DOCUMENT RESUME

ED 450 131 TM 032 321

AUTHOR Glas, Cees A. W.; Vos, Hans J.
TITLE Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory. Research Report.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO RR-00-06
PUB DATE 2000-00-00
NOTE 34p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Bayesian Statistics; Classification; *Computer Assisted Testing; Item Response Theory; *Mastery Tests; *Test Construction; Test Items
IDENTIFIERS Multidimensionality (Tests); *Sequential Testing; Testlets

ABSTRACT

This paper focuses on a version of sequential mastery testing (i.e., classifying students as a master/nonmaster or continuing testing and administering another item or testlet) in which response behavior is modeled by a multidimensional item response theory (IRT) model. First, a general theoretical framework is outlined that is based on a combination of Bayesian sequential decision theory and multidimensional IRT. Then how multidimensional IRT-based sequential master testing can be generalized to adaptive item- and testlet-selection rules is discussed for the case where the choice of the next item or testlet to be administered is optimized using the information from previous responses. Both compensatory and conjunctive loss structures are considered. Simulation studies are used to evaluate: (1) the performance, in terms of average loss, of multidimensional IRT-based sequential mastery testing as a function of the number of items administered per testing stage; (2) the effects on average loss when turning the sequential procedure into an adaptive sequential procedure; and (3) the impact on average loss when the multidimensional structure is ignored and a unidimensional IRT model is used in the decision procedure. (Contains 9 tables and 20 references.) (Author/SLD)

TM

Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory

Research Report 00-06

Cees A.W. Glas
Hans J. Vos

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory

Cees A.W. Glas

Hans J. Vos

Abstract

In this article, a version of sequential mastery testing (i.e., classifying students as a master/non-master or to continue testing and administering another item or testlet) is studied where response behavior is modeled by a multidimensional item response theory (IRT) model. First, a general theoretical framework is outlined that is based on a combination of Bayesian sequential decision theory and multidimensional IRT. Then it is pointed out how multidimensional IRT-based sequential mastery testing can be generalized to adaptive item- and testlet-selection rules, that is, to the case where the choice of the next item or testlet to be administered is optimized using the information from previous responses. Both compensatory and conjunctive loss structures are considered. Simulation studies are used to evaluate (1) the performance, in terms of average loss, of multidimensional IRT-based sequential mastery testing as a function of the number of items administered per testing stage, (2) the effects on average loss when turning the sequential procedure into an adaptive sequential procedure, (3) the impact on average loss when the multidimensional structure is ignored and a unidimensional IRT model is used in the decision procedure.

Key words: adaptive testing, Bayesian sequential decision theory, mastery testing, item response theory, multidimensional item response theory.

Introduction

In an adaptive mastery test (AMT), the decision is to classify a student as a master, a non-master, or to continue testing and administering another item or testlet (i.e., items within a batch that are strongly related). In the sequel, we will assume that another testlet rather than another item is presented in case of continuing testing. Adaptive mastery tests are designed with the goal of maximizing the probability of making correct classification decisions (i.e., declaring mastery or non-mastery) while at the same time minimizing test length (Lewis & Sheehan, 1990). For instance, Lewis and Sheehan (1990) showed in a simulation study that average test lengths could be reduced by half without sacrificing classification accuracy. In AMT, both the stopping rule (i.e., termination criterion) and testlet selection mechanism are adaptive. In other words, test takers with a low and high level of ability are classified as non-master and master, respectively, whereas those with an intermediate level of ability are presented another testlet. Furthermore, student's ability measured on a latent continuum is estimated after each response, and the next testlet is selected such that its difficulty matches student's last ability estimate. Doing so, able students can avoid doing too many easy items and less able students can avoid being exposed to too many difficult items. An implicit assumption is that items have unequal difficulty implying that the probability to answer an item correctly is not equal for all items in the pool, that is, response behavior is modeled by an item response theory (IRT) model. In case the termination criterion is determined using Bayesian sequential decision theory (e.g., De Groot, 1970; Lehmann, 1986), Vos and Glas (2000) denote an AMT as an adaptive sequential mastery test (ASMT), which combines the strong points of both approaches.

Three basic elements can be identified in Bayesian sequential decision theory. In addition to a measurement model relating the probability of a correct response to student's (unknown) ability and a loss function evaluating the total costs and benefits for each possible combination of decision outcome and ability, cost of test administration ('cost per observation') must be explicitly specified in this approach. Doing so, maximum expected losses associated with the non-mastery and mastery decisions can now be calculated straightforward at each stage of testing. As far as the maximum expected loss associated with continuing testing concerns, this quantity is determined by averaging the maximum expected losses associated with each of the possible future decision outcomes with weights equal to the probability of observing those outcomes (i.e., the posterior predictive distributions). Optimal rules (i.e., Bayesian sequential rules) are now obtained by minimizing the posterior expected losses associated with all possible decision rules at each stage of testing using techniques of dynamic programming (i.e., backward

induction). Backward induction starts by considering the final stage of testing (where no option to continue testing is available) and then works backward to the first stage of testing. Decision rules are hereby prescriptions specifying for each possible response pattern what decision (i.e., declare master/non-mastery or to continue testing) has to be taken. The Bayes principle assumes that prior knowledge about student's ability is available and can be characterized by a probability distribution called the prior. This prior probability represents our best prior beliefs concerning student's ability, that is, before any testlet yet has been administered.

The impact of IRT-based sequential mastery testing (SMT), that is, the next item to be administered is randomly selected within the Bayesian sequential decision-theoretic framework, and ASMT on average loss, proportion correct classification decisions, and proportion testlets given was investigated by Vos and Glas (2000) in a number of simulation studies using the 1PL as well as the 3PL testlet model. Two different dependence structures of testlet responses were introduced for the 3PL testlet model. First, it was assumed that all item responses were independent, given student's ability. Secondly, a hierarchical IRT model was used reflecting a greater similarity of responses to items within than between testlets. For the loss structure involved, a linear loss function was adopted implying that the distance between student's ability and the cut- off point θ_c, which is determined in advance by the decision-maker on the underlying latent ability θ using standard-setting techniques, is taken into account.

The results of the simulation studies indicated that the average loss in the SMT and ASMT conditions decreased considerably compared to the fixed test condition, mainly due to a significant decrease of testlets administered. The number of correct decisions remained relatively stable. With the 3PL model, ASMT produced considerably better results than SMT, while with the 1PL model the results of ASMT were only slightly better. When testlet response behavior was simulated by a hierarchical IRT model with within-person ability variance, average loss increased. Ignoring the within-person variance in the decision procedure resulted in a further inflation of losses. Across studies, the minimal variance criterion (i.e., maximizing the expected reduction in the variance of the difference between the losses of the mastery and non-mastery decision) and selection of testlets with maximum information near the cut-off point θ_c produced the best results, but the difference with the maximum information at the EAP estimate of ability was very small.

The purpose of this article is to study a version of ASMT where response behavior is modeled by a multidimensional 1PL testlet model. The loss structure involved will be considered for both conjunctive (i.e., minimal requirements for each ability) and compensatory (i.e., low performance on one ability can be compensated by high performance on another

ability) testing strategies. The article concludes with a simulation study that aims on the gain of an SMT over a fixed-length mastery test and, in turn, the gain of an ASMT over an SMT using a multidimensional 1PL testlet model. As in Vos and Glas (2000), gain will be defined in terms of average loss, the average number of testlets administered, and the percentage of correct classification decisions.

Definition of the decision problem

In the following, it will be assumed that the variable-length mastery problem consists of S ($S \geq 1$) stages labeled $s = 1, ..., S$ and at each stage a testlet can be administered. This testlet consists of one or more items indexed with i and the observed item responses for a randomly sampled student will be denoted by a discrete random variable U_i, with realization u_i. Let the vector of item responses $\mathbf{u}_s$ be the response pattern to the s-th testlet. For $s = 1, ..., S$ the decisions will be based on a statistic $\mathbf{w}_s$ which is a function of the response patterns $\mathbf{u}_s$, that is, $\mathbf{w}_s = f(\mathbf{u}_1, ..., \mathbf{u}_s)$. In many cases, $\mathbf{w}_s$ will be the response pattern $\mathbf{u}_1, ..., \mathbf{u}_s$ itself. However, below it will become clear that some computations are only feasible if the information of the complete response pattern is aggregated. At each stage of testing s ($s = 1, ..., S-1$) a decision rule $d(\mathbf{w}_s)$ can be defined as

$$d(\mathbf{w}_s) = \begin{cases} m & \text{mastery decision} \\ n & \text{non-mastery decision} \\ c & \text{testing is continued.} \end{cases} \quad (1)$$

At the final stage of testing, stage S, only the two mastery classification decisions m and n are available. Mastery will be defined in terms of the latent proficiency continuum of the IRT model.

Multidimensional IRT models

Multidimensional IRT models are IRT models for response behavior where the responses depend on more than one latent ability. Multidimensional IRT models for dichotomously scored items were first presented by McDonald (1967) and Lord and Novick (1968). These authors use a normal ogive to describe the probability of a correct response. McDonald (1967,1997) developed an estimation procedure based on an expression for the association between pairs of items derived from a polynomial expansion of the normal ogive. The procedure is implemented in NOHARM (Normal-Ogive Harmonic Analysis Robust

Method, Fraser, 1988). An alternative approach using all information in the data, and therefore labeled "Full Information Factor Analysis" was developed by Bock, Gibbons, and Muraki (1988). This approach is a generalization of the marginal maximum likelihood (MML) and Bayes modal estimation procedures for unidimensional IRT models (see, Bock & Aitkin, 1981, Mislevy, 1986), and has been implemented in TESTFACT (Wilson, Wood, & Gibbons, 1991). A Bayesian estimation procedure using a Markov Chain Monte Carlo (MCMC) technique has been presented by Béguin and Glas (1998).

A comparable model using a logistic rather than a normal-ogive representation has been studied by Andersen (1985), Glas (1992), Reckase (1985, 1997) and Ackerman (1996a and 1996b). In the present article, the logistic version of the model will be used. In the logistic version, the probability of a correct response is given by

$$p(U_i = 1 \mid \theta_1, ..., \theta_Q, a_{i1}, ..., a_{iQ}, b_i, c_i) = c_i + (1 - c_i)\frac{\exp(\sum_q a_{iq}\theta_q - b_i)}{1 + \exp(\sum_q a_{iq}\theta_q - b_i)}, \qquad (2)$$

where $\theta_1, ..., \theta_Q$ are ability parameters, $a_{i1}, ..., a_{iQ}$ factor-loadings, b_i the item difficulty and c_i the guessing parameter. The probability of a response pattern

$$p(\mathbf{u} \mid \mathbf{a}, \mathbf{b}, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int, ..., \int p(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}, \mathbf{c}) g(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\theta}, \qquad (3)$$

with $p(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}, \mathbf{c})$ the probability of a response pattern given $\boldsymbol{\theta}$, which is derived from (2) using the assumption of local independence, and $g(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ the Q-variate normal distribution.

Compensatory and Conjunctive-Disjunctive Loss Functions

In the framework of the analysis of dichotomous dominance data, Coombs and Kao (1955, also see Coombs, 1960) make an important distinction between conjunctive-disjunctive and compensatory multidimensional models. The IRT model discussed above is a compensatory model because in determining the probability the ability dimensions are weighted with the factor loadings. However, the distinction between conjunctive, disjunctive and compensatory relations between latent variables and manifest variables can also be applied to define a loss structure.

Compensatory loss functions First an example of a compensatory loss structure will be given. Consider two dimensions. Let θ_1, θ_2 and θ_{1c} and θ_{2c} denote test taker's proficiency level and some pre-specified cut-off points in the latent space, respectively. Consider a line in the two-

dimensional proficiency space defined by $A_1(\theta_1 - \theta_{1c}) + A_2(\theta_2 - \theta_{2c}) = 0$. This line divides the latent space into two subspaces, persons with a proficiency in one subspace are masters, the persons in the other subspace are non-masters. The loss function for the master and non-master decision is given by

$$L(m, \theta_1, \theta_2) = max\{sC, sC + A_1(\theta_1 - \theta_{1c}) + A_2(\theta_2 - \theta_{2c})\} \tag{4}$$

with $A_1, A_2 < 0$ and

$$L(n, \theta_1, \theta_2) = max\{sC, sC + B_1(\theta_1 - \theta_{1c}) + B_2(\theta_2 - \theta_{2c})\}, \tag{5}$$

with $B_1, B_2 > 0$; C is the cost of delivering one testlet, sC is the cost of delivering s tests. To ensure that $B_1(\theta_1 - \theta_{1c}) + B_2(\theta_2 - \theta_{2c}) = 0$ defines the same line as $A_1(\theta_1 - \theta_{1c}) + A_2(\theta_2 - \theta_{2c}) = 0$, the additional constraint $A_1/A_2 = B_1/B_2$ is imposed. Notice that the loss structure is compensatory in the sense that a proficiency below a cut-off score on one dimension can be compensated by a proficiency above a cut-off score on the other dimension.

In Q dimensions, the loss function becomes

$$L(m, \boldsymbol{\theta}) = max\{sC, sC + \mathbf{A}'(\boldsymbol{\theta} - \boldsymbol{\theta}_c)\} \tag{6}$$

and

$$L(n, \boldsymbol{\theta}) = max\{sC, sC + \mathbf{B}'(\boldsymbol{\theta} - \boldsymbol{\theta}_c)\}, \tag{7}$$

where $\mathbf{A}$ and $\mathbf{B}$ are vectors of weights with all elements negative and positive, respectively, and $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_c$ are the ability vector and a vector of cut-off points, respectively. An additional constraint is that $\mathbf{A}'(\boldsymbol{\theta} - \boldsymbol{\theta}_c) = 0$ and $\mathbf{B}'(\boldsymbol{\theta} - \boldsymbol{\theta}_c) = 0$ define the same $(Q-1)$-dimensional linear sub-space.

Conjunctive loss functions In a conjunctive loss structure, a test taker is considered a master if the proficiency is above a cut-off point on all dimensions, and is considered a non-master if proficiency is below a cut-off point on any dimension. In two dimensions, this could be

translated into the following loss-function. Define

$$L(m,\theta_1,\theta_2) = \begin{cases} sC + A_1(\theta_1 - \theta_{1c}) + A_2(\theta_2 - \theta_{2c}) & \text{if } \theta_1 \leq \theta_{1c} \text{ and } \theta_2 \leq \theta_{2c} \\ sC + A_2(\theta_2 - \theta_{2c}) + A_3(\theta_1 - \theta_{1c})(\theta_2 - \theta_{2c}) & \text{if } \theta_1 > \theta_{1c} \text{ and } \theta_2 < \theta_{2c} \\ sC + A_1(\theta_1 - \theta_{1c}) + A_4(\theta_1 - \theta_{1c})(\theta_2 - \theta_{2c}) & \text{if } \theta_1 < \theta_{1c} \text{ and } \theta_2 > \theta_{2c} \\ sC & \text{if } \theta_1 > \theta_{1c} \text{ and } \theta_2 > \theta_{2c} \end{cases} \tag{8}$$

and

$$L(n,\theta_1,\theta_2) = \begin{cases} sC + (\theta_1 - \theta_{1c})^{B_1}(\theta_2 - \theta_{2c})^{B_2} & \text{if } \theta_1 > \theta_{1c} \text{ and } \theta_2 > \theta_{2c} \\ sC & \text{otherwise,} \end{cases} \tag{9}$$

for $A_1, A_2, A_3, A_4 < 0$ and $B_1, B_2 > 0$. Both loss functions are continuous, $L(n,\theta_1,\theta_2)$ is strictly positive and increasing on the space where $L(m,\theta_1,\theta_2)$ is equal to sC, in the same manner, $L(m,\theta_1,\theta_2)$ is strictly positive and decreasing on the space where $L(n,\theta_1,\theta_2)$ is sC. Notice that $L(m,\theta_1,\theta_2) = sC + A_1(\theta_1 - \theta_{1c})$ on the line $\theta_2 = \theta_{2c}$, and $L(m,\theta_1,\theta_2) = A_2(\theta_2 - \theta_{2c})$ on the line $\theta_1 = \theta_{1c}$.

Coombs and Kao (1955) show that conjunctive and disjunctive models are isomorph and only one mathematical model needs to be developed for the analysis of the problem. In the present case it is easily verified that choosing (8) as the definition for $L(n,\theta_1,\theta_2)$,(9) for the definition of $L(m,\theta_1,\theta_2)$ and setting $A_1, A_2, A_3, A_4 > 0$ and $B_1, B_2 < 0$ defines the loss structure for the disjunctive case.

At stage s, the decision whether the respondent is a master or a non-master, or whether another testlet will be administered, is based on the expected losses of the three possible decisions given the observation $\mathbf{w}_s$. The expected losses of the first two decisions are computed as

$$E(L(m,\boldsymbol{\theta}) \mid \mathbf{w}_s) = \int, ..., \int L(m,\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{w}_s) d\boldsymbol{\theta} \tag{10}$$

and

$$E(L(n,\boldsymbol{\theta}) \mid \mathbf{w}_s) = \int, ..., \int L(n,\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{w}_s) d\boldsymbol{\theta}, \tag{11}$$

where $p(\boldsymbol{\theta} \mid \mathbf{w}_s)$ is the posterior density of $\boldsymbol{\theta}$ given $\mathbf{w}_s$. The expected loss of the third possible decision is computed as the expected risk of continuing testing. If the expected risk of continuing testing is smaller than the expected loss of a master or a non-master decision, testing will be continued. The expected risk of continuing testing is defined as follows.

Let $\{\mathbf{w}_{s+1} \mid \mathbf{w}_s\}$ be the range of $\mathbf{w}_{s+1}$ given $\mathbf{w}_s$. Then, for $s = 1, ..., S-1$, the expected risk of continuing testing is defined as

$$E(R(\mathbf{w}_{s+1}) \mid \mathbf{w}_s) = \sum_{\{\mathbf{w}_{s+1}|\mathbf{w}_s\}} R(\mathbf{w}_{s+1}) p(\mathbf{w}_{s+1} \mid \mathbf{w}_s), \tag{12}$$

where the so-called posterior predictive distribution $p(\mathbf{w}_{s+1} \mid \mathbf{w}_s)$ is given by

$$p(\mathbf{w}_{s+1} \mid \mathbf{w}_s) = \int, ..., \int p(\mathbf{w}_{s+1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{w}_s) d\boldsymbol{\theta}, \tag{13}$$

and risk is inductively defined as

$$\begin{aligned} R(\mathbf{w}_{s+1}) = min\{&E(L(m, \boldsymbol{\theta}) \mid \mathbf{w}_{s+1}), \\ &E(L(n, \boldsymbol{\theta}) \mid \mathbf{w}_{s+1}), E(R(\mathbf{w}_{s+2}) \mid \mathbf{w}_{s+1})\}. \end{aligned} \tag{14}$$

The risk associated with the last testlet is defined as

$$R(\mathbf{w}_S) = min\{E(L(m, \boldsymbol{\theta}) \mid \mathbf{w}_S), E(L(n, \boldsymbol{\theta}) \mid \mathbf{w}_S)\}. \tag{15}$$

So, given an observation $\mathbf{w}_s$, the expected distribution of $\mathbf{w}_{s+1}, \mathbf{w}_{s+2}, ..., \mathbf{w}_S$ is generated and an inference about future decisions is made. Based on these inferences, the expected risk of continuation (12) is computed and compared with the expected losses of a mastery or non-mastery decision. If the risk of continuation is smaller than these two expected losses, testing is continued. If this is not the case the classification decision with the smallest expected loss is made.

Notice that the definitions (12) through (15) imply a recursive definition of the expected risk of continuation. In practice, the computation of the expected risk of continuing testing can be done by backward induction as follows. First, the risk of the last testlet is computed for all possible values of $\mathbf{w}_S$. Then the posterior predictive distribution $p(\mathbf{w}_S \mid \mathbf{w}_{S-1})$ is computed using (13), followed by the expected risk $E(R(\mathbf{w}_S) \mid \mathbf{w}_{S-1})$ defined in (12). This, in turn, can be used for computing the risk $R(\mathbf{w}_{S-1})$ for all $\mathbf{w}_{S-1}$ using (14), and this iterative process continues until s is reached and the decision can be made whether to administer testlet $s+1$, or to decide on mastery or non-mastery.

The Compound Multidimensional Rasch model

The theory presented thus far is applicable to the broad class of multidimensional IRT models defined above. The theory of adaptive sequential mastery testing will now be worked out in detail for a special case of the general model. In this so-called compound multidimensional Rasch model (Glas, 1992), it is assumed that the complete test, or, in the present case, the complete testlet, consists of Q sub-tests, where every sub-test relates to a specific ability θ_q, $q = 1, ..., Q$. Further, it is assumed that the ensemble of person parameters $\theta_1, ..., \theta_Q$ has a Q-variate normal distribution with a mean equal to zero and a covariance matrix Σ.

Given $\theta_1, ..., \theta_Q$, the probability of a response pattern $\mathbf{u}_1, ..., \mathbf{u}_Q$ is given by

$$\begin{aligned} p(\mathbf{u}_1, ..., \mathbf{u}_Q \mid \theta_1, ..., \theta_Q) &= \prod_{q=1}^{Q} \prod_{i=1}^{K_q} \frac{\exp(u_{qi}(\theta_q - b_{qi}))}{1 + \exp(\theta_q - b_{qi})} \\ &= \prod_{q=1}^{Q} \exp(t_q \theta_q) \exp(-\mathbf{u}_q' \mathbf{b}_q) P_{q0}(\theta_q), \end{aligned} \tag{16}$$

where $\mathbf{b}_q = (b_{1q}, ..., b_{qK_q})'$ is a vector of item parameters, $\mathbf{u}_q' \mathbf{b}_q$ is the inner product of $\mathbf{u}$ and t_q, $t_q = \sum_i u_{qi}$ is the sum score, and

$$P_{q0}(\theta_q) = \prod_{i=1}^{K_q} (1 + \exp(\theta_q - b_{qi}))^{-1}. \tag{17}$$

Notice that t_q is the minimal sufficient statistic for θ_q. Further, it is easily verified that $P_{q0}(\theta_q)$ is the probability, given θ_q, of a response pattern with all item responses equal to zero. The probability of observing t_q given θ_q is given by

$$\begin{aligned} p(t_q \mid \theta_q) &= \sum_{\{\mathbf{u}_q \mid t_q\}} p(\mathbf{u}_q \mid \theta_q) \\ &= \sum_{\{\mathbf{u}_q \mid t_q\}} \exp(t_q \theta_q - \mathbf{u}_q' \mathbf{b}_\mathbf{q}) P_{q0}(\theta_q) \\ &= \gamma_{t_q}(\mathbf{b}_q) \exp(t_q \theta_q) P_{q0}(\theta_q), \end{aligned}$$

with $\gamma_{t_q}(\mathbf{b}_q)$ an elementary symmetric function defined by

$$\gamma_{t_q}(\mathbf{b}_q) = \sum_{\{\mathbf{u}_q \mid t_q\}} \exp(-\mathbf{u}_q' \mathbf{b}_\mathbf{q}),$$

and where $\{\mathbf{u}_q \mid t_q\}$ stands for the set of all possible response patterns resulting in a sum score t_q.

Given $\boldsymbol{\theta} = (\theta_1, ..., \theta_Q)$, the probability of a response pattern $\mathbf{u} = (\mathbf{u}_1, ..., \mathbf{u}_Q)$ is given by

$$\begin{aligned} p(\mathbf{u} \mid \boldsymbol{\theta}) &= \prod_{q=1}^{Q} \exp(t_q \theta_q) \exp(-\mathbf{u}_q' \mathbf{b}_q) P_{q0}(\theta_q) \\ &= \exp(\mathbf{t}'\boldsymbol{\theta}) \exp(-\mathbf{u}'\mathbf{b}) P_0(\boldsymbol{\theta}), \end{aligned}$$

where $\mathbf{b} = (\mathbf{b}_1, ..., \mathbf{b}_Q)$ is a vector of item parameters, $\mathbf{t} = (t_1, ..., t_Q)$ and

$$P_0(\boldsymbol{\theta}) = \prod_{q=1}^{Q} P_{q0}(\theta_q).$$

The probability of observing $\mathbf{t}$ given $\boldsymbol{\theta}$ is given by

$$p(\mathbf{t} \mid \boldsymbol{\theta}) = \Gamma_{\mathbf{t}}(\mathbf{b}) \exp(\mathbf{t}'\boldsymbol{\theta}) P_0(\boldsymbol{\theta})$$

with $\Gamma_{\mathbf{t}}(\mathbf{b})$ is a product of the elementary symmetric functions $\gamma_{t_q}(\mathbf{b}_q)$ for $q = 1, ..., Q$. Below, $\Gamma_{\mathbf{t}}(\mathbf{b})$ will be referred to as a compound elementary symmetric function.

Usually the prior $\boldsymbol{\theta}$ is standard normal, so let $g(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})$ be the normal density with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Then

$$p(\boldsymbol{\theta} \mid \mathbf{t}) = \frac{p(\mathbf{t} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})}{p(\mathbf{t})} = \frac{\exp(\mathbf{t}'\boldsymbol{\theta}) P_0(\boldsymbol{\theta}) g(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})}{\int, ..., \int \exp(\mathbf{t}'\boldsymbol{\theta}) P_0(\boldsymbol{\theta}) g(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}) d\boldsymbol{\theta}}.$$

Notice that $\Gamma_{\mathbf{t}}(\mathbf{b})$ cancels from the nominator and denominator.

Applying the general framework of the previous section to the Rasch model boils down to choosing the minimal sufficient statistics for $\boldsymbol{\theta}$, that is, the unweighted sum scores for the statistics $\mathbf{w}_s$. So let t_{sq} be the score pattern on the q-th sub-test for the s-th occasion. Further, define $\mathbf{r}_s$ as a Q-vector with elements $r_{sq} = \sum_{d=1}^{s} t_{dq}$. Let $p(\boldsymbol{\theta} \mid \mathbf{r}_s)$ stand for the posterior density of proficiency given $\mathbf{r}_s$. Then the expected losses (10), (11) and the expected risk (12) can be written as $E(L(m,\boldsymbol{\theta}) \mid \mathbf{r}_s)$, $E(L(n,\boldsymbol{\theta}) \mid \mathbf{r}_s)$ and $E(R(\mathbf{r}_{s+1}) \mid \mathbf{r}_s)$. More specifically, the

expected risk is given by

$$E(R(\mathbf{r}_{s+1}) \mid \mathbf{r}_s) = \sum_{\mathbf{r}_{s+1}|\mathbf{r}_s} p(\mathbf{r}_{s+1} \mid \mathbf{r}_s) R(\mathbf{r}_{s+1}), \tag{18}$$

where the summation is over all scores $\mathbf{r}_{s+1}$ compatible with $\mathbf{r}_s$. Defining $\mathbf{z}_{s+1} = \mathbf{r}_{s+1} - \mathbf{r}_s$, the posterior predictive distribution (13) specializes to

$$\begin{aligned} p(\mathbf{r}_{s+1} \mid \mathbf{r}_s) &= \int, \ldots, \int p(\mathbf{r}_{s+1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{r}_s) d\boldsymbol{\theta} \\ &= \int, \ldots, \int \Gamma_{\mathbf{z}_{s+1}} \exp(\mathbf{z}'_{s+1}\boldsymbol{\theta}) P_{0(s+1)}(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{r}_s) d\boldsymbol{\theta}, \end{aligned} \tag{19}$$

where $\Gamma_{\mathbf{z}_{s+1}}$ is a shorthand notation for a compound elementary symmetric function of the item parameters of occasions $s+1$ and $P_{0(s+1)}(\boldsymbol{\theta})$ is equal to (17) evaluated using the item parameters of test $s+1$. That is, $P_{0(s+1)}(\boldsymbol{\theta})$ is equal to the probability of a zero response pattern on test $s+1$, given $\boldsymbol{\theta}$.

Simulation studies

A simulation study was designed to investigate the following four research questions. (1) What is the performance, in terms of average loss, of multidimensional IRT-based sequential mastery testing as a function of the number of items administered per testing stage? (2) What are the effects on average loss when turning the sequential procedure into an adaptive sequential procedure? (3) How is average loss in the sequential procedure influenced when ignoring the multidimensional structure and using a unidimensional IRT model? And finally, (4) how does ignoring the multidimensional structure affect the adaptive sequential procedure in terms of average loss?

Compensatory loss functions For all simulations pertaining to compensatory loss functions, a three-dimensional compound Rasch model was used. The parameters of the loss function were $(A_1, A_2, A_3) = (-1, -1, -1)$ and $(B_1, B_2, B_3) = (1, 1, 1)$, while the cost of administering one item was set equal to 0.02. The cut-off point was set equal to $\theta_c = 0$.

In the studies, the following aspects were varied:

- The correlation between the latent dimensions. The three-dimensional compound Rasch model was simulated in two conditions: a high-homogenity condition were the correlation

between all three dimensions was $\rho = 0.80$ and a low-homogenity condition were this correlation was $\rho = 0.40$.

- The test administration design. In the test procedure 27 items could be delivered. These items could be delivered as a fixed test of 27 items, or in a sequential design with 3 stages with 9 items per stage, 9 stages of 3 items, and 27 stages of one item.
- The test administration mode. Test administration could be either sequential or adaptive sequential. For the sequential procedure, the item difficulties b_i were drawn from a standard normal distribution. Further, the items were evenly distributed over the three ability dimensions, that is, a third of the items loaded on the first dimension, a third on the second, and a third on the third dimension. Finally, also within a stage the items were evenly distributed over the three dimensions, with the exception of the one-item stages, were items alternately loaded on a dimension. The item parameters were redrawn in every replication. For the adaptive sequential mode, a testlet bank was generated in such a way that it could be expected that it supported selection of testlets with differential optimal measurement properties. For the design of 27 stages of one items each, this was simply translated into drawing 375 item difficulties for each ability dimension from the standard normal distribution and choosing the optimal item via a selection criterium that will be outlined below. For the procedures with 3 and 9 stages, the following procedure was adopted:

  - define the grid $\{\mathbf{h}\} = \{h_1, h_2, h_3\} = \{h(i), h(j), h(k) | i, j, k = 1, ..., 5, h(n) = -1.0 + 0.5(n-1)\}$. Notice that this grid has 5^3, that is 125 points.
  - for each point $\mathbf{h} \in \{\mathbf{h}\}$, draw 3 item difficulties from the multivariate normal distribution defined by $\mathcal{N}(\mathbf{h}, 0.2\mathbf{I})$. Each item is assumed to load on a different dimension. This is repeated 3 times for each point $\mathbf{h} \in \{\mathbf{h}\}$. For the procedure with 3 stages, the 9 items form one testlet, for the procedure with 9 stages, three testlets of 3 items are formed. In this manner the total number of items available for the three procedures (27, 9 and 3 stages) remains constant, that is, equal to 1125.

  Also for the adaptive mode, the item difficulties were redrawn in every replication.

The choice of a criterium for adaptive testlet selection in a multidimensional framework is more complicated that in a unidimensional framework. In the latter framework, Vos and Glas (2000) studied three selection criteria. The first two entailed the choice of the testlet with maximum information at the cut-off point and at the expected-a-posteriori estimate of ability, respectively. In the multi-dimensional framework, these two criteria are less plausible. In one dimension, both the running estimate of ability and the cut-off point are on the same continuum, and any test with high information between these two points will be informative for the decision that has

to be taken. In a multidimensional framework, the test taker's ability is a point in Q-dimensional space and the boundary between masters and non-masters becomes a line in two dimensions, or a linear manifold in more than two dimensions. Therefore, in this case the relation between the position of the test taker in the support of the loss function and the optimal testlet will be much more complicated, and remains a point of further study.

As an alternative, the third criterion studied by Vos and Glas (2000) will be used. This approach is motivated by the fact that one is primarily interested in minimizing possible losses due to misclassifications.. The sequential procedure is based on comparing $L(m, \theta)$ and $L(n, \theta)$ to come to a decision. If, for every possible follow-up testlet $s + 1$, the observation $\mathbf{w}_{s+1}$ is available, a natural choice for the follow-up test is the testlet were the posterior variance of the difference between $L(m, \theta)$ and $L(n, \theta)$, say $var(L(m, \theta) - L(n, \theta) \mid \mathbf{w}_{s+1})$, was minimal. However, the observation $\mathbf{w}_{s+1}$ is not yet available, so a prediction must be made of the likelihood of $\mathbf{w}_{s+1}$. This likelihood is obtained via the predictive distribution $p(\mathbf{w}_{s+1} \mid \mathbf{w}_s)$. So if $\{\mathbf{w}_{s+1}|\mathbf{w}_s\}$ is the set of all possible values $\mathbf{w}_{s+1}$ given $\mathbf{w}_s$, the criterion for selection of the next testlet becomes

$$\sum_{\{\mathbf{w}_{s+1}|\mathbf{w}_s\}} var(L(m, \theta) - L(n, \theta) \mid \mathbf{w}_{s+1}) p(\mathbf{w}_{s+1} \mid \mathbf{w}_s), \tag{20}$$

that is, a testlet is chosen such that the expected variance of the difference between the losses of the mastery and non-mastery decision is minimal. In the study on the unidimensional case by Vos and Glas (2000) the performance of the three selection criteria was comparable, with a slight advantage for the procedure based on maximum information at the cut-off point.

Insert Table 1 and 2 about here

The results for the simulation studies for $\rho = 0.80$ and $\rho = 0.40$ are reported in Table 1 and Table 2, respectively. The results shown are a result of 1000 replications. For every replication a true ability θ was drawn from the standard normal distribution. At the end of every replication, loss was computed using the true ability value. In Table 1, it can be seen that the mean loss decreased with the number of items in a testlet. This decrease can be attributed to a decrease in the number of items given. The proportion of correct decisions did not decrease, in fact, it slightly increased. Finally, it can be seen that using an adaptive testlet selection procedure further decreased mean loss, but this decrease was far less important

than the decrease attributable to decrease of the testlet size. These findings are analogous to the findings of Vos and Glas (2000) for the unidimensional case.

The results for the study in the condition with $\rho = 0.40$ are shown in Table 2. It can be seen that the results are analogous to the results in Table 1, with the exception that all mean losses are systematically larger than in the condition where $\rho = 0.80$. This is explained by the fact that in the case of a homogeneous item pool, item responses are informative with respect to all ability dimensions, while in the heterogenous case, item responses are mainly informative with respect to the ability on which they load.

Insert Table 3 and 4 about here

In Table 3 and Table 4, the results are given for the conditions where the multidimensional ability structure is ignored in the computations supporting the sequential and adaptive sequential procedure. In this condition, response behavior was generated and the final mean losses were computed using the 'true' item and 'true' multidimensional ability parameters, while the computations supporting the sequential and adaptive sequential procedure were made using a standard unidimensional Rasch model with the 'true' item difficulties b_i and unidimensional standard normally distributed ability parameters. One could view this unidimensional approximation of multidimensional response behavior as an approximation based on the assumption that the correlation between the latent abilities is equal to one, i.e., $\rho = 1.0$. Therefore, in the unidimensional case, the losses (6) and (7) were computed using $\theta_1 = \theta_2 = \theta_3 = \theta$, where θ has a standard normal prior, and $\theta_{c1} = \theta_{c2} = \theta_{c3} = \theta_c = 0$. The results for the condition with $\rho = 0.80$ are shown in Table 3, the results for the condition with $\rho = 0.40$ are shown in Table 4.

It can be seen that, in general, the mean losses were higher than the analogous losses in Table 1 and 2, but the increase of the loss remained limited. Therefore, it must be concluded that the unidimensional approximation based on the assumption $\rho = 1.0$ worked quite well. Further, one might expect that the approximation in the case where $\rho = 0.40$ would be worse, but this expectation was not confirmed by the results. An important exception was the case of adaptive testlet selection with 27 testlets of one item each. In that case, the average loss for the adaptive sequential procedure became higher than the average loss in the non-adaptive sequential testlet selection procedure. So there the combination of a unidimensional approximation of ability with the circumstance that the testlets only loaded on one ability dimension resulted in a relatively poor performance.

Conjunctive loss functions For all simulations pertaining to conjunctive loss functions, a two-dimensional compound Rasch model was used. The parameters of the loss function $A_1, ..., A_4$ were all equal to -0.5,and the paparmeters B_1 and B_2 wire both equal to 1. The cost of administering one item was set equal to 0.01 and the cut-off point was set equal to $\theta_c = 0$.

In the studies, the following aspects were varied:

- The correlation between the latent dimensions: $\rho = 0.80$ and $\rho = 0.40$.
- The test administration design. In the test procedure 32 items could be delivered. These items could be delivered as a fixed test of 32 items, or in a sequential design with 4 stages with 8 items per stage, 8 stages of 4 items, and 32 stages of one item.
- The test administration mode: sequential or adaptive sequential. For the sequential procedure, the item difficulties b_i were drawn from a standard normal distribution. Further, the items were evenly distributed over the two ability dimensions, that is, half of the items loaded on the first dimension and half loaded on the second dimension. Finally, also within a stage the items were evenly distributed over the two dimensions, with the exception of the one-item stages, were items alternately loaded on a dimension. The item parameters were redrawn in every replication. For the adaptive sequential mode, a testlet bank was generated in such a way that it could be expected that it supported selection of testlets with differential optimal measurement properties. For the design of 32 stages of one items each, this was simply translated into drawing 100 item difficulties for each ability dimension from the standard normal distribution and choosing the optimal item via a selection criterium that will be outlined below. For the procedures with 32, 8 and 4 stages, the following procedure was adopted:
  - define the grid $\{\mathbf{h}\} = \{h_1, h_2\} = \{h(i), h(j) | i, j = 1, ..., 5, h(n) = -1.0 + 0.5(n-1)\}$. Notice that this grid has 5^2, that is 25 points.
  - for each point $\mathbf{h} \in \{\mathbf{h}\}$, draw 2 item difficulties from the multivariate normal distribution defined by $\mathcal{N}(\mathbf{h}, 0.2\mathbf{I})$. Each item is assumed to load on a different dimension. This is repeated 4 times for each point $\mathbf{h} \in \{\mathbf{h}\}$. For the procedure with 4 stages, the 8 items form one testlet, for the procedure with 8 stages, two testlets of 4 items are formed. In this manner the total number of items available for the two procedures (32, 8 and 4 stages) remains constant, that is, equal to 200.

Also for the adaptive mode, the item difficulties were redrawn in every replication.

Insert Table 5 to 8 about here

Contrary to the unidimensional approximation of the compensatory model, the unidimensional approximation does not work well for the conjunctive model. The reason for the poor performance for the unidimensional approximation of the two-dimensional conjunctive model is that there are many non-masters in the region $\{\theta_1 > \theta_{1c}$ and $\theta_2 < \theta_{2c}\}$ and $\{\theta_1 < \theta_{1c}$ and $\theta_2 > \theta_{2c}\}$ that still obtain a sum score to make them eligible for a mastery decision in the unidimensional approximation through a compensatory process where a low ability and a low sum score on one dimension is compensated by a high ability and sum score on the other dimension. This can be seen in Table 9, where the proportion of wrongly identified non-masters is much higher than the proportion of non-masters wrongly identified. Notice that this proportion is negatively related to the correlation. In the compensatory model, the error-proportions are symmetric and approximately equal to 0.10.

Insert Table 9 about here

Conclusions and Further Research

In this article, a general theoretical framework for non-adaptive and adaptive sequential testing based on a combination of Bayesian sequential decision theory and multidimensional IRT was presented. This framework was applied to the compound Rasch model. In this model it is assumed that the test items can be split up into a number of subsets related to specific ability dimensions and the relation between the dimensions is modeled by a covariance structure. Using this model, a number of simulation studies were performed which showed that augmentation of the number of stages in a sequential mastery procedure resulted in a marked decrease of average loss. Moving to adaptive sequential mastery testing further reduced average loss, but the effect was far less important than the effect of a non-adaptive sequential procedure. For the compensatory model, the results of the simulation studies showed that ignoring the multidimensional structure and using a unidimensional approximation to the multi-dimensional model did not generally result in an important increase in average losses. An exception was adaptive sequential testing with only one item per testlet and a low correlation of the ability dimensions. In that case, the average loss was higher that in the analogous case without adaptive item selection. For the conjunctive model, the unidimensional approximation was very poor.

For application of the general framework for non-adaptive and adaptive sequential testing presented here to more general multidimensional IRT models, two important issues will

require further research. Firstly, the computation of the multiple integrals is done using Gauss-Hermite quadrature, which becomes very time-consuming when more than three dimensions are involved (see, for instance, Glas, 1992). Therefore, problems of higher dimensionality will need simulation methods for the evaluation of the multiple integrals. Secondly, many multidimensional IRT models, like, for instance, the "Full Information Factor Analysis" model by Bock, Gibbons, and Muraki (1988) have no sufficient statistics for $\boldsymbol{\theta}$, and will need alternative choices for $\mathbf{w}_s = f(\mathbf{u}_1, ..., \mathbf{u}_s)$. For the unidimensional 3PL model, Vos and Glas (2000) show that using unweighted sum scores results in a feasible procedure that produces acceptable results. A generalization to a multi-dimensional framework would probably be based on a Q-dimensional vector of partial sum scores, but this remains a point of further study.

References

Ackerman, T.A. (1996a). Developments in multidimensional item response theory. *Applied Psychological Measurement, 20,* 309-310.

Ackerman, T.A. (1996b). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20,* 311-329.

Andersen, E.B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50,* 3-16.

Béguin, A.A. & Glas, C.A.W. (1998). *MCMC estimation of multidimensional IRT models.* [Research Report 98-14], University of Twente, Enschede.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika, 46,* 443-459.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Coombs, C.H. (1960). *A theory of data.* Ann Arbor, MA: Mathesis Press.

Coombs, C.H. & Kao, R.C. (1955). *Nonmetric factor analysis.* [Engng. Res. Bull., No.38]. Ann Arbor: University of Michigan Press.

DeGroot, M.H. (1970). *Optimal statistical decisions.* New York NJ: McGraw-Hill.

Fraser, C. (1988). NOHARM: *A Computer Program for Fitting Both Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory.* NSW: University. of New England.

Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson, (Ed.), *Objective measurement: Theory into practice, Vol. 1* (pp.236-258). New Jersey, NJ: Ablex Publishing Corporation.

Lehmann, E.L. (1986). *Testing statistical hypothesis. (second edition).* New York, NJ: Wiley.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* 367-386.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric monographs, No.15.*

McDonald, R.P. (1997). Normal-ogive multidimensional model. In W.J. van der Linden and R.K. Hambleton (eds.): *Handbook of Modern Item Response Theory.* (pp.257-269). New York: Springer.

Reckase, M.D. (1983). A procedure for decision making using tailored testing. In D.J. Weiss (Ed.): *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237-257). New York, NJ: Academic Press.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden and R.K. Hambleton (eds.). *Handbook of Modern Item Response Theory.* (pp.271-286). New York, NJ: Springer.

Vos, H.J., & Glas, C.A.W. (2000) Testlet-Based Adaptive Mastery Testing. In W.J.van der Linden & C.A.W.Glas (eds.): *Computerized adaptive testing: Theory and practice.* (pp.289-310). Boston, MA: Kluwer.

Wilson, D.T., Wood, R., & Gibbons, R. (1991) *TESTFACT: Test scoring, Item statistics, and Item Factor Analysis.* (Computer Software). Chicago: Scientific Software International, Inc.

Table 1
Relation between selection method and loss
compensatory model, $\rho = 0.80$

Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
1	27	Fixed Test	0.81	1.00	0.7079
3	9	Sequential	0.79	0.38	0.4443
3	9	Adaptive	0.79	0.27	0.3777
9	3	Sequential	0.78	0.25	0.3972
9	3	Adaptive	0.78	0.25	0.3408
27	1	Sequential	0.79	0.25	0.3446
27	1	Adaptive	0.80	0.22	0.3060

Table 2
Relation between selection method and loss
compensatory model, $\rho = 0.40$

Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
1	27	Fixed Test	0.77	1.00	0.7654
3	9	Sequential	0.72	0.38	0.5387
3	9	Adaptive	0.73	0.26	0.4696
9	3	Sequential	0.73	0.25	0.4652
9	3	Adaptive	0.73	0.22	0.4169
27	1	Sequential	0.73	0.22	0.4109
27	1	Adaptive	0.73	0.21	0.3988

Table 3
Relation between selection method and loss
when multidimensionality is ignored
compensatory model, $\rho = 0.80$

Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
1	27	Fixed Test	0.81	1.00	0.6985
3	9	Sequential	0.81	0.41	0.4248
3	9	Adaptive	0.81	0.43	0.4138
9	3	Sequential	0.77	0.28	0.4074
9	3	Adaptive	0.80	0.27	0.3457
27	1	Sequential	0.80	0.27	0.3721
27	1	Adaptive	0.80	0.24	0.3295

Table 4
Relation between selection method and loss
when multidimensionality is ignored
compensatory model, $\rho = 0.40$

Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
1	27	Fixed Test	0.76	1.00	0.8200
3	9	Sequential	0.73	0.40	0.5781
3	9	Adaptive	0.73	0.43	0.5017
9	3	Sequential	0.70	0.29	0.4838
9	3	Adaptive	0.75	0.27	0.4484
27	1	Sequential	0.76	0.27	0.4023
27	1	Adaptive	0.71	0.23	0.4429

Table 5
Relation between selection method and loss
conjunctive model, $\rho = 0.80$

Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
1	32	Fixed Test	0.85	1.00	0.3549
4	8	Sequential	0.82	0.30	0.1475
4	8	Adaptive	0.80	0.26	0.1396
8	4	Sequential	0.78	0.22	0.1306
8	4	Adaptive	0.80	0.21	0.1302
32	1	Sequential	0.79	0.20	0.1277
32	1	Adaptive	0.80	0.20	0.1270

Table 6
Relation between selection method and loss
conjunctive model, $\rho = 0.40$

Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
1	32	Fixed Test	0.80	1.00	0.3999
4	8	Sequential	0.81	0.30	0.1765
4	8	Adaptive	0.81	0.24	0.1588
8	4	Sequential	0.81	0.23	0.1570
8	4	Adaptive	0.82	0.20	0.1377
32	1	Sequential	0.80	0.19	0.1375
32	1	Adaptive	0.81	0.19	0.1373

Table 7
Relation between selection method and loss
when multidimensionality is ignored
conjunctive model, $\rho = 0.80$

Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
1	32	Fixed Test	0.47	1.00	0.6208
4	8	Sequential	0.46	0.30	0.4581
8	4	Sequential	0.49	0.24	0.4247
32	1	Sequential	0.49	0.28	0.4340

Table 8
Relation between selection method and loss
when multidimensionality is ignored
conjunctive model, $\rho = 0.40$

Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
1	32	Fixed Test	0.41	1.00	0.6523
4	8	Sequential	0.43	0.31	0.5040
8	4	Sequential	0.41	0.25	0.5136
32	1	Sequential	0.44	0.29	0.4878

Table 9
Pattern of correct and incorrect decisions for unidimensional approximation of conjunctive model

Correlation				
0.40		Decision		
	State	Mastery	Non-mastery	Total
	Mastery	0.23	0.08	0.31
	Non-mastery	0.51	0.18	0.69
	Total	0.74	0.26	1.00
0.80		Decision		
	State	Mastery	Non-mastery	Total
	Mastery	0.29	0.10	0.39
	Non-mastery	0.43	0.18	0.61
	Total	0.72	0.28	1.00

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

RR-98-09 B.P. Veldkamp, *Multiple Objective Test Assembly Problems*

RR-98-08 B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

RR-98-07 W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

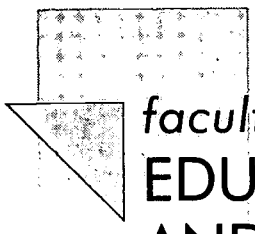
RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.



faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM032321

NOTICE

REPRODUCTION BASIS

This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").