ON THE USE OF MCMC CAT WITH EMPIRICAL PRIOR INFORMATION TO IMPROVE THE EFFICIENCY OF CAT[1]

MARIAGIULIA MATTEUCCI

DEPARTMENT OF STATISTICAL SCIENCES,

UNIVERSITY OF BOLOGNA, ITALY

e-mail: m.matteucci@unibo.it

BERNARD P. VELDKAMP

RESEARCH CENTER FOR EXAMINATION AND CERTIFICATION,

UNIVERSITY OF TWENTE, THE NETHERLANDS

e-mail: B.P.Veldkamp@gw.utwente.nl

1 Corresponding author: Mariagiulia Matteucci, Department of Statistics "P. Fortunati", University of Bologna, via Belle Arti 41, 40126 Bologna (Italy). E-mail: m.matteucci@unibo.it

ON THE USE OF MCMC CAT WITH EMPIRICAL PRIOR INFORMATION TO IMPROVE THE EFFICIENCY OF CAT

Abstract

In this paper, empirical prior information about the candidate is applied in computerized adaptive testing (CAT). The main objective of CAT is to improve efficiency of test administration. In this paper, it is shown how the inclusion of background variables both in the initialization and the ability estimation is able to improve the accuracy of ability estimates. In particular, a Gibbs sampler scheme is proposed in the phases of interim and final ability estimation. By using both simulated and real data, it is demonstrated that the method produces more accurate ability estimates, especially for short tests and when reproducing boundary abilities. This implies that operational problems of CAT related to weak measurement precision under particular conditions, can be reduced as well. In the empirical example, the methods were applied to CAT for intelligence testing in the area of personnel selection. Other promising applications would be in the medical world, where testing efficiency is of paramount importance as well.

Key words: adaptive testing, empirical prior information, Gibbs sampler, measurement precision.

Introduction

In recent years, we have seen a rapid development of computer-based testing in the field of psychological measurement, especially in adaptive testing. The practice of conducting the test administration via adaptive testing is becoming more and more well-established. Since the early 1970s (Lord, 1970; 1971; Owen, 1969;1975), studies have been conducted to develop the theoretical framework of computerized adaptive testing (CAT) (see e.g., van der Linden and Glas, 2000; Wainer et al., 2000). The basic idea of CAT is to adapt the difficulty of the items to the estimated ability level of the candidate. In this way, the behavior of a real oral examiner during a testing occasion is simulated. In fact, most oral examinations start with an initial item and, depending on the examinee's response, proceed with a more difficult or easier item, until the examinee's grade of proficiency becomes sufficiently precise. Analogously, in computerized adaptive testing a first item is submitted to the test-taker: if the item is endorsed, a more difficult item is presented, otherwise an easier one is selected by the algorithm to be submitted. The procedure ends when a pre-specified criterion is met. Finally, the estimated ability is reported as a measure of the examinee's proficiency.

CAT relies strongly on item response theory (IRT), developed in order to estimate individual and item characteristics after a test administration (see e.g., Lord and Novick, 1968). In fact, items are selected from an item pool that is calibrated with a particular IRT model, based on data nature and fit, and the response process is assumed to follow the IRT model.

Despite the wide use of computerized adaptive testing, the method has a number of operational problems like item pool maintenance (Ariel, van der Linden, and Veldkamp,

2006; Belov and Armstrong, 2009), test assembly (van der Linden, 2005), item exposure control (e.g. Sympson and Hetter, 1985; van der Linden and Veldkamp, 2004, 2007), and recovery from unforced errors during the beginning of CAT (Guyer, 2008). Furthermore, technical issues, such as initialization and ability estimation, might be improved, especially when only a restricted number of items can be submitted.

In this study, the focus is on the use of collateral information about the candidate in CAT. In many situations, much information about the candidate is available. For example, bio data, educational level, and information about work experience might be available in a job selection context. In educational settings, results on previous tests, social economic status, or the educational level of the parents might be available. Besides, it often happens that a whole battery of tests is administered to the candidate during an exam, or during a psychological screening. The question arises how all of this information could be used to improve the efficiency of the CAT.

Collateral information may be included in CAT in two different stages. Firstly, the initialization of ability estimate can make use of prior information (see Gialluca and Weiss, 1979; van der Linden, 1999). As a consequence, a better provisional ability estimate is provided and the first item is selected closer to the true ability of the person. Secondly, background variables may be included in the estimation process through an empirical prior distribution. Two different problems will be solved by using complementary information in CAT. First of all, the test length will be reduced. Additionally, bias due to unforced errors during the beginning of CAT (Guyer, 2008) will be reduced as well, since the impact of these errors on ability estimation is much smaller due to the use of an informative prior.

A natural way of developing this approach is represented by Bayesian statistics, where likelihood and prior distributions are combined in order to obtain the posterior

distribution of interest. Recently, Markov chain Monte Carlo (MCMC) methods, and particularly the Gibbs sampler (Geman and Geman, 1984), have been applied extensively in IRT estimation because they are able to provide flexible algorithms for a large variety of models, such as unidimensional models (Albert, 1999; Johnson and Albert, 1999; Patz and Junker, 1999), multidimensional models (Béguin and Glas, 2001; Sheng and Wikle, 2007; 2008) and models with a hierarchical structure (Fox and Glas, 2001; Sheng and Wikle, 2008; Natesan, Limbers, and Varni, 2010). Basically, the advantages of using MCMC are twofold. Firstly, the method is able to integrate all dependencies between variables and allows the specification of different prior distributions depending on the researcher's previous knowledge. This particular aspect makes the Gibbs sampler a flexible and powerful statistical tool. Secondly, MCMC is free from the technical limitations of the Gaussian quadrature involved in the marginal maximum likelihood (MML) estimation (Béguin and Glas, 2001; Sheng and Wikle, 2007). Moreover, with modern computers, MCMC computer-intensiveness has been strongly reduced.

By introducing the empirical prior within MCMC, the posterior distribution becomes candidate-tailored and more precise ability estimates can be obtained. In the paper of van der Linden (1999) it is shown how prior information can be included in the ability initialization. The purpose of this paper is to show how collateral information can be used even more efficiently by introducing it both in initialization and ability estimation. Furthermore, the paper describes how the empirical prior can be integrated in the estimation process within the Gibbs sampler scheme.

The paper first gives an overview of how prior information can be included in CAT. Then, it is shown how the Gibbs sampler can be implemented in computerized adaptive testing effectively in order to integrate information coming from both likelihood and

prior distributions. The advantages of introducing background variables in CAT administration are discussed through a set of comparative simulation studies, by using first a variable-length termination criterion, and then a fixed-length one. The number of items needed to complete the CAT and the level of ability precision are evalutated in case empirical priors are introduced instead of standard priors. The issue of convergence of the MCMC algorithm is addressed briefly. Finally, the results of an empirical CAT application are presented in the context of intelligence testing for personnel selection.

Adaptive Testing with Empirical Prior Information

In testing occasions, besides the candidates' responses on a target test, a set of individual covariates may be available. Background variables may include scores obtained by the examinees on other tests or testlets, socio-economic, or demographical variables. Moreover, response times can represent an effective source of information about individual ability (van der Linden, 2008; van der Linden and Pashley, 2010). Given the availability of such information, its inclusion in the investigation of candidates' ability might make sense. Whether and how collateral information about examinees may be included in IRT ability initialization and estimation has been discussed by various authors (e.g., Zwinderman, 1991; 1997; van der Linden, 1999; Matteucci and Veldkamp, 2011; Matteucci, Mignani, and Veldkamp, 2009). As reported in van der Linden and Pashley (2010), one reason for introducing collateral information about the candidates in adaptive testing is CAT weakness in ability estimation when dealing with short tests, caused by a possible bad start in the ability initialization. Even if it is well known that the convergence of the algorithm is not affected by the choice of starting values, a rough initial inference about ability may cause a very slow

convergence (Guyer, 2008). In the following, the different steps of CAT with empirical prior are described. A particular section is dedicated to the ability estimation.

The Phases of CAT

In computerized adaptive testing, the item parameters are typically treated as known and the main purpose of test administration is the ability estimation of test takers. Item parameters are estimated on the basis of a particular IRT model, and stored in an item pool. The IRT model should be able to reproduce the individuals' response process; therefore, it describes the mathematical function linking the response probability to a set of item parameters and ability. Once the item parameters have been estimated with sufficient precision, items with target features are included in the item pool to be administered. The choice of the model depends on different issues such as item format, dimensionality specification, and fit. For the purpose of this study, the unidimensional two-parameter normal ogive (2PNO) model (Lord, 1952; Lord and Novick, 1968) is assumed to underlie the response process. The model has been designed for binary observed data, employing a cumulative standard normal distribution to express the probability of a correct response to an item *j*, with *j=1,...,k* items, as a function of ability and item parameters, as follows

$$P(Y_j = 1 \mid \theta) = \Phi(\alpha_j \theta - \delta_j) = \int_{-\infty}^{\alpha_j \theta - \delta_j} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \qquad (1)$$

where Y_j is the random response variable for item *j*, taking the value *1* for a correct response and *0* otherwise, α_j and δ_j are the item discrimination and difficulty respectively, and θ is the unidimensional ability. Model (1) assumes unidimensionality,

i.e., a single latent trait accounts for the individual responses. Depending on the data characteristics, other models are possible and have been employed in CAT.

Once the items have been calibrated according to an IRT model, computerized adaptive testing works with the following steps:

1. Ability initialization
2. Item selection
3. Item administration
4. Ability estimate update.

Steps 2-4 are repeated iteratively until a stopping rule is satisfied and a final estimate of the candidate's ability is obtained. An empirical prior may be introduced both in the initialization of the algorithm (step 1) and in the interim-final ability estimation (step 4).

In order to introduce empirical information in the first step, a relation between the ability θ and a set of P individual covariates $\{X_p\}$, with $p=1,\ldots,P$, is assumed in the form of a linear regression, as follows

$$\theta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_P X_{iP} + \varepsilon_i, \qquad (2)$$

where the error terms are assumed to be independent and normally distributed as $\varepsilon_i \sim \mathrm{N}(0,\sigma^2)$, with $i=1,\ldots,n$ individuals. The assumption of a linear regression model is translated into a normal conditional distribution of θ_i given the covariates, as

$$\theta_i \mid X_{i1},\ldots,X_{iP} \sim N(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_P X_{iP};\sigma^2). \qquad (3)$$

Equation (3) represents the informative prior distribution for ability. When regression (2) is estimated with satisfying precision and the quality of the background variables is

good (i.e., they are highly informative predictors), the estimated regression coefficients may be used in order to initialize the ability in CAT for a generic examinee *i* with realizations ($x_{i1},\ldots,x_{iP}$), as follows

$$\hat{\theta}_{i0} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_P x_{iP}. \quad (4)$$

The advantage of using *ad hoc* information to initialize the algorithm is mainly to shorten the procedure. Within this approach, initial values may be much more reliable and accurate initial inferences about ability could be able to shorten time to convergence significantly. As discussed in van der Linden and Pashley (2010), the choice of the prior distribution should be taken carefully. In fact, in the initial phase of CAT no response data are available and the choice of the first item is completely determined by the empirical information. When the prior is not reliable, the examinee's initial ability may be located far from the true ability and needs more time to be recovered. However, this consideration is also valid for fixed initialization: when $\hat{\theta}_0 = 0$ is imposed as the initial ability estimate for all candidates, the recovery of the true ability for examinees with high or low θ values is seriously compromised within short tests.

Before proceeding with item selection (step 2), the following notation on CAT is introduced. Given *J* calibrated items in the pool, indexed by $j=1,\ldots,J$, denote the rank of selected items as $k=1,\ldots,K$. Hence, when choosing the *k*th item to be administered: j_k is the index of the chosen item, $S_{k-1}=\{j_1,j_2,\ldots,j_{k-1}\}$ is the set of selected items and $R_k=\{1,\ldots,J\}\backslash S_{k-1}$ is the set of remaining items in the pool. In the following, the index $i=1,\ldots,n$ of examinees is omitted and the test administration is referred to a generic candidate *i* implicitly.

In order to proceed with the item selection (step 2), various criteria have been proposed in the literature. A classical and straightforward method which is also applied in linear testing is the maximum-information criterion (Birnbaum, 1968). When selecting the *k*th item, the method works choosing the item which maximizes Fisher's expected information function at the current ability value $\theta = \hat{\theta}_{k-1}$, as follows

$$j_k \equiv \arg\max_j \{I_j(\hat{\theta}_{k-1}); j \in R_k\}. \quad (5)$$

The form of the information function depends on the particular chosen IRT model. According to model (1), the information function becomes

$$I_j(\hat{\theta}_{k-1}) = \alpha_j^2 \frac{\{(2\pi)^{-1/2} \exp(-\eta_j^2/2)\}^2}{\Phi(\eta_j)[1-\Phi(\eta_j)]}, \quad (6)$$

where $\eta_j = \alpha_j \hat{\theta}_{k-1} - \delta_j$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The method is widely used; nevertheless, the maximum-information criterion associated with a fixed ability initialization leads to the problem of item overexposure, because the same item is always selected as the first one.

Following the CAT algorithm through step 3, the chosen item is administered to the test-taker and the answer is recorded. The response is subsequently used in step 4, when ability should be estimated. Steps 2-4 of the algorithm are repeated iteratively until a stopping rule is satisfied, as a fixed test length or a pre-specified level of precision for the ability estimate.

One crucial issue in CAT certainly is the measurement precision of ability estimates. Typically, standard errors of ability score estimates are not negligible and efforts in the

direction of improving the accuracy of ability estimates should be done. In fact, the task of obtaining an accurate ability estimate is particularly hard when poor information comes from the responses or when the examinee's level of proficiency is extreme (very high or very low).

In adaptive testing, a number of methods for the ability estimation are in use. These include maximum likelihood (ML) procedures or Bayesian methods (see van der Linden and Pashley, 2010). Because ML estimates stay undetermined until a mixed response pattern is observed, Bayesian methods could be preferred for ability estimation. Therefore, due to its growing and relatively new use in IRT, a Gibbs sampler scheme is implemented for ability estimation in CAT. The algorithm, as shown in Matteucci, Mignani, and Veldkamp (2009), is able to integrate efficiently data coming from individual responses and empirical prior information. The method is illustrated in detail in the next section.

MCMC Ability Estimation

To perform a Bayesian ability estimation in CAT, the Gibbs sampler (Geman and Geman, 1984) is implemented. The algorithm belongs to the family of MCMC methods which introduce simulation for the purpose of reproducing a target distribution by using one or more sequences of correlated random variables. According to the Bayesian approach, both ability and item/regression parameters are regarded as random variables. Once all components of the joint posterior distribution of interest have been individuated, the single conditional distributions should be specified. The Gibbs sampler works by creating suitable samples from each single conditional distribution iteratively until convergence. Among others, Albert (1999), Béguin and Glas (2001), Fox and Glas (2001), and Matteucci, Mignani, and Veldkamp (2009) dealt with Gibbs

sampler estimation within item response theory models. In the current work, the algorithm is modified in order to estimate ability in adaptive testing with the inclusion of an informative empirical prior.

Generally, the presence of the binary response variable Y_j can be modeled by introducing continuous underlying variables Z_j, which are independent and identically distributed as $Z_j \sim N(\alpha_j\theta - \delta_j; 1)$. The relation between the observed and the underlying variables is the following

$$Y_j = \begin{cases} 1 & if \quad Z_j > 0, \\ 0 & if \quad Z_j \leq 0. \end{cases} \tag{7}$$

According to Equation (7), the continuous variable *Z* is greater than zero if and only if the corresponding observed response is a success, i.e. *Y=1*; the *underlying variable* approach (Bartholomew, 1987; Bartholomew and Knott, 1999) describes the partition of the continuous variable *Z* in order to represent the dichotomy of *Y*.

From a fully Bayesian perspective, the joint posterior distribution of interest is

$$P(\mathbf{Z}, \theta, \xi, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}, \mathbf{X}) = P(\mathbf{Z} \mid \theta, \xi, \mathbf{Y}) P(\theta \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}) P(\xi) P(\boldsymbol{\beta}) P(\sigma^2), \tag{8}$$

where ξ is the vector including all item parameters. In linear testing, given the data on the responses and the observed covariates, the Gibbs sampler would have worked iteratively sampling from the following single conditional distributions:

1. $\mathbf{Z} \mid \theta, \xi$
2. $\theta \mid \mathbf{Z}, \xi, \boldsymbol{\beta}, \sigma^2$
3. $\xi \mid \theta, \mathbf{Z}$

4. $\boldsymbol{\beta} \mid \theta, \sigma^2$
5. $\sigma^2 \mid \theta, \boldsymbol{\beta}$.

On the other hand, in adaptive testing both item and regression parameters are treated as known; therefore, their conditional distributions are not needed in the scheme. In CAT, the Gibbs sampler works only with the conditional distribution of the underlying response variables Z_j (distribution in step 1) and the posterior distribution of the ability θ (distribution in step 2), in order to proceed with the ability estimation. The single conditional distributions, compared to the joint posterior, are treatable and easy to draw samples from.

With regard to the first conditional distribution, a classical result (see e.g., Johnson and Albert, 1999, chapter 3) is that the distribution of each Z_j given the ability and the item parameters is a truncated normal, as follows

$$Z_j \mid \theta, \xi \sim \begin{cases} N(\eta_j;1) \quad with \quad Z_j > 0 \quad if \quad Y_j = 1, \\ N(\eta_j;1) \quad with \quad Z_j \leq 0 \quad if \quad Y_j = 0. \end{cases} \tag{9}$$

The conditional distribution of the underlying variables Z_j is normal, with expected value equal to $\eta_j = \alpha_j \theta - \delta_j$ and variance 1, truncated by 0 to the left if $Y_j=1$ (correct response to item j) and to the right if $Y_j=0$ (incorrect response to item j).

The second conditional distribution is obtained combining the likelihood and the informative prior distribution, according to Bayesian conjugate families of distributions. Starting from the normal regression model $Z_j = \alpha_j \theta - \delta_j + \upsilon_j$ for $j=1,\ldots,J$, we obtain

$$Z_j + \delta_j = \alpha_j \theta + \upsilon_j, \tag{10}$$

where υ_j are *independent and identically distributed as N(0;1)*. Equation (10) is simply the regression of the terms on the left side $Z_j+\delta_j$ on the independent variable α_j, where θ is the regression coefficient. Hence, the likelihood function of the ability θ follows a normal distribution, as

$$\theta \sim \mathrm{N}(\hat{\theta}; v), \tag{11}$$

where $\hat{\theta} = (\alpha_j^{'}\alpha_j)^{-1}\alpha_j^{'}(Z_j + \delta_j)$ is the least square estimate of θ and $v = (\alpha_j^{'}\alpha_j)^{-1}$ is the variance. Practically, the variance can be calculated as $v = 1/\sum_{j=1}^{J}\alpha_j^2$ and the expected value as $\hat{\theta} = \sum_{j=1}^{J}\alpha_j(Z_j + \delta_j)/\sum_{j=1}^{J}\alpha_j^2$. The prior distribution for the ability is the empirical normal prior (3) and the combination of likelihood and prior leads to a normal posterior distribution, as follows

$$\theta \mid \mathbf{Z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}\left(\frac{\hat{\theta}/v + \mathbf{X}\boldsymbol{\beta}/\sigma^2}{1/v + 1/\sigma^2}; \frac{1}{1/v + 1/\sigma^2}\right). \tag{12}$$

After the *k*th item has been administered, the Gibbs sampler is able to simulate ability as follows:

1. Start with known item parameters $\boldsymbol{\xi}$ and a provisional estimate of $\theta_k^{(0)}$, $\theta_k^{(0)} \equiv \theta_{k-1}$, and sample $\boldsymbol{Z}^{(0)}$ from distribution (9), with $j \in S_k$.
2. Use $\boldsymbol{Z}^{(0)}$ and known $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, σ^2 to sample $\theta_k^{(1)}$ from distribution (12).
3. Repeat steps 1-2 with the updated values, iteratively.

The steps describe the estimation of the interim ability. Simply, after the last item has been administered, the same steps may be applied with the updated likelihood in order

to obtain the final ability estimate. The Gibbs sampler has been implemented in the software MATLAB 7.1 (The MathWorks Inc., 2005) .

In order to compare the accuracy of ability estimates in adaptive testing by using different criteria for the initialization and the ability estimation, simulation studies are conducted under different conditions.

Simulation Studies

Several simulation studies were conducted. In CAT different stopping rules can be applied (Wainer et al., 2000). In variable length CAT, items are being administered until the measurement error is below a certain threshold, whereas in fixed length CAT, a fixed number of items is being administered. Fixed length CAT is often applied when the test has to meet a number of specifications with respect to content, or other attributes. The first simulation study is designed to compare the performances of the algorithm with and without empirical prior for a variable length CAT. In the second study, the focus is on the impact of empirical prior information for fixed length CAT of different lengths. In the third study, different settings are evaluated for a short test of length equal to 10. In particular, the estimation results are compared for the MCMC CAT proposed by the authors, CAT without empirical prior, and CAT with only empirical ability initialization. Finally, the issue of the algorithm convergence is taken into account.

Prior in Use: a Comparison in a Variable Length CAT

The purpose of the first simulation study is to show the potentiality of the empirical prior in reducing the test length within the Gibbs sampler scheme. To this aim, two

different CAT designs are compared: the first one follows the common practice of initializing the ability at zero and assuming a standard normal as a prior for the ability distribution, while the second one adopts an empirical prior both in the initialization and in the ability estimation, as shown in the previous section. For simplicity of description, the former approach is denominated *standard* while the latter is called *fully empirical*. In both cases, item selection is conducted by using the maximum-information criterion.

In the study, an item bank of 500 items is employed, with item parameters sampled as $\alpha_j \sim U(0.7;2)$ and $\delta_j \sim U(-4;4)$, for *j=1,...,k* . When the fully empirical approach is adopted, the linear relation $\theta = 0.2 + 0.7X + \varepsilon$ with $\varepsilon \sim N(0;0.3)$ is assumed between the ability θ and a single covariate *X*. Responses are simulated for different levels of ability from -3 to 3 according to model (1). Given the true θ, the *X*-values are simulated for each replication from $(\theta - 0.2 - \varepsilon)/0.7$.

The Gibbs sampler with a chain length of 5000 iterations and burn-in of 500 is employed for the ability estimation. The output consists of the mean and standard deviations sampled from the posterior distribution of ability. The choice of the chain length and the number to discard iterations is motivated by the convergence study described in the end of this section. All chains showed fast convergence and good mixing properties. In order to compare the efficiency of the two different approaches, especially in terms of number of items needed to complete the CAT algorithm, the stopping rule is set to a test information above 10 at the current ability estimate.

For all ability levels within each approach, a number of 100 replications have been conducted. The mean number of items needed to complete the CAT over replications has been recorded together with the corresponding standard deviation (s.d. items). With respect to ability, the expected posterior estimate, bias and standard deviation (s.d.) are reported. The results of the simulations are shown in Table 1.

[INSERT TABLE 1 ABOUT HERE]

As can be seen from the mean test length, the fully empirical solution is able to reduce the mean number of items needed respect to the standard one, and the two approaches are comparable only for ability levels close to zero. By using empirical information, CAT tests are shortened and, as a consequence, item overexposure is also reduced. Furthermore, the recovery of the true ability is more precise in the fully empirical approach in terms of both bias and estimate stability, which can be assessed by looking at the standard deviation (s.d.). In fact, the standard solution fails to recover the ability levels when deviating from $\theta=0$.

Prior in Use: a Comparison with Different Test Lengths

In the second simulation study, the same item pool and conditions of the previous study are maintained, but a fixed length CAT is used. In fact, in order to get results for tests consisting of different numbers of items, the CAT stopping rule is defined fixing the test length at 10, 15 or 20 items. As usual, a number of 100 replications have been conducted in the simulation.

Besides the expected a posterior estimate and the standard deviation, also the average bias and the root mean square errors (RMSE) have been calculated. Table 2 provides the results of the simulation study in case of a short test consisting of 10 items.

[INSERT TABLE 2 ABOUT HERE]

As can be easily noticed, compared with the standard version of CAT, the parameter recovery of empirical CAT is more accurate in terms of RMSE, and the estimates are more stable because the are associated with lower standard deviations, especially when deviating from $\theta=0$. Bias is comparable between the two approaches.

Table 3 and 4 show the results of the simulations conducted for adaptive tests of 15 and 20 items, respectively.

[INSERT TABLE 3 ABOUT HERE]

[INSERT TABLE 4 ABOUT HERE]

Due to the increasing number of items, standard CAT becomes more precise, and the two approaches become comparable, even if for T=15 the fully empirical approach maintains lower standard deviation and RMSE, especially for extreme abilities. The comparison of true and simulated values for central abilities suggests that there are no considerable differences in reproducing the ability values between the two approaches.

From this simulation study it can be learned that the introduction of an informative prior leads to an improvement of measurement precision in the individual ability assessment. This improvement becomes very evident for short tests and when shifting to boundary ability values. This cannot be generalized to the case of longer test (e.g., more than 20 items): when the test length increases, the prior distribution lacks in strength and the two solutions become more and more similar.

Introduction of Prior Information at Different Levels

According to the findings of the previous study, the use of prior information in CAT shows its maximum effectiveness in case of short tests. In this simulation study, the focus is on the comparison of different levels of prior information for a target test consisting of 10 items. Results of Table 2 regarding fully empirical and standard CAT are compared to an intermediate solution, named *empirical initialization*, where empirical information is used only in the initialization of the ability estimate. Table 5 illustrates the results of the simulation.

[INSERT TABLE 5 ABOUT HERE]

The empirical initialization CAT shows an intermediate behavior with respect to the other two approaches. This approach obtains standard deviations which are more comparable to the fully empirical approach than the standard one. On the other hand, estimates are biased, even more seriously than the standard solution especially for $\theta=-3$ and $\theta=3$. As can be clearly seen in Figure 1, which shows the RMSEs across the ability true values for the three approaches, the empirical initialization solution performs better than the standard approach but worse than the fully empirical one.

[INSERT FIGURE 1 ABOUT HERE]

For the fully empirical solution, the RMSE curve is always below or at most close to the curves associated with the standard and the empirical initialization approaches. The difference in precision is particularly significant for ability levels in the tails of the distribution.

A Note on the Algorithm Convergence

One of the most critical issues in MCMC estimation is assessing the convergence of the algorithm. A large number of researchers have approached the problem turning out with different, sometimes conflicting solutions (for a review, see Cowles and Carlin, 1996). When simulating a MCMC chain, the first thing is to check the trace plot of the simulated random draws. Even if convergence cannot be ensured by simply looking at the iteration history, a clearly critical situation of non-convergence can be detected immediately. After computing the posterior mean and the standard deviation, a measure of the standard error of estimate should be calculated. As suggested in Gelman, Carlin, Stern and Rubin (2004, chap. 10), an approximate measure of the accuracy of the sample mean estimate is the standard deviation divided by the square root of the number of simulations, which is nothing but the posterior deviance. Moreover, an estimate of the Monte Carlo standard error should be computed. One possibility is to calculate the square root of the spectral density variance estimate divided by the number of actual iterations (time-series estimate), as proposed by Geweke (1992) in order to provide an estimate of the asymptotic standard error. As a rule of thumb, the estimated Monte Carlo error should be less than 5% of the standard deviation.

In order to decide the necessary number of iterations for obtaining an acceptable accuracy, a study has been conducted by simulating single chains. In particular, the simulation design of the second study is drawn on in the case of ability $\theta=0$ and test length $T=10$. The purpose of this convergence study is to evaluate the accuracy of the posterior mean in the simulations by using different number of iterations (1000, 2000, 5000 and 10000). Table 6 shows the results both for the fully empirical and the standard approaches.

[INSERT TABLE 6 ABOUT HERE]

The number of iterations is specified in the first column, while the number to discard iterations (burn-in phase) is contained in column 2. Besides the posterior mean and the standard deviation, an estimate of the Monte Carlo error (MC error) is reported, which has been calculated by using the R package BOA.

One single replication, depending on the number of iterations in the chain, took only few seconds to complete (from 1 to 7 seconds) on a 2.66 GHz Intel Core2 Quad desktop. The simulations conducted by using 1000 iterations do not satisfy the accuracy condition of MC error less than 5% of the standard deviation, while the solution with 2000 iterations slightly satisfies it. On the other hand, running 5000 or 10000 iterations turns out with MC errors significantly lower than the 5% of standard deviation and are thereby considered a good standard of accuracy.

As a consequence of these results, the adopted number of iterations was settled to 5000. For each replication of the simulation studies described in the section, the MC error was assessed to be less than 5% of standard deviation. All chains showed fast convergence and good mixing properties. The chosen chain length represents a good compromise between the estimate accuracy and the time needed to complete the algorithm. Figure 2 shows the trace plot of the simulation with 5000 iterations when prior information is included.

[INSERT FIGURE 2 ABOUT HERE]

Clearly, the plot shows a random fluctuation of the sample values around the mean. The absence of autocorrelation (at least at a lag higher than 5) is confirmed by the autocorrelation plot reported in Figure 3.

[INSERT FIGURE 3 ABOUT HERE]

Usually, one of the main drawbacks of MCMC is the time consuming and slow convergence of the algorithm; however, adopting the above mentioned features for the chain, the simulation represents a good compromise between speed and accuracy. Of course, we should also mention that the model implemented is rather simple, because it is a unidimensional model for binary indicators. Probably, the extension of the algorithm to more complicated model, as multidimensional models, would come out with a slower convergence.

Empirical Example

The MCMC CAT described in previous sections provides a useful strategy for improving the quality of measurement precision and has a good potentiality in real applications of adaptive testing. In order to show the effectiveness of the method in practice, a case study was chosen in the field of intelligence testing. Data regarding a computer adaptive intelligence tests for personnel selection, the Connector Ability (Maij- de Meij, Schakel, Smid, Verstappen, and Jaganjac, 2008) were available. The complete test consists of three different subscales: Number series, Figure series, and Raven's matrices. This test has been developed for applications in the area of HRM, for example for job selection or for career development. In our example, the focus is on the role of the Raven's matrices (RM) ability as predictor for the performance in the

Number series (NS) test. In Matteucci, Mignani and Veldkamp (2009) a Bayesian procedure for concurrent estimation of both the person parameters, item parameters, and the empirical prior on the person parameters has been described. Following this approach, the relation between the RM and NS subscales was estimated, resulting in the following empirical prior distribution

$$\theta \mid X_1 \sim N(-0.243 + 0.394 X_1 : 0.414), \tag{13}$$

where θ is the ability in the NS subscale and X_1 is the ability in the RM subscale. Given the standard normal scale of ability, the estimated regression coefficient $\hat{\beta}_1 = 0.394$ shows a positive and moderate effect of the RM ability on the performance in the NS subscale.

To determine whether the introduction of the prior distribution (13) is effective in this case study, an adaptive version of the NS test is simulated for a group of 660 real examinees. The full item bank consisted of 499 calibrated Number series items. Some descriptive statistics on the item parameters included in the item bank are shown in Table 7.

[INSERT TABLE 7 ABOUT HERE]

Discrimination parameters vary in the interval [0.180; 1.470], with a mean value around 0.7, while difficulty parameters are included in the range [-2.290; 2.300] with a mean of -0.4. Discrimination and difficulty parameters are treated as known in the adaptive test administration, while the abilities previously estimated in the NS test for the 660 examinees are considered as true abilities in the simulation. For each examinee, the adaptive test is replicated 10 times, and the ability estimation is performed by using

5000 MCMC iterations with the usual burn-in of length 500. The algorithm stopping rule is established as test information at the current ability estimate above 10, which is the equivalent of a standard error less or equal to 0.32 for a population with a standard normal ability distribution. For each candidate, the mean number of submitted items over replications is recorded. As usual, the three MCMC CAT approaches (fully empirical, empirical initialization and standard) are compared. The simulation results for the three different approaches are shown in Table 8.

[INSERT TABLE 8 ABOUT HERE]

Before looking at the mean number of items needed in CAT, a remark on the setting of the item parameters with respect to the examinees being simulated is needed. As can be observed from the first column of Table 8, 16 equal spaced intervals of ability from -2.4 to 2.4 are constructed in order to present aggregated results. The second column shows the number of items with difficulty parameters falling in each interval while the third column contains the number of simulees in each ability range. Three items in the bank have difficulty parameters in the range [-2.4; -2.1], but no examinees in the same ability range were simulated. Eight items in the bank had difficulty parameters above 1.5, where also no examinees were simulated.

With regards to low ability intervals, the fully empirical solution performs better than the others, with a mean number of items needed in test administration sensibly lower while the standard solution presents the worst results. While approaching intermediate ability levels, the number of items needed in the simulation reduces and the three approaches show similar performances, even if the empirical initialization and the standard solutions still seem the weakest. For high ability intervals, the fully empirical

solution performed better than the empirical initialization and the standard CAT. The results of the MCMC CAT applied to a real item bank regarding intelligence tests show that the inclusion of empirical prior information, especially in the estimation of the candidate's ability, is effective in reducing the test length for the same test information level. The application also demonstrates that the quality of results depends much on the quality of the item bank itself in terms of size and item properties.

Discussion

The study focused on increased efficiency of computerized adaptive testing. It also introduced the problem of ability estimation in computerized adaptive testing under particular situations of uncertainty about the candidate's level of proficiency. Examples are CAT consisting of a small number of items or candidates with latent ability far from average. The introduction of prior information in the algorithm resulted in more accurate ability estimates or, analogously, in a reduction of the test length at a given level of precision, and strengthened the applicability of CAT for extreme ability levels and for short CATs. This approach is developed within the MCMC methods, particularly adopting the Gibbs sampler to integrate likelihood with empirical prior information about the candidate. The use of MCMC in ability estimation allows to overcome both the technical limitations of the Gaussian quadrature in estimation and the problem of non-mixed patterns in CAT.

The main purpose of the study was to compare the precision of ability estimates among different specifications and uses of prior distributions. Therefore, a fixed-length termination rule was applied in the simulation studies more intensively. However, a

study was conducted also adopting a variable-length termination rule which was used to compare the number of items needed in order to obtain the same precision of measurement.

The findings of simulation studies suggest that the introduction of informative priors is effective in improving the accuracy of ability estimates, especially when dealing with rather short tests and when the ability is far from zero. In particular, the measurement precision is improved when empirical priors are introduced both to initialize and to estimate ability. The use of empirical information is highly recommended with rather short tests, where the standard approaches based on a standard normal prior fail to reproduce stable ability estimates. When using a variable length CAT, it was demonstrated that the test could be shortened and, as a consequence, the item overexposure could be reduced as well.

Despite the great availability of background variables concerning the individuals, the quality of information remains a fundamental issue. The usefulness of the described approach depends highly on the predictive capability of the collateral variables.

In many applications in psychological measurement, it would be acceptable to use background variables to increase measurement precision. For example, in personnel selection, companies are just interested in selecting the best candidates based, and test efficiency is a major issue. Besides, adaptive tests are becoming more and more used in the area of medicine, where tailored tests are proposed to patients in order to infer their physical and mental health. Covariates about patients such as psychological status can be introduced as empirical prior information in these settings. In many medical, clinical or diagnostic applications, reducing the burden of test administration for both patients and doctors/psychologists is an important topic. In educational applications, it might be an issue to use collateral information. In high-stakes tests like exams or admission tests,

the use of collateral information would not be accepted. However, when such problems of fairness arise and empirical information cannot be used in the ability estimation, an initial inference which is as close as possible to the true ability value is recommended, i.e., an empirical CAT initialization is desirable. This approach solves the issue of overexposure of the first item, observed in CAT combining a fixed initialization (e.g., ability equal to zero) and maximum-information criterion for item selection. Because good performances of MCMC CAT have been recorded when background variables are used both in the initialization and in the ability estimation, another possibility would be to exclude the use of prior information only from the final ability estimation in order to prevent the method from potential criticism due to fairness issues.

MCMC CAT might also provide other advantages which can be used in further research. In the current study, the item parameters were assumed to be fixed and known. However, these parameters result from a calibration study and have been estimated with uncertainty. In a Bayesian estimation procedure, this uncertainty can be taken into account. In this way, unrealistically high precision of ability estimates due to the assumption of known item parameters might be dealt with in future applications. Moreover, the Gibbs sampler represents a flexible tool which can be implemented for more complex IRT models and with different specifications for the prior distribution, depending on the available empirical covariates.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269.

Ariel, A., van der linden, W. J., & Veldkamp, B. P. (2006). A strategy for optimizing item pool management. *Journal of Educational Measurement, 43,* 85-96.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. New York: Oxford University Press.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis.* London: Arnold Publishers.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66,* 541-562.

Belov, D. I., & Armstrong, R. D., (2009). Direct and inverse problems of item pool design for computerized adaptive testing. *Educational and Psychological Measurement, 69,* 533-547.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association, 91,* 883-904.

Fox, J. -P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271-288.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis, 2nd edition*. Boca Raton, Florida: Chapman and Hall/CRC.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721-741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J.M. Bernardo, J. Berger, A.P. Dawid & A.F.M. Smith (Eds.), *Bayesian statistics 4* (pp. 169-193). Oxford,U.K.: Oxford University Press.

Gialluca, K. A., & Weiss, D. J. (1979). *Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement*. Research Report 79-6. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Guyer, R. D. (2008). *Effect of Early Misfit in Computerized Adaptive Testing on the Recovery of Theta*. University of Minnesota (MN): Unpublished doctoral dissertation.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York: Springer-Verlag.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph, 7*.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139-183). New York: Harper and Row.

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

Matteucci, M., & Veldkamp, B. P. (2011). Including empirical prior information in test administration. In B. Fichet, D. Piccolo, R. Verde, & M. Vichi (Eds.), *Classification and multivariate analysis for complex data structures* (pp. 171-179). Berlin Heidelberg: Springer-Verlag.

Matteucci, M., Mignani, S., & Veldkamp, B. P. (2009). Issues on item response theory modelling. In M. Bini, P. Monari, D. Piccolo & L. Salmaso (Eds.), *Statistical methods for the evaluation of educational services and quality of products* (pp. 29-45). Berlin Heidelberg: Springer-Verlag.

Maij- de Meij, A. M., Schakel, L., Smid, N., Verstappen, N., & Jaganjac, A. (2008). *Connector Ability; Professional Manual*. Utrecht, The Netherlands: PiCompany B.V.

Natesan, P., Limbers, C., & Varni, J. W. (2010). Bayesian estimation of graded response multilevel models using Gibbs sampling: formulation and illustration. *Educational and Psychological Measurement*, *70*, 420-439.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* . Research Report 69-92. Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351-356.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, *67*, 899-919.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement, 68*, 413-430.

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* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

The MathWorks Inc. (2005). MATLAB 7.1 [Computer program]. Natick, MA: The MathWorks, Inc.

van der Linden, W. J. (1999). Empirical initialization of the trait estimation in adaptive testing. *Applied Psychological Measurement, 23,* 21-29.

van der Linden, W. J. (2005). *Linear models for optimal test design.* New York: Springer Verlag.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33,* 5-20.

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer Academic Publishers.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3-30). New York: Springer.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure rates in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29,* 273-291.

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32,* 398-417.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., & Steinberg, L. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika, 56,* 589-600.

Zwinderman, A. H. (1997). Response models with manifest predictors. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 245-256). New York: Springer-Verlag.

TABLE 1

Final test length and ability parameter recovery for fully empirical and standard solutions.

True θ	Fully empirical					Standard				
	Mean n. items	s.d. items	$\hat{\theta}$	Bias	s.d.	Mean n. items	s.d. items	$\hat{\theta}$	Bias	s.d
-3	9.91	1.84	-3.04	-0.04	0.24	12.49	2.85	-2.79	0.21	0.31
-2.5	6.69	1.14	-2.50	0.00	0.23	9.42	1.84	-2.33	0.17	0.31
-2	5.45	0.64	-1.99	0.01	0.25	7.65	0.98	-1.89	0.11	0.30
-1.5	5.11	0.40	-1.54	-0.04	0.29	6.71	0.74	-1.41	0.09	0.26
-1	5.17	0.45	-1.02	-0.02	0.28	6.16	0.53	-0.95	0.05	0.28
-0.5	5.46	0.87	-0.49	0.01	0.23	5.77	0.75	-0.46	0.04	0.27
0	5.24	0.45	0.04	0.04	0.26	5.29	0.56	0.02	0.02	0.25
0.5	5.28	0.55	0.47	-0.03	0.25	5.32	0.63	0.46	-0.04	0.25
1	5.17	0.43	1.03	0.03	0.26	5.58	0.83	0.84	-0.16	0.33
1.5	5.18	0.41	1.52	0.02	0.28	6.38	0.84	1.40	-0.10	0.31
2	5.49	0.64	2.03	0.03	0.23	7.64	1.03	1.89	-0.11	0.31
2.5	7.05	1.47	2.49	-0.01	0.28	9.72	1.84	2.37	-0.13	0.31
3	10.15	2.11	3.05	0.05	0.30	12.51	2.59	2.77	-0.23	0.33

TABLE 2

Ability parameter recovery for fully empirical and standard solutions (T=10).

True θ	Fully empirical				Standard			
	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE
-3	-3.06	0.25	-0.06	0.25	-2.94	0.36	0.06	0.36
-2.5	-2.57	0.25	-0.07	0.26	-2.45	0.29	0.05	0.29
-2	-2.01	0.22	-0.01	0.22	-1.93	0.27	0.07	0.28
-1.5	-1.47	0.18	0.03	0.18	-1.44	0.24	0.06	0.25
-1	-0.98	0.22	0.02	0.22	-0.97	0.25	0.03	0.25
-0.5	-0.52	0.20	-0.02	0.20	-0.45	0.19	0.05	0.20
0	-0.01	0.22	-0.01	0.22	-0.01	0.24	-0.01	0.24
0.5	0.52	0.18	0.02	0.18	0.49	0.22	-0.01	0.22
1	1.00	0.19	0.00	0.19	0.94	0.24	-0.06	0.25
1.5	1.51	0.21	0.01	0.21	1.46	0.20	-0.04	0.20
2	2.06	0.24	0.06	0.25	1.98	0.28	-0.02	0.28
2.5	2.60	0.26	0.10	0.27	2.47	0.30	-0.03	0.30
3	3.05	0.27	0.05	0.27	2.94	0.33	-0.06	0.34

TABLE 3

Ability parameter recovery for fully empirical and standard solutions (T=15).

True θ	Fully empirical				Standard			
	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE
-3	-3.05	0.21	-0.05	0.22	-2.91	0.26	0.09	0.28
-2.5	-2.54	0.21	-0.04	0.21	-2.46	0.22	0.04	0.22
-2	-2.03	0.18	-0.03	0.18	-1.96	0.21	0.04	0.21
-1.5	-1.51	0.15	-0.01	0.15	-1.45	0.19	0.05	0.20
-1	-1.00	0.14	0.00	0.14	-1.01	0.20	-0.01	0.20
-0.5	-0.48	0.17	0.02	0.17	-0.46	0.15	0.04	0.15
0	0.01	0.18	0.01	0.18	0.01	0.17	0.01	0.16
0.5	0.48	0.18	-0.02	0.18	0.47	0.17	-0.03	0.17
1	0.98	0.17	-0.02	0.17	1.01	0.14	0.01	0.14
1.5	1.52	0.17	0.02	0.18	1.48	0.18	-0.02	0.18
2	2.05	0.20	0.05	0.21	2.00	0.23	0.00	0.23
2.5	2.58	0.23	0.08	0.24	2.50	0.29	0.00	0.29
3	3.08	0.28	0.08	0.29	2.96	0.30	-0.04	0.31

TABLE 4

Ability parameter recovery for fully empirical and standard solutions (T=20).

True θ	Fully empirical				Standard			
	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE
-3	-3.07	0.20	-0.07	0.21	-2.96	0.23	0.04	0.23
-2.5	-2.53	0.21	-0.03	0.21	-2.50	0.19	0.00	0.19
-2	-1.98	0.17	0.02	0.17	-1.97	0.17	0.03	0.17
-1.5	-1.51	0.15	-0.01	0.15	-1.47	0.13	0.03	0.14
-1	-0.96	0.14	0.04	0.15	-0.96	0.16	0.04	0.16
-0.5	-0.52	0.15	-0.02	0.15	-0.46	0.16	0.04	0.16
0	-0.02	0.15	-0.02	0.16	0.02	0.14	0.02	0.14
0.5	0.49	0.14	-0.01	0.14	0.49	0.16	-0.01	0.16
1	1.01	0.16	0.01	0.16	1.02	0.15	0.02	0.15
1.5	1.52	0.14	0.02	0.14	1.45	0.16	-0.05	0.16
2	2.06	0.18	0.06	0.19	2.00	0.17	0.00	0.17
2.5	2.58	0.19	0.08	0.20	2.53	0.22	0.03	0.23
3	3.06	0.24	0.06	0.24	2.98	0.21	-0.02	0.21

TABLE 5

Ability parameter recovery for fully empirical, empirical initialization and standard solutions (T=10).

True θ	Fully empirical				Empirical initialization				Standard			
	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE
-3	-3.06	0.25	-0.06	0.25	-2.92	0.26	0.08	0.27	-2.94	0.36	0.06	0.36
-2.5	-2.57	0.25	-0.07	0.26	-2.45	0.25	0.05	0.25	-2.45	0.29	0.05	0.29
-2	-2.01	0.22	-0.01	0.22	-1.92	0.25	0.08	0.26	-1.93	0.27	0.07	0.28
-1.5	-1.47	0.18	0.03	0.18	-1.44	0.21	0.06	0.22	-1.44	0.24	0.06	0.25
-1	-0.98	0.22	0.02	0.22	-0.91	0.21	0.09	0.23	-0.97	0.25	0.03	0.25
-0.5	-0.52	0.20	-0.02	0.20	-0.46	0.20	0.04	0.21	-0.45	0.19	0.05	0.20
0	-0.01	0.22	-0.01	0.22	-0.02	0.26	-0.02	0.26	-0.01	0.24	-0.01	0.24
0.5	0.52	0.18	0.02	0.18	0.48	0.22	-0.02	0.22	0.49	0.22	-0.01	0.22
1	1.00	0.19	0.00	0.19	1.01	0.23	0.01	0.23	0.94	0.24	-0.06	0.25
1.5	1.51	0.21	0.01	0.21	1.44	0.20	-0.06	0.21	1.46	0.20	-0.04	0.20
2	2.06	0.24	0.06	0.25	1.95	0.23	-0.05	0.23	1.98	0.28	-0.02	0.28
2.5	2.60	0.26	0.10	0.27	2.46	0.25	-0.04	0.25	2.47	0.30	-0.03	0.30
3	3.05	0.27	0.05	0.27	2.88	0.30	-0.12	0.33	2.94	0.33	-0.06	0.34

TABLE 6

Estimated accuracy of simulation across different number of iterations.

N. iter	Burn-in	Fully empirical				Standard			
		$\hat{\theta}$	s.d.	5% s.d.	MC error	$\hat{\theta}$	s.d.	5% s.d.	MC error
1000	100	-0.119	0.393	0.020	0.023	0.070	0.422	0.021	0.025
2000	200	-0.101	0.391	0.020	0.013	-0.048	0.417	0.021	0.018
5000	500	0.303	0.411	0.021	0.011	-0.135	0.427	0.021	0.008
10000	1000	0.048	0.373	0.019	0.006	-0.107	0.410	0.021	0.008

TABLE 7

Descriptive statistics on the item parameters included the item bank.

	Discrimination parameters	Difficulty parameters
Mean	0.745	-0.411
Median	0.727	-0.410
Standard deviation	0.309	0.748
Minimum	0.180	-2.290
Maximum	1.470	2.300

TABLE 8

Results on the mean number of items needed in CAT simulation.

			Fully empirical		Empirical initialization		Standard	
Ability range	N. items with difficulty parameter in the range	N. examinees in the range	Mean n. items	s.d.	Mean n. items	s.d.	Mean n. items	s.d.
-2.4 -| -2.1	3	0	-	-	-	-	-	-
-2.1 -| -1.8	12	2	13.750	0.212	15.350	2.616	16.150	0.636
-1.8 -| -1.5	20	8	11.713	0.732	13.325	1.029	13.638	0.905
-1.5 -| -1.2	34	42	10.655	0.681	11.210	0.854	11.569	0.833
-1.2 -| -0.9	67	54	9.620	0.434	10.174	1.283	10.006	0.511
-0.9 -| -0.6	64	97	9.136	0.154	9.344	0.226	9.481	0.951
-0.6 -| -0.3	78	132	9.085	0.107	9.239	0.172	9.198	0.149
-0.3 -| 0.0	78	123	9.322	0.259	9.533	0.369	9.498	0.293
0.0 -| 0.3	61	86	9.920	0.420	10.303	0.585	10.307	0.567
0.3 -| 0.6	44	61	11.290	0.756	12.077	1.133	11.874	0.997
0.6 -| 0.9	21	30	13.600	1.642	15.217	1.642	15.540	1.835
0.9 -| 1.2	7	16	17.681	2.119	21.244	2.589	20.431	3.034
1.2 -| 1.5	1	9	24.356	3.207	29.622	4.757	29.611	2.930
1.5 -| 1.8	3	0	35.843	6.097	44.871	6.054	46.129	5.559
1.8 -| 2.1	5	0	-	-	-	-	-	-
2.1 -| 2.4	1	0	-	-	-	-	-	-

FIGURE 1

Root mean square error (RMSE) for the three different approaches (fully empirical, empirical initialization and standard) when the test consists of 10 items.

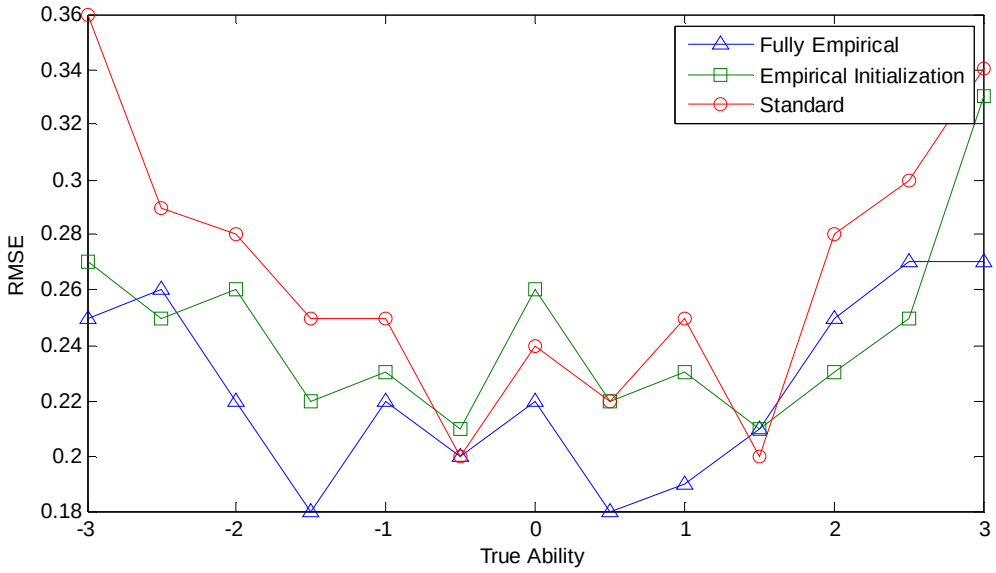

FIGURE 2

Trace plot of a single chain, in the case of T=10 and empirical information introduced.

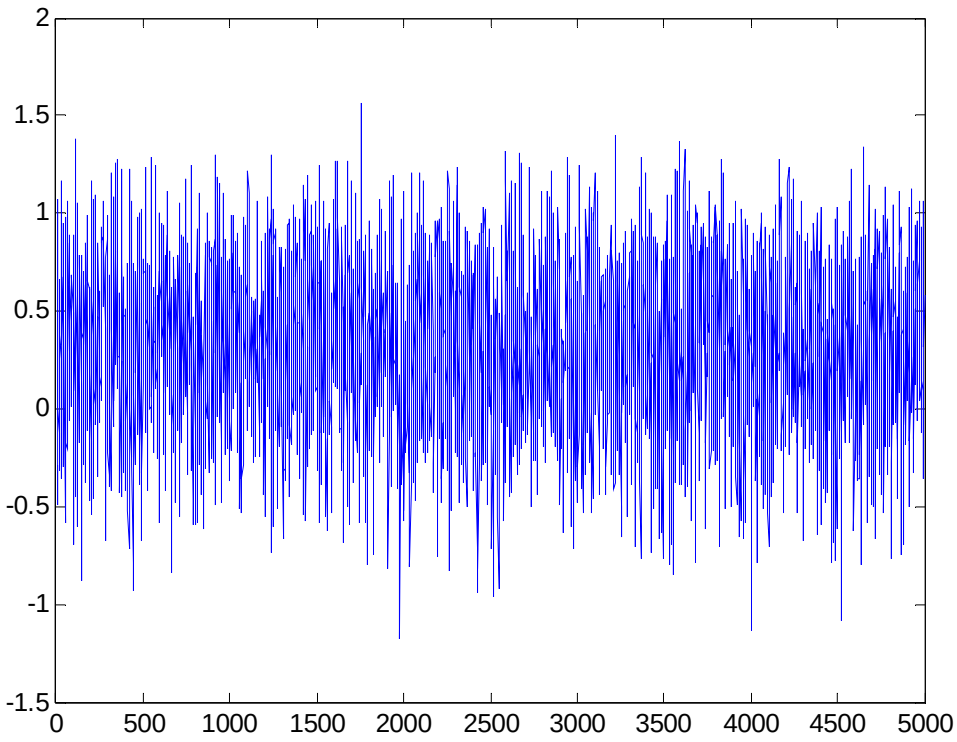

FIGURE 3

Autocorrelation plot.