two stopping rules were used: fixed test length of 30 items and maximum standard error of the ability estimate of 0.20 with an absolute maximum test length of 99 items. Each stopping rule was used in combination with a different ability distribution. Fixed test length was used in combination with a standard normal distribution of ability, and maximum standard error of 0.20 was used in combination with seven discrete ability values: {-3.0, -2.0, -1.0, 0.0, 1.0, 2.0, 3.0}. A random sample of $N = 5000$ simulees was drawn from the standard normal ability distribution. In addition, 500 simulees were selected for each of the seven discrete ability values resulting in a total of $N = 3500$ simulees. Ability estimates were obtained using the weighted likelihood estimation (WLE) procedure. WLE provides better (ability) estimates than either maximum likelihood estimation (MLE) or Bayesian modal estimation (BME) because it removes the first-order bias term from MLE (Warm, 1989).

CAT simulation procedure

The CAT started with two pseudo randomly selected items from the item bank, after which one of the four item selection procedures was started. The simulee's response was generated as follows: after an item was selected for administration, the probability of every score given the simulee's ability was calculated. Next, the cumulative probabilities of the scores were determined. A pseudo random number between 0 and 1 was then drawn from a uniform distribution. Finally, it was determined in which interval of the cumulative probabilities of the scores this number was. The score assigned to the simulee was the score belonging to this interval. The CAT was terminated after the used stopping rule was reached.

Evaluation criteria

Accuracy, i.e., mean bias (Bias), and precision, i.e., root mean square (RMS) of the ability estimates, were used to evaluate the CATs. Bias and RMS are defined as

$$Bias = \frac{1}{N}\sum_{j=1}^{N}(\theta_j - \hat{\theta}_j), \tag{6.1}$$

$$RMS = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\theta_j - \hat{\theta}_j)^2}. \tag{6.2}$$

Bias is the mean over all simulees of the difference between the ability and the ability estimate. RMS is the root of the mean squared differences between the ability and the ability estimate. Mean test length was only used to evaluate the CATs when the stopping rule was a maximum standard error of 0.20. Item exposure was investigated only when mixed item banks were used so that the exposure of items with different categories could be compared. Item exposure is defined as the number of times an item was administered divided by the total number of test administrations (note that this definition only applies to CATs with a fixed length).

7 Results

Although the results did not show any substantial differences in bias and RMS of the ability estimates between MPI, MII-1.0, and MII-0.5, there were some differences, however. In most cases, MPI performed equal to or better than MII, and only in some cases MII performed slightly better. In general, the differences between the three experimental item selection procedures were small compared to the differences between those three and random item selection.

The other independent variables (size of item bank, number of item categories, item parameter distribution) did not interact with any of the three experimental item selection procedures in terms of differences in bias and RMS. Because of the absence of interactions, not all the results will be presented. The main effects will be illustrated with examples.

In Figure 3 and 4, bias and RMS are plotted against the number of administered items for all four item selection procedures. In the CATs, item banks containing 500 trinary items with standard normally distributed item category parameters were used. It can be seen that MII-0.5 showed somewhat less RMS than MPI and MII-1.0.

Figure 3

Bias of CATs using 500-item bank with trinary items and $b_i \sim N(0, 1)$, $N = 5000$.

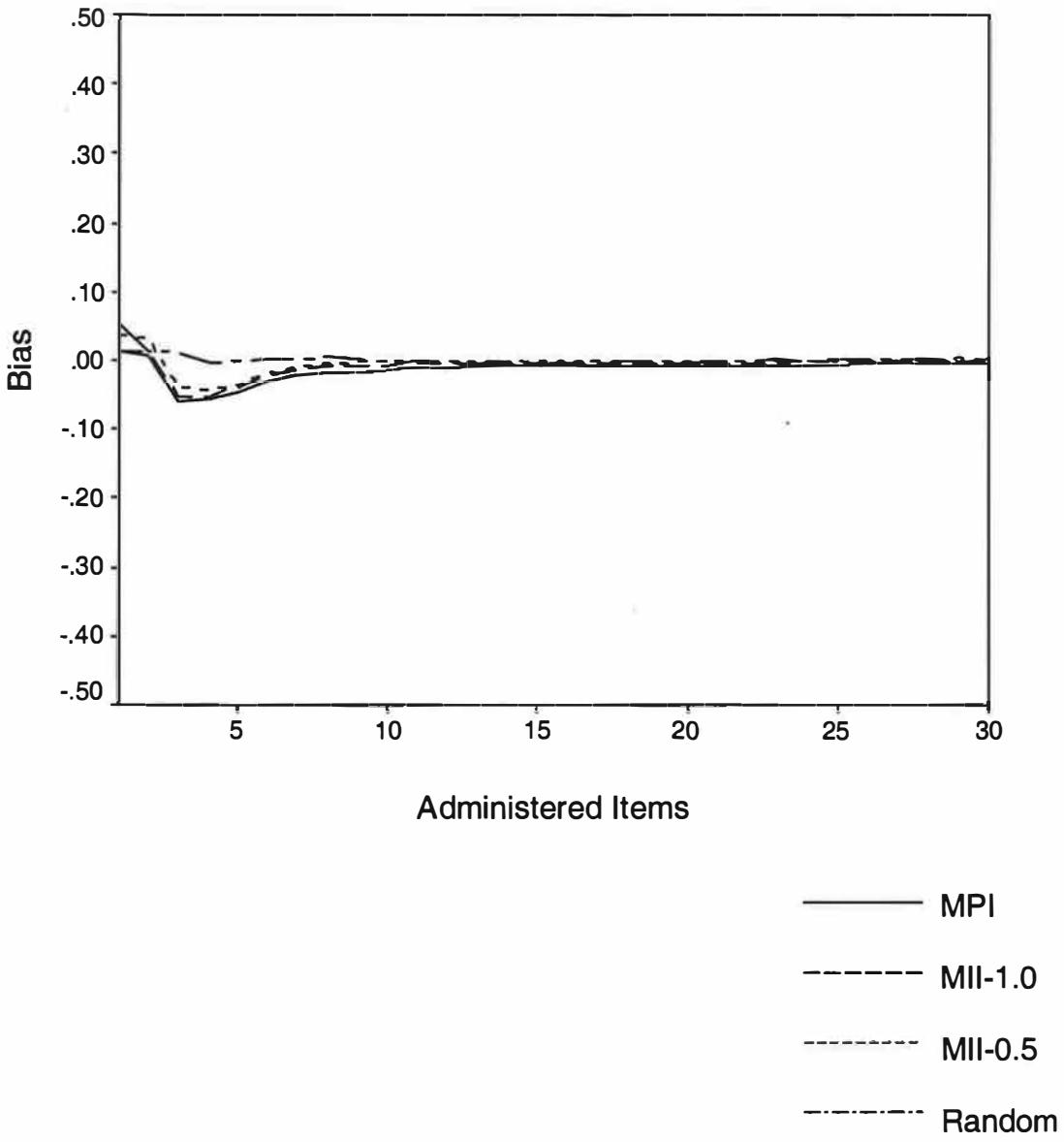
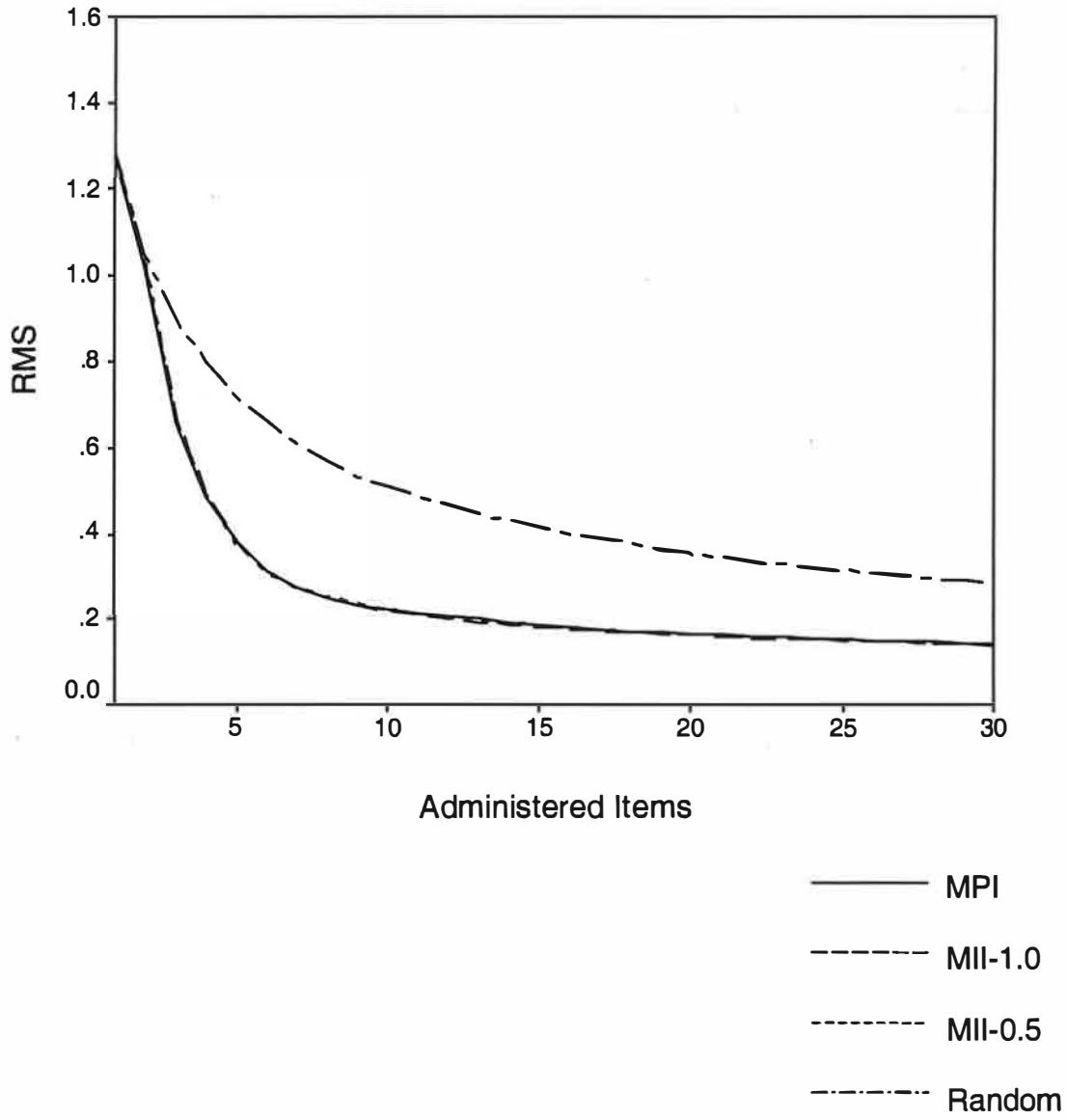


Figure 4

RMS of CATs using 500-item bank with trinary items and $b_i \sim N(0, 1)$, $N = 5000$.



The effects of the number of items in an item bank are presented in Figure 5 and 6. Bias and RMS are plotted against the administered items for the item banks with 500 and 150 items. The CATs used MII-0.5, trinary items and standard normally distributed item category parameters. CATs selecting items from the item bank with 500 items showed less bias in the beginning and less RMS at every stage.

Figure 5

Bias of CATs using MII-0.5 with trinary items and $b_i \sim N(0, 1)$, $N = 5000$.

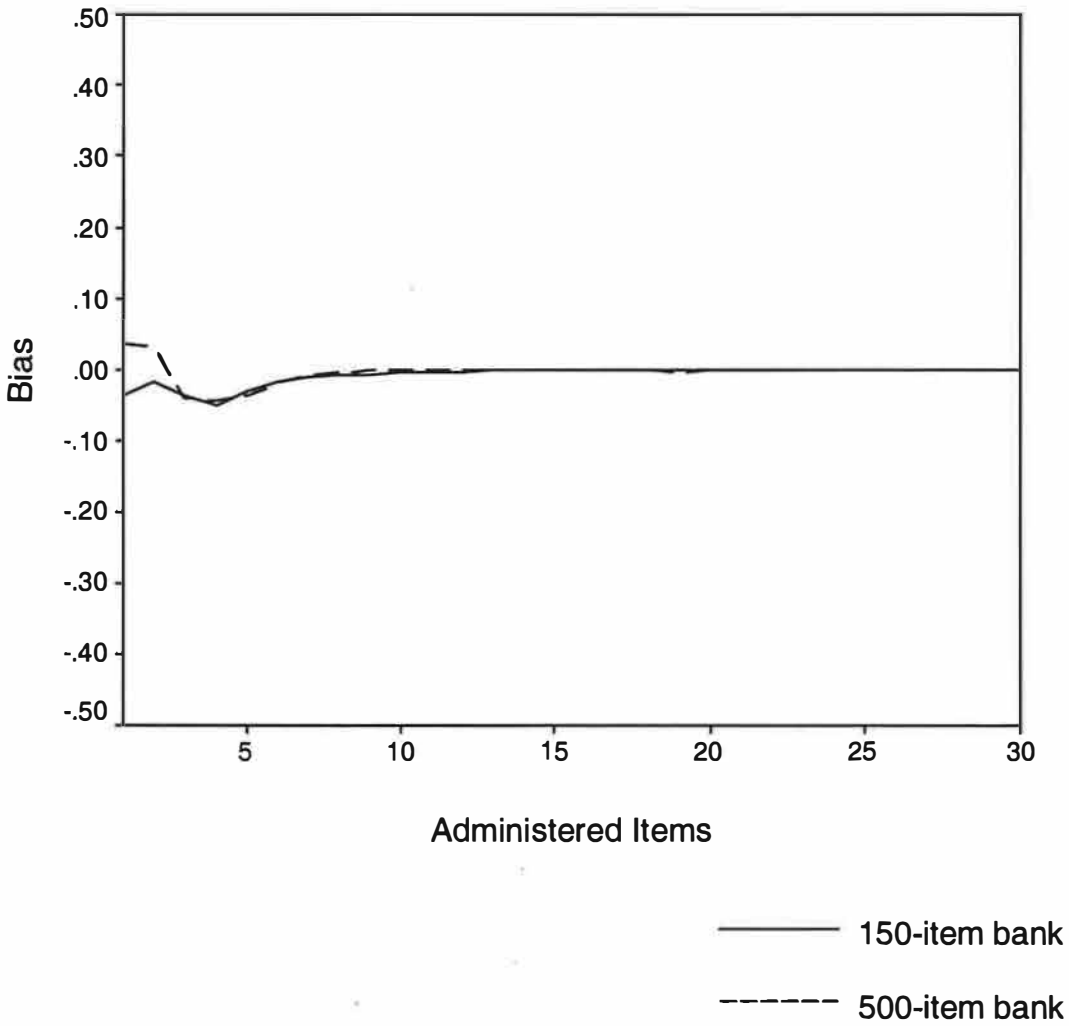


Figure 6

RMS of CATs using MII-0.5 with trinary items and $b_i \sim N(0, 1)$, $N = 5000$.

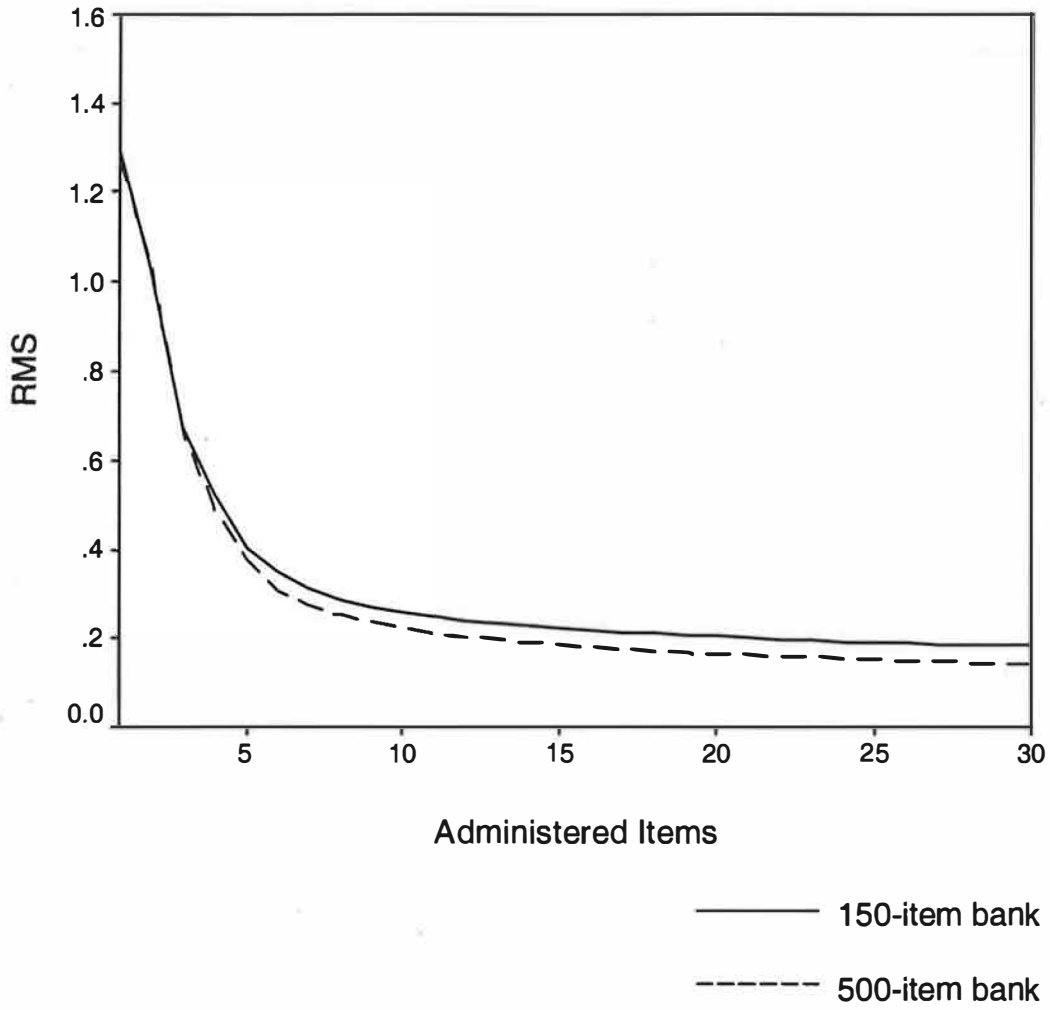


Figure 7 and 8 show the bias and RMS of CATs with the three item category parameter distributions using MII-0.5. The CATs used the 500-item banks with trinary items. If the item category parameter distribution matched the ability distribution, the CATs showed no bias. The two mismatched item category parameter distributions did show considerable bias in the beginning of a CAT. This bias, however, disappeared when the number of administered items was increased. When the distributions of ability and the item category parameters matched, the CATs showed less RMS.

Figure 7

Bias of CATs using MII-0.5, 500-item bank, and trinary items, N =5000.

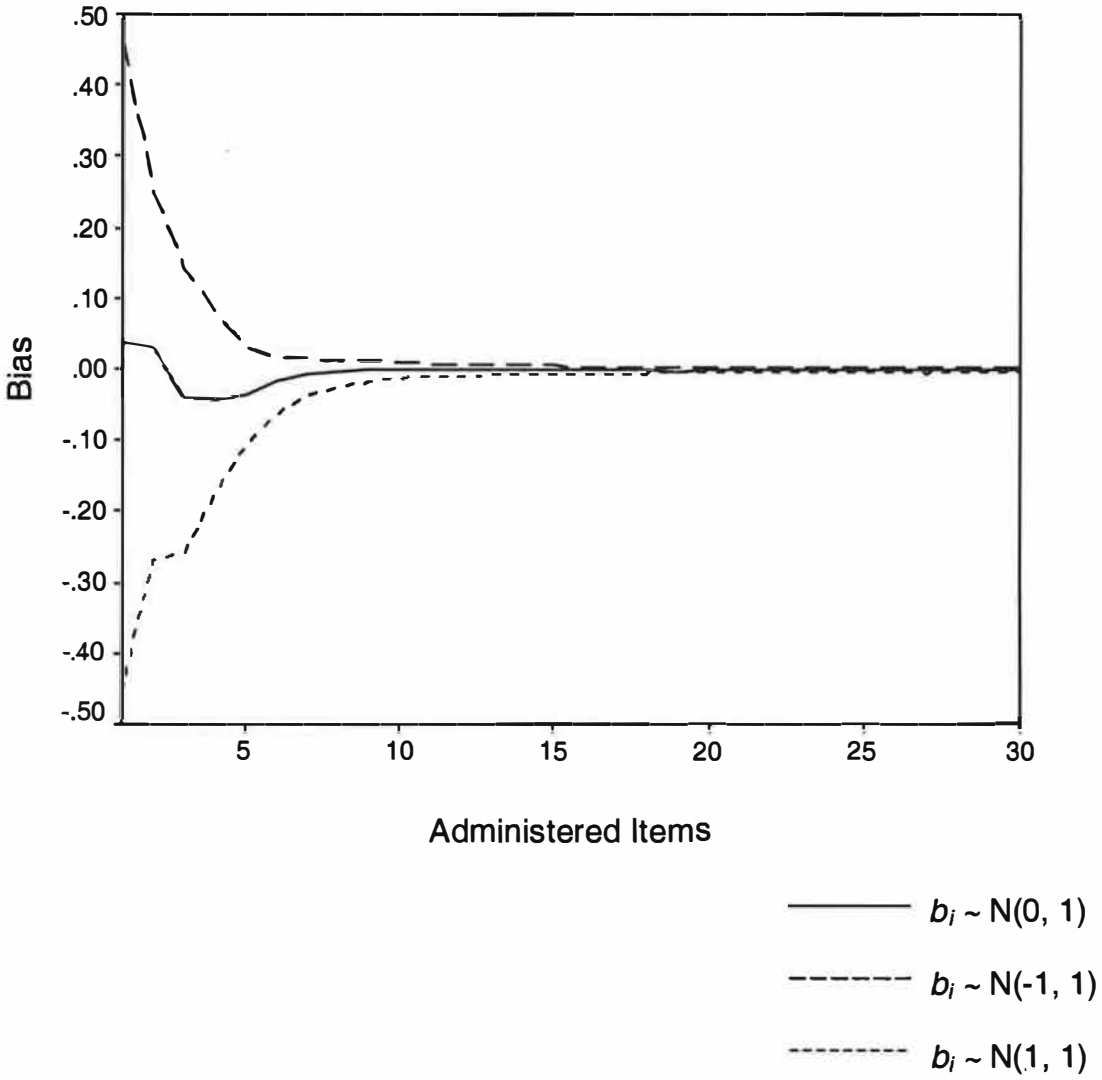
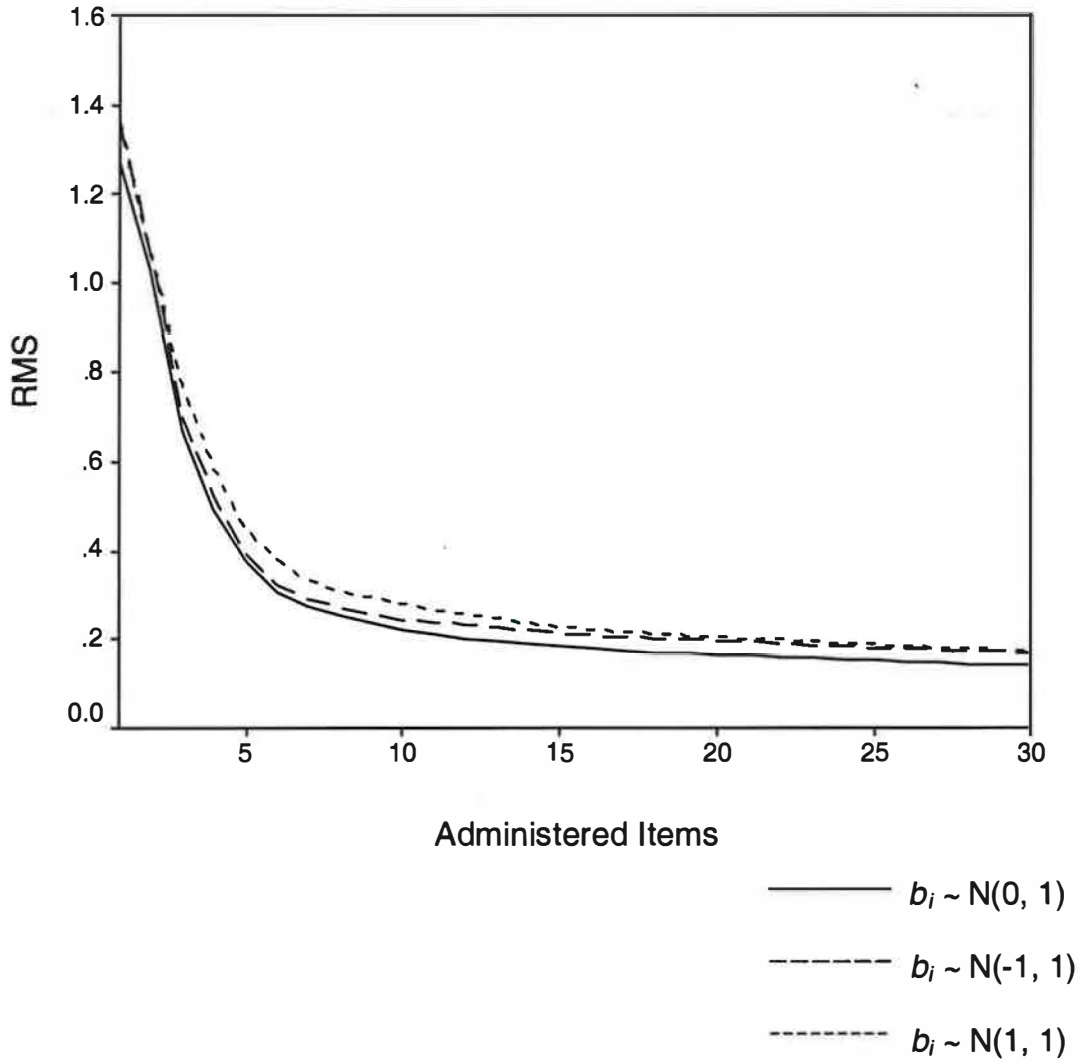


Figure 8

Bias of CATs using MII-0.5, 500-item bank, and trinary items, N=5000.



CATs using item banks with dichotomous items showed more bias in the beginning than CATs using item banks with items with three, four categories, and mixed item banks, as can be seen in Figure 9. (Note that dichotomous items were only studied in this specific case and were originally not included in the design of the study.) All these CATs used MII-0.5 and standard normally distributed item category parameters. The CATs using items with four categories showed the least RMS. This is shown in Figure 10. The CATS using item banks with mixed categories performed slightly better than the item bank with trinary items in the beginning of the CATs, but worse than the item bank with items with four categories.

Figure 9

Bias of CATs using MII-0.5 with $b_i \sim N(0, 1)$, $N = 5000$.

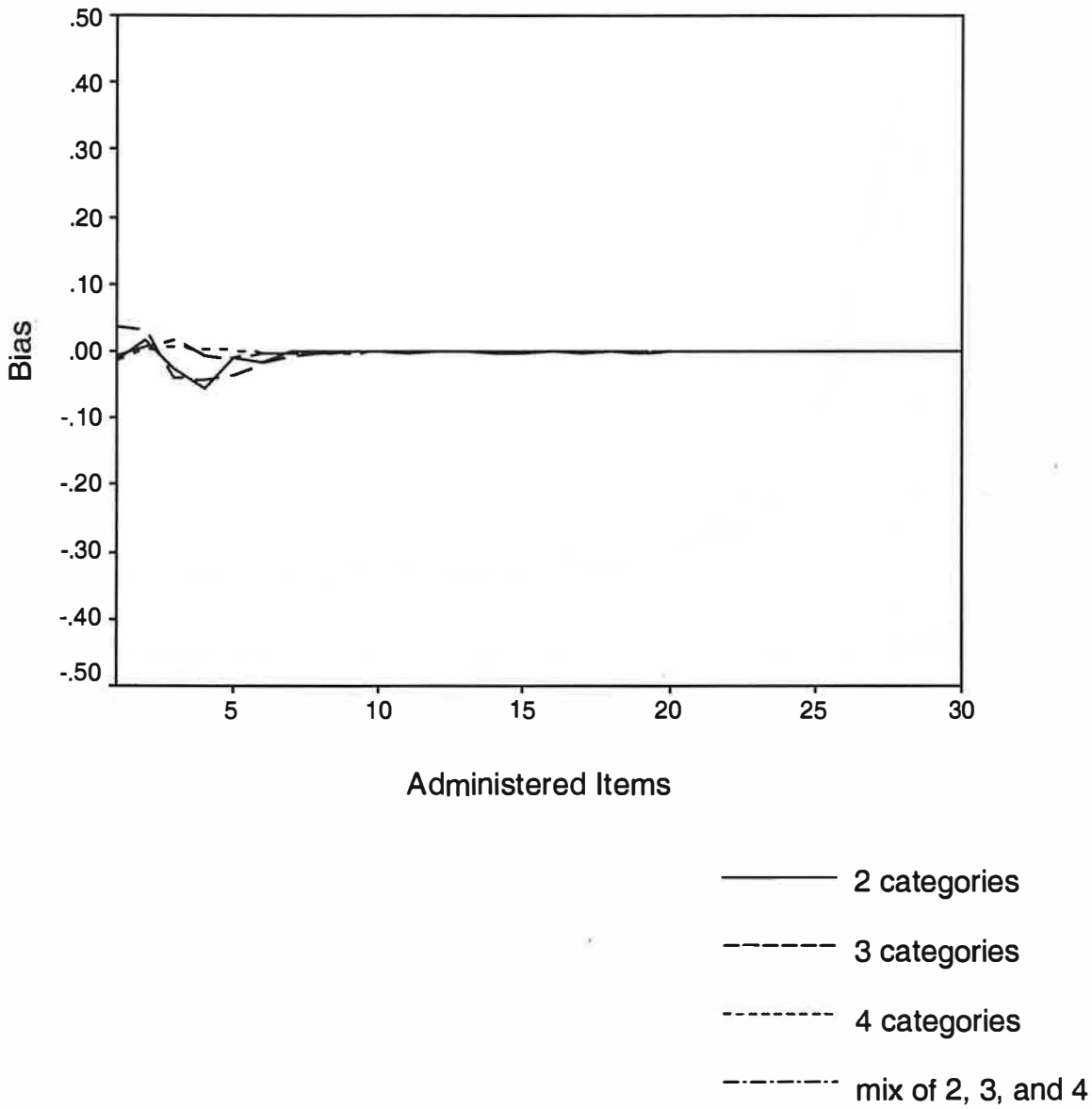
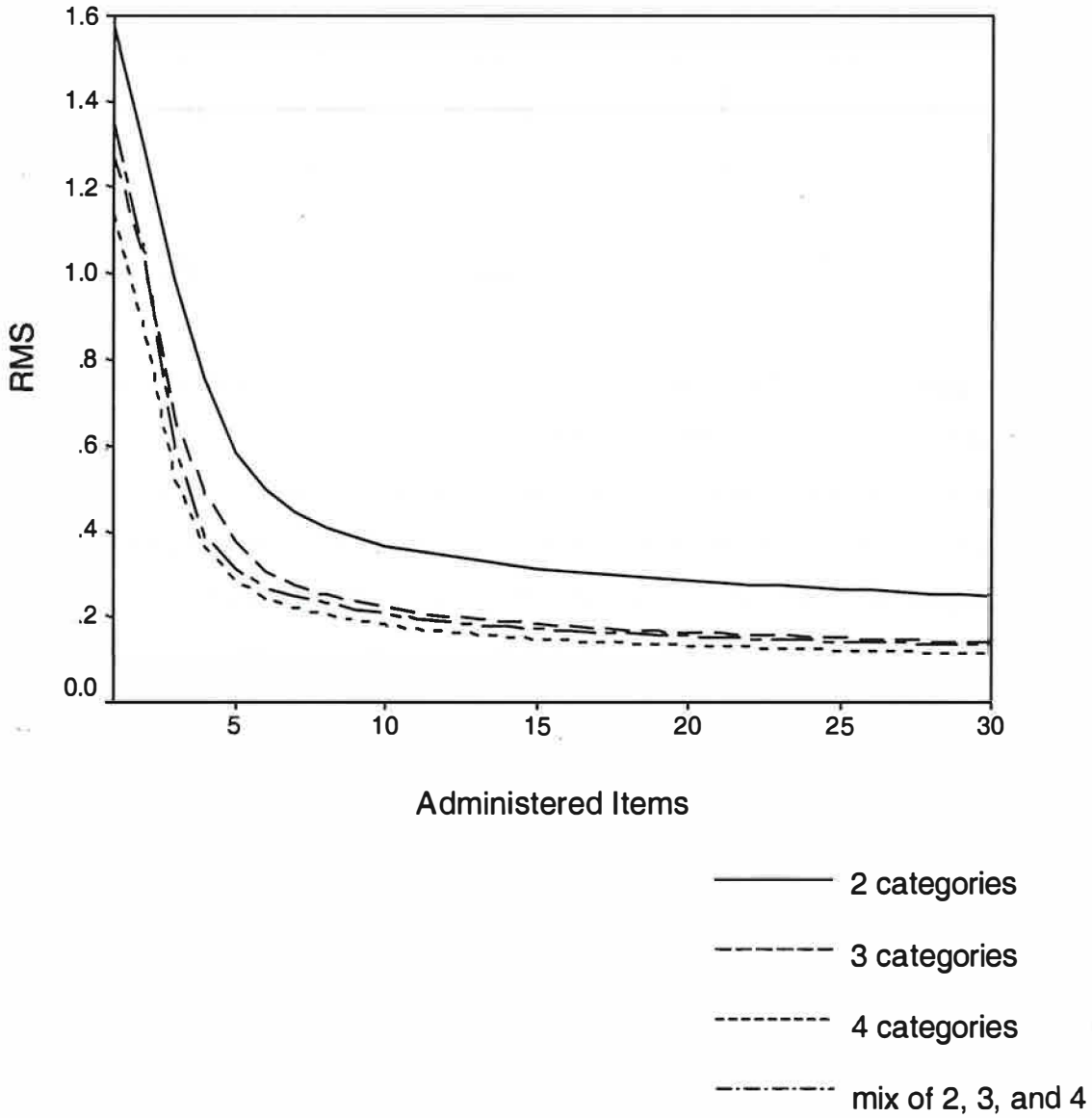


Figure 10

RMS of CATs using MII-0.5 with $b_i \sim N(0, 1)$, $N = 5000$.



When mixed item banks were used, the items with more categories were selected more often in terms of item exposure, with $F_{contrast}(2, 497) = 24.53$, $p = .000$ (see Table 2). Note that this 500-item bank was used for CATs with a fixed length of 30 (explaining the total mean exposure of 0.06). More categories did not produce equally better CATs when considering RMS (the total score of three trinary items is equal to the total score of two items with four categories). An improvement of 3/2 was definitely not the case. In general, item banks with trinary items showed the best results when taking into account the total maximum score.

Table 2

Means and standard deviations of item exposures in a CAT using an item bank with mixed categories.

Number of Categories	*Mean*	*Std. Dev.*	*N*
2	0.0176	0.0570	167
3	0.0396	0.1108	166
4	0.1228	0.2174	167
Total	0.0600	0.1514	500

In Figure 11, 12, and 13, bias, mean test length, and RMS are plotted against ability for CATs using MPI, MII-0.5, and MII-1.0 with trinary items and standard normally distributed item category parameters. The stopping rule was a maximum standard error of 0.20 with an absolute maximum test length of 99 items. Bias and RMS for all three item selection procedures are approximately equal for all ability values after the stopping rule was reached. Test lengths differed substantially, however. Simulees with large positive and negative ability values had to be administered considerably longer tests.

Figure 11

Bias of CATs with maximum standard error of 0.20 as a stopping rule with absolute maximum length of 99 items, $N = 3500$.

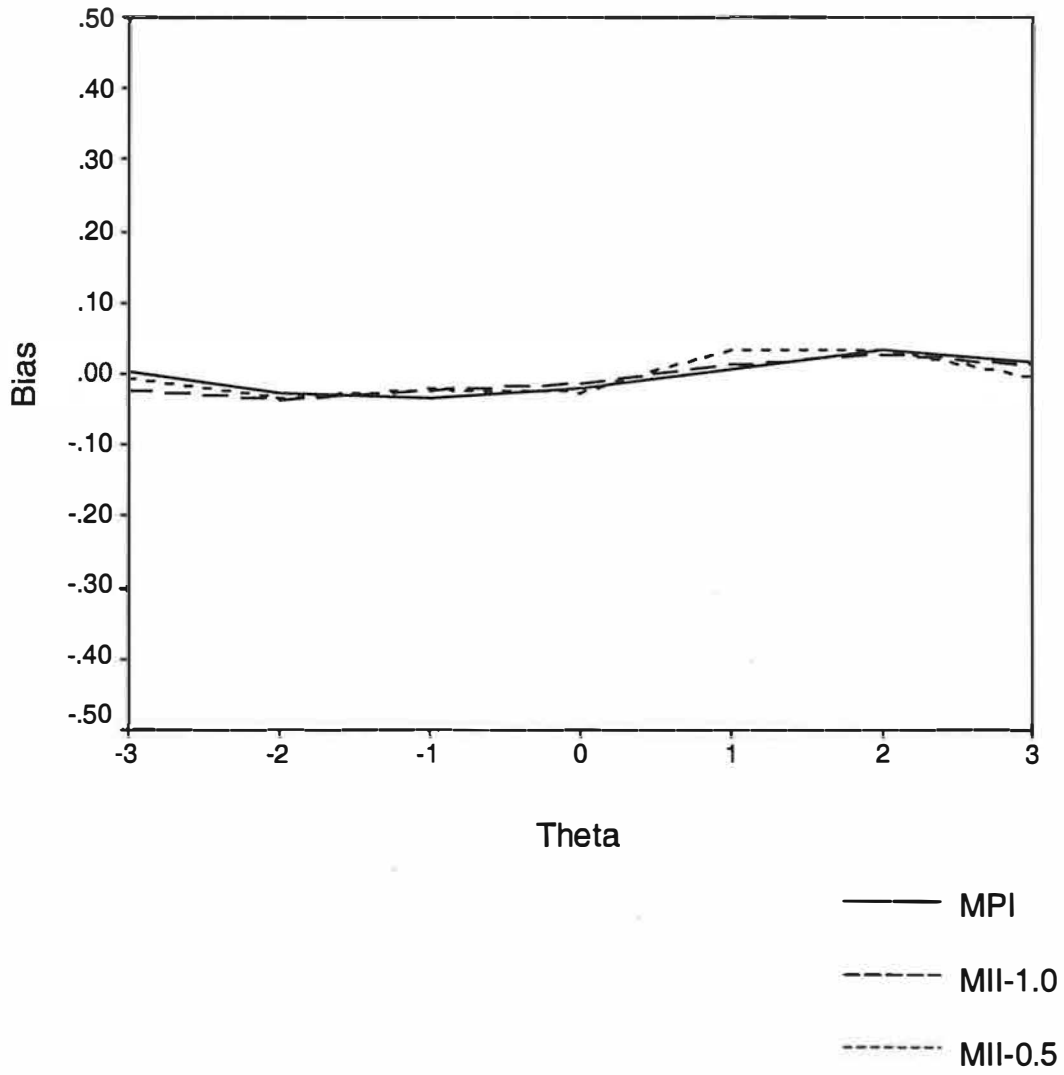


Figure 12

RMS of CATs with maximum standard error of 0.20 as a stopping rule with absolute maximum length of 99 items, $N = 3500$.

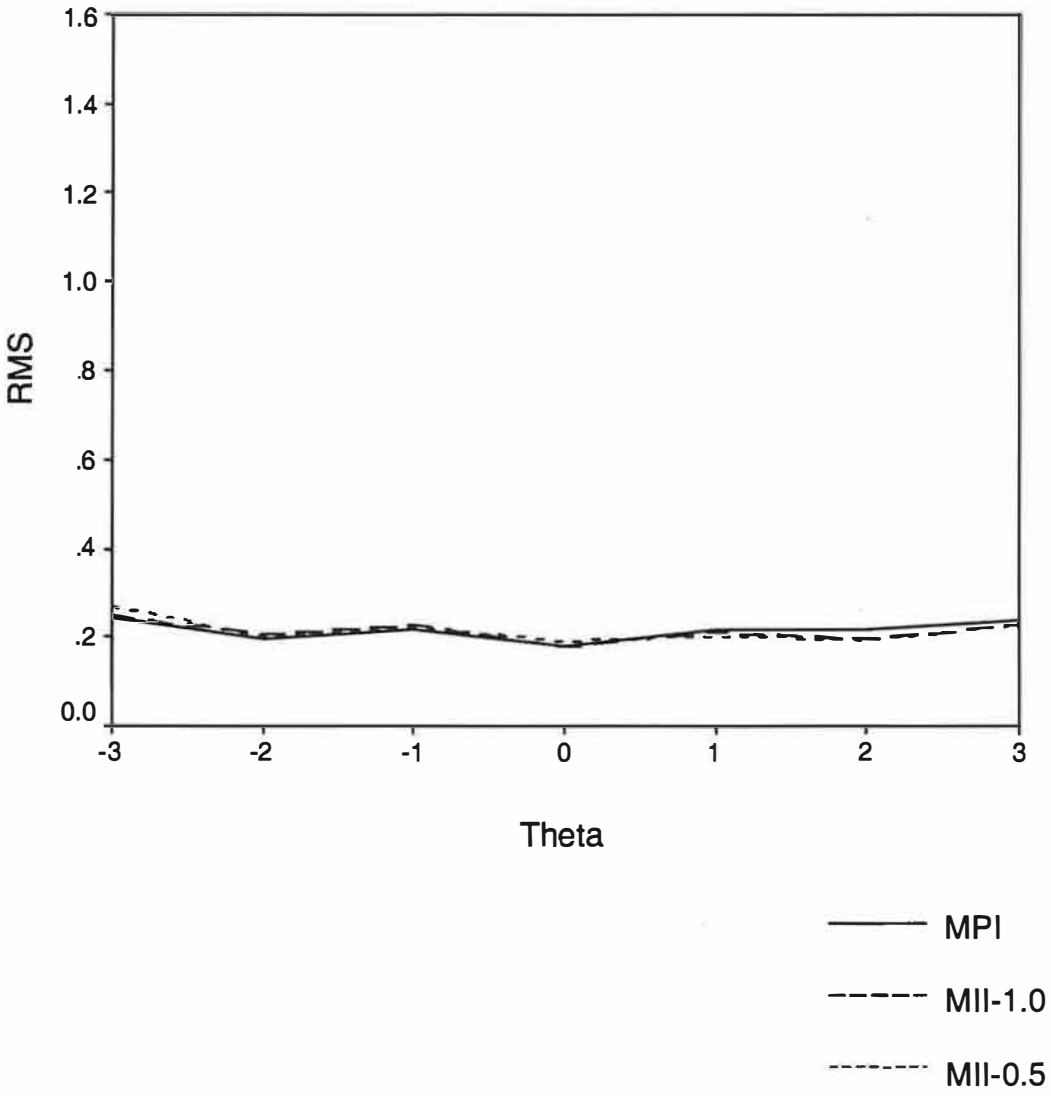
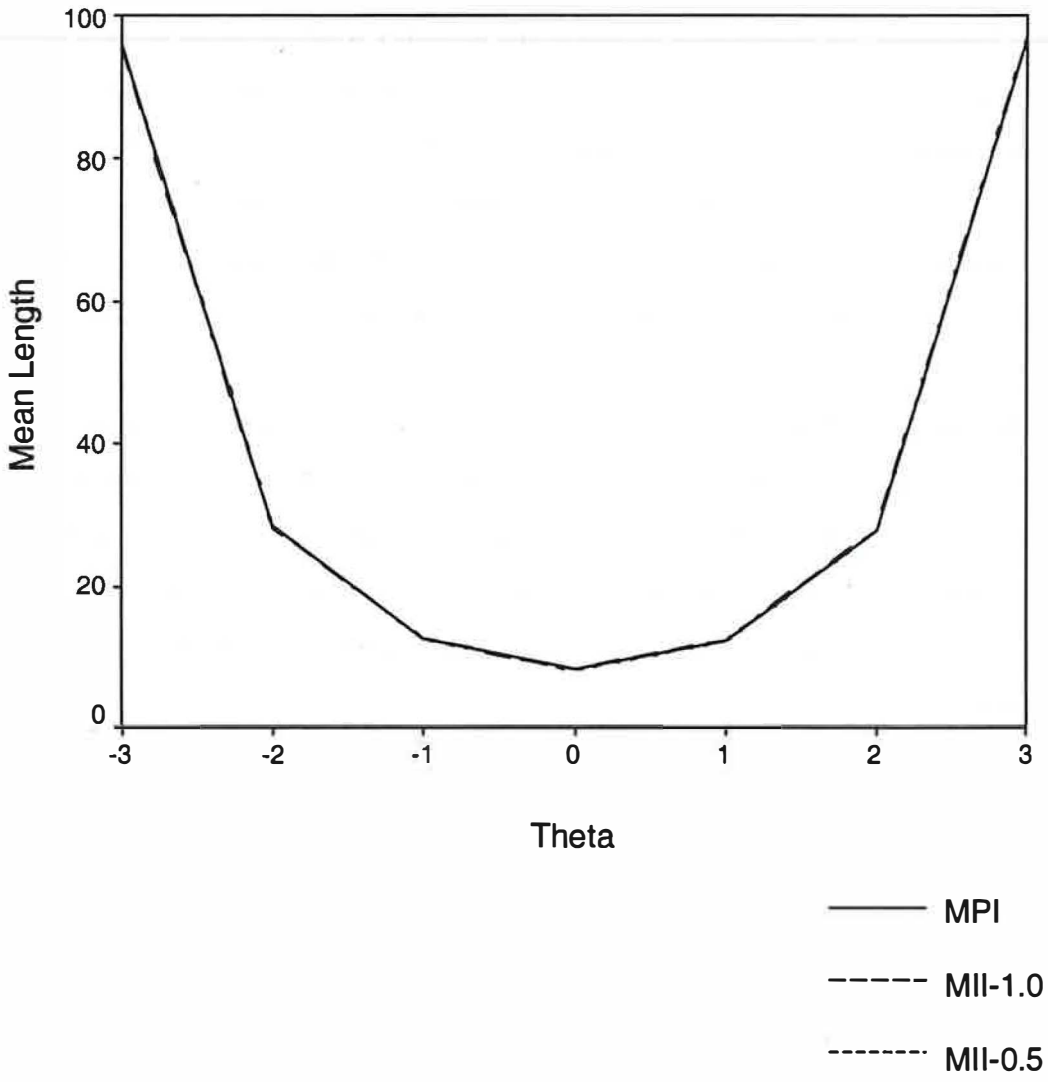


Figure 13

Mean length of CATs with maximum standard error of 0.20 as a stopping rule with absolute maximum length of 99 items, $N = 3500$.



8 Conclusions and Discussion

The main conclusion of this study is that the two maximum interval information (MII) item selection procedures, in general, did not improve the quality of a CAT in terms of accuracy and precision of the ability estimate (in comparison to the maximum point information (MPI) selection procedure). While, in some cases, MII performed slightly worse and, in others, slightly better, the

differences were small. This result is in accordance with the study by Berger and Veerkamp (1997) in which dichotomous items were used. The reason for these small differences is that MII, like MPI, depends very much on the a_i-parameter. MII selects roughly the same items as MPI, particularly the ones with a high a_i parameter, and many item categories when mixed item banks are used. Therefore, the differences between MII and MPI were too small to have an effect on the quality of a CAT in terms of accuracy and precision of the ability estimate. Apparently, the situation in which MII selects a different item than MPI does, as was shown in Figure 2, does not occur as often as is needed to result in CATs of different quality. The interval width to calculate the area under the information function did not influence the quality of the CATs. Despite some differences between MII-1.0 and MII-0.5, there was no indication that either one should be preferred.

The idea of MII as an average of ability values is more attractive than the idea of MPI, especially when the ability estimate is uncertain. When these two item selection procedures perform similarly, MPI may be preferred because it is easier to compute. Note that other weight functions (e.g., the likelihood function of the ability estimate or a function that cancels out the dependency on a_i) can be used in the general weighted information criterion, which may result in better CATs.

The item selection procedures that were investigated in this study are highly dependent on the a_i parameter. In computerized adaptive testing, this dependence is not desirable (see Section 5), although discrimination between different abilities is. In the beginning of a CAT administration, when the ability estimate is relatively uncertain, items that give information over a range of ability values are desirable. To overcome the dependency on a_i, Kullback-Leibler information which is less dependent on a_i may be used for polytomous item selection. Chang and Ying (1996) found that, for dichotomous items, item selection using Kullback-Leibler information resulted in CATs (evaluated on several criteria) at least equal to CATs using MPI and in several cases better than CATs using MPI. With computers getting better and faster, Bayesian item selection, which requires a substantial amount of computing time when using dichotomous items (Van der Linden, 1998), could be realized for polytomous items. Bayesian item selection is also less dependent on the a_i parameter.

Although CATs using item banks with items with more categories resulted in better CATs, when taking into account the maximum score of an item, the results indicated that the optimal

number of categories of an item is three. This seems plausible given the decomposition property of polytomous GPCM items and the consequences of decomposition for item information discussed in Section 4. However, with the two stopping rules used here, items with four categories are selected more often than items with two and three categories (when mixed item banks are used). A different stopping rule may be used to overcome this problem, namely, that of a fixed maximum total score. This stopping rule may result in CATs of different lengths but with the same maximum total score. The use of this stopping rule may have effects on item selection procedures when mixed item banks are used because one has to select an optimal item given the current ability estimate and the number of scoring points left.

The administration of a CAT is subject not only to criteria such as maximum standard error of the ability estimate and maximum test length. Other criteria such as item exposure (Stocking & Lewis, 1998; Sympson & Hetter, 1985), item content (Kingsbury & Zara, 1991), and test specifications should also be considered in order to obtain reliable and valid tests (Chang & Ying, 1996). In general, these additional criteria have an adverse effect on the properties of a CAT such as accuracy and precision of the ability estimate (Van der Linden, 1998). From a practical point of view, however, a study of the effects of these criteria on polytomous item selection and on the resulting CATs is of interest.

9 References

Akkermans, W., & Muraki, E. (1997). Item information and discrimination functions for trinary PCM items. *Psychometrika, 62*, 569-578.

Baek, S-G. (1997). Computerized adaptive testing using the partial credit model for attitude measurement. In: M. Wilson, G. Engelhard jr., & K. Draney (ed.), *Objective measurement: Theory into practice, vol. 4*. Norwood, NJ: Ablex.

Berger, M.P.F., & Veerkamp, W.J.J. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-26.

Birnbaum, A. (1968). Some latent trait models. In: F.M. Lord, & M.R. Novick, *Statistical theories of mental test scores* (p. 395-479). Reading, MA: Addison-Wesley.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

De Ayala, R.J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development, 25*, 172-189.

Dodd, B.G., De Ayala, R.J., & Koch, W.R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5-22.

Donoghue, J.R. (1994). An empirical examination of the IRT information function of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 31*, 295-311.

Eggen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-261.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhof.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331-347.

Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometrika, 59*, 111-119.

Huynh, H. (1996). Decomposition of a Rasch partial credit item into independent binary and indecomposable trinary items. *Psychometrika, 61*, 31-39.

Kingsbury, G.G., & Zara, A.R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*, 241-261.

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

Mellenbergh, G.J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*, 91-100.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*, 351-363.

Priestley, H.A. (1997). *Introduction to Integration.* Clarendon Press, Oxford.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded responses. *Psychometrika Monograph,* No. 17.

Stocking, M.L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Sympson, J.B., & Hetter, R.D. (1985). Controlling item exposure rates in computerized adaptive testing. Proceedings of the 27th annual meeting of the Military Testing Association (973-977). San Diego, CA.

Tang, L.T. (1996, April). *A comparison of the traditional maximum information method and the global information method in CAT item selection.* Paper presented at the annual meeting of the NCME, New York.

van der Linden, W.J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216.

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

A. Appendix

For the item selection criteria used in this paper, it is useful to find out what the total area under the Fisher information function of a polytomous item is. It will be shown that it is equal to $m_i a_i$.[1] This only has to be shown for trinary GPCM items, because of the decomposition property discussed in section 4. The area under the information function of 2-PL items is equal to a_i (Birnbaum, 1968, p. 460). With this knowledge, the area under the information function of any GPCM item can be calculated.

First, we can write the information function of a trinary GPCM (see Equation 3.2) item as follows:

$$I_i(\theta) = a_i\left(P_{i1}(\theta) + 4P_{i2}(\theta) - \left((P_{i1}(\theta)) + 2(P_{i2}(\theta))\right)^2\right) \tag{A.1}$$

Next, the indefinite integral is calculated for each part. For the first term, $P_{il}(\theta)$, the integral becomes

$$\int P_{i1}(\theta)d\theta = \int \frac{\exp(a_i(\theta - b_{i1}))}{1 + \exp(a_i(\theta - b_{i1})) + \exp(a_i(2\theta - b_{i1} - b_{i2}))} d\theta, \tag{A.2}$$

which can be rewritten as

$$\int \frac{2\exp(a_i b_{i2})}{a_i\left(\left(2\exp(a_i\theta) + \exp(a_i b_{i2})\right)^2 + \left(4\exp(a_i(b_{i1} + b_{i2})) - \exp(2a_i b_{i2})\right)\right)} d\left(2\exp(a_i\theta) + \exp(a_i b_{i2})\right). \tag{A.3}$$

By applying the following substitutions:

$x = 2\exp(a_i\theta) + \exp(a_i b_{i2})$, and
$u = \sqrt{4\exp(a_i(b_{i1} + b_{i2})) - \exp(2a_i b_{i2})}$,

[1] A great deal of this appendix was calculated with the help of the Maple V computer program.

the integral becomes

This is a standard integral (except for the constant factor) and its solution is

$$\frac{2\exp(a_i b_{i2})\arctan\left(\frac{x}{u}\right)}{a_i u}.$$

$$\int\left(\frac{2\exp(a_i b_{i2})}{a_i}\right)\left(\frac{1}{x^2+u^2}\right)dx. \tag{A.4}$$

$$\int P_{i1}(\theta)d\theta = \frac{2\exp(a_i b_{i2})\arctan\left(\frac{2\exp(a_i\theta)+\exp(a_i b_{i2})}{\sqrt{4\exp(a_i(b_{i1}+b_{i2})-\exp(2a_i b_{i2})}}\right)}{a_i\sqrt{4\exp(a_i(b_{i1}+b_{i2})-\exp(2a_i b_{i2})}}+C, \tag{A.5}$$

By filling in the substitutions, the integral and its solution are given by

where C is a constant.

The second part of the integral of the information function can be written as follows:

This results in

$$\int P_{i2}(\theta)d\theta = -0.5\int P_{i1}(\theta)d\theta - \int\frac{0.5\exp(a_i(\theta-b_{i1}))-\exp(a_i(2\theta-b_{i1}-b_{i2}))}{1+\exp(a_i(\theta-b_{i1}))+\exp(a_i(2\theta-b_{i1}-b_{i2}))}d\theta. \tag{A.6}$$

The other parts can be integrated in the same way. The definite integral of the information function

$$\int P_{i2}(\theta)d\theta = -\frac{\exp(a_i b_{i2})\arctan\left(\frac{\exp(a_i b_{i2})+2\exp(a_i\theta)}{\sqrt{4\exp(a_i(b_{i1}+b_{i2})-\exp(2a_i b_{i2})}}\right)}{a_i\sqrt{4\exp(a_i(b_{i1}+b_{i2})-\exp(2a_i b_{i2})}} \tag{A.7}$$

$$\int I_i(\theta)d\theta = +\frac{\ln(\exp(a_i(b_{i1}+b_{i2})+\exp(a_i(\theta+b_{i2})+\exp(2a_i\theta))}{2a_i}+C.$$

$$\frac{a_i\left(-8\exp(2a_i(b_{i1}+b_{i2}))-4\exp(a_i(\theta+b_{i1}+2b_{i2}))+2\exp(a_i(b_{i1}+3b_{i2}))+\exp(a_i(\theta+3b_{i2}))\right)}{\left(4\exp(a_i(b_{i1}+b_{i2}))-\exp(2a_i b_{i2})\right)\left(\exp(a_i(\theta+b_{i2}))+\exp(a_i(b_{i1}+b_{i2}))+\exp(2a_i\theta)\right)}+C \tag{A.8}$$

of a trinary GPCM item is (after simplifying)

In order to calculate the area under the total IIF, the extended fundamental theorem of the calculus (FTC) is needed (see Priestley, 1997, Section 16.3), which amounts to

(A.9)

$$\int_{-\infty}^{\infty} F' = \lim_{b \to \infty} F(b) - \lim_{a \to -\infty} F(a).$$

The total area under the information function of a GPCM item is equal to $2a_i$ because if Equation A.8 is evaluated in $a \rightarrow -\infty$, the integral goes to $-2a_i$ and $b \rightarrow \infty$, it goes to 0. Because of the decomposition property of the GPCM discussed earlier, and since the area under the IIF is equal to a_i for dichotomous items, the area under the IIF of any GPCM item is $m_i a_i$.