Stat Methods Appl
DOI 10.1007/s10260-012-0216-1

On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency

Mariagiulia Matteucci · Bernard P. Veldkamp

Accepted: 1 October 2012

Abstract The paper deals with the introduction of empirical prior information in the estimation of candidate's ability within computerized adaptive testing (CAT). CAT is generally applied 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 proved 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 examples, the methods were applied to CAT for intelligence testing in the area of personnel selection and to educational measurement. Other promising applications would be in the medical world, where testing efficiency is of paramount importance as well.

Keywords Adaptive testing · Empirical prior information · Gibbs sampler · Measurement precision

1 Introduction

In recent years, we have seen a rapid development of computer-based testing in the field of educational and psychological measurement, especially in adaptive testing.

M. Matteucci (✉)
Department of Statistical Sciences, University of Bologna, via Belle Arti, 41 40126 Bologna, Italy
e-mail: m.matteucci@unibo.it

B. P. Veldkamp
Research Center for Examination and Certification, University of Twente, P.O. Box 217 7500 AE, Enschede, Netherlands
e-mail: B.P.Veldkamp@gw.utwente.nl

Published online: 17 October 2012

The basic idea of computerized adaptive testing (CAT) is to adapt the item difficulty to the estimated ability level of the candidate, so that the test becomes person-tailored. In this way, the test length is shortened with respect to standard test administration, such as paper and pencil tests, keeping the same measurement accuracy.

Despite its increasing use, CAT has a number of operational problems like item pool maintenance (Ariel et al. 2004, 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; Barrada et al. 2009), item parameter uncertainty (De Jong et al. 2009; Veldkamp 2012), and recovery from unforced errors during the beginning of CAT (Guyer 2008).

From a statistical point of view, one of the most important issues is the choice of the ability estimator. The maximum likelihood (ML) estimator, as well as the Warm (1989) weighted likelihood estimator (WLE), does not produce finite estimates for perfect response patterns (all endorsed or all not endorsed items). On the other hand, Bayesian estimators, such as expected a posteriori (EAP), maximum a posteriori (MAP), and Bayes modal (BM), depend on the choice of the prior distribution for the ability. Especially in the beginning of the test, or when the test length is fixed to be short, the choice of the prior distribution is crucial. In fact, a wrongly located prior may lead the ability estimate far from the true value which will be hardly recovered in the subsequent few steps (see e.g. van der Linden and Pashley 2010).

In order to overcome these limitations, in this paper we propose a fully Bayesian estimation of the ability via Markov chain Monte Carlo (MCMC) methods with empirical prior information about the candidates. A Bayesian approach is chosen to avoid the problem of ML estimation for perfect response patterns, while collateral information is introduced to guide the prior distribution especially in the initial ability estimation. Nowadays, collateral information about the candidates is regularly collected during assessments and its inclusion in CAT administration is straightforward. 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 the initialization and in the ability estimation. Furthermore, the paper describes how the empirical prior can be integrated in the estimation process within the Gibbs sampler scheme. By using simulated data, we show how efficiency of CAT is improved under different settings, e.g. fixed and variable length, and we compare our proposal to alternatives where prior information is introduced at different levels. To further prove the effectiveness of the approach, an empirical application with personnel selection test data is discussed.

The paper is organized as follows. Section 2 reviews the main features of computerized adaptive testing, with a particular reference to the current state of methodological and computational aspects involved in the CAT phases. In Sect. 3, our proposal is introduced by first motivating the introduction of empirical information in CAT, and then developing a Gibbs sampler scheme for ability estimation with an empirical prior distribution. In Sect. 4, the advantages of our approach 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 evaluated in case empirical priors are introduced instead of standard priors. In Sect. 5, the results of two empirical CAT applications are

presented in the context of intelligence testing for personnel selection and of educational testing. Finally, Sect. 6 concludes the paper with a discussion.

2 Computerized adaptive testing

The theoretical framework of computerized adaptive testing (CAT) was developed since the early 1970s (Lord 1970; Owen 1969, 1975) and it is now formalized (see, e.g., van der Linden and Glas 2000, 2010; Wainer et al. 2000). Unlike linear testing, where the same set of items is submitted to a sample of individuals, CAT is based on adaptive item selection and administration in analogy to an oral examination. 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 on the presence of an item pool containing items with particular psychometric properties. Therefore, item response theory (IRT) models are employed (see, e.g., Lord and Novick 1968). IRT models describe the mathematical function linking the individual response probability to a set of item parameters, denoting the item psychometric characteristics, and the individual ability. After a test administration, a particular IRT model is chosen to estimate the item parameters based on data nature and fit. Once the item parameters have been estimated with sufficient precision, items with target features are included in the item pool. Moreover, during CAT administration, the response process is assumed to follow the chosen IRT model. 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 $j = 1, \ldots, k$ items, as a function of ability and item parameters, as follows

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

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. The model is identified by fixing the mean ability equal to zero and its variance equal to one. Depending on the data characteristics, other models such as multidimensional models are possible

and have been employed in CAT (Segall 1996; Veldkamp and van der Linden 2002; Reckase 2009, chapter 10).

2.1 The phases of CAT

Once the items have been calibrated according to an IRT model, the item parameters are typically treated as known. Afterwards, CAT works with the following different steps:

1. ability initialization
2. item selection
3. item administration
4. ability estimate update.

Given J calibrated items in the pool, indexed by $j = 1, \ldots, J$, the rank of selected items is denoted as $k = 1, \ldots, K$. Hence, when choosing the kth 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 $R_k = \{1, \ldots, J\} \setminus S_{k-1}$ 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.

The first phase deals with the ability initialization, where an initial proficiency level of the candidate is defined. A typical choice is to initialize the ability to its expected mean value, i.e. $\theta_0 = 0$. An alternative to a fixed initialization is a random one. Otherwise, when information about the candidate can be inferred from a set of covariates, an empirical initialization is possible as well (van der Linden 1999). 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). Clearly, the efficiency of CAT is strongly affected by the ability initialization when a short number of items is submitted.

In order to proceed with the item selection (Step 2), various criteria have been proposed in literature. The most popular method is the maximum-information criterion (Birnbaum 1968). When selecting the kth item, the method works by choosing the item which maximizes Fisher's expected information function at the current ability value $\theta = \hat{\theta}_{k-1}$, as follows

$$j_k \equiv \arg\max_j \{I_j(\hat{\theta}_{k-1});\ j \in R_k\}. \tag{2}$$

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

$$I_j(\hat{\theta}_{k-1}) = \alpha_j^2 \frac{[(2\pi)^{-1/2} \exp(-\eta_j^2/2)]^2}{\Phi(\eta_j)[1 - \Phi(\eta_j)]}, \tag{3}$$

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 also in linear testing, where the most informative items are included in operational tests. Nevertheless, within an adaptive environment,

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. Alternative item selection rules are based on Kullback–Leibler information (Chang and Ying 1996; Lehman and Casella 1998, Section 1.7), likelihood-weighted information criterion (Veerkamp and Berger 1997), and Bayesian criteria (Owen 1969, 1975; van der Linden 1998; Veldkamp 2010). For a review on item selection criteria, see van der Linden and Glas (2007), van der Linden and Pashley (2010).

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. 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. After the administration of *k-1* items, the ML estimator is defined as

$$\hat{\theta}_{k-1}^{ML} \equiv \arg\max_{\theta}\{L(\theta|y_{j_1}\ldots y_{j_{k-1}}) : \theta \in (-\infty, +\infty)\}. \tag{4}$$

An alternative is the weighted likelihood estimator (WLE) proposed by Warm (1989). The maximum of the likelihood function is not always unique for particular IRT models, such as the three-parameter logistic model. In linear testing, the ML estimator owns the properties of correctness and asymptotic efficiency. In CAT, the ML estimators cannot rely on asympotic properties. Small-sample properties depend on the item distribution within the pool and the item selection criterion (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. Bayesian methods are based on the posterior distribution of the ability θ. The full posterior distribution of θ can be used as ability estimator, and its variance as a measure of uncertainty about θ. However, updating the posterior distribution after the responses to $k-1$ items, the formulation of the likelihood based on IRT models such as model (1) does not allow to choose a prior from a conjugate family. A restricted Bayesian approach was proposed by Owen (1969), assuming a normal prior distribution for θ and replacing the updated posterior distribution by a normal distribution with the same parameters as the true posterior. However, Owen's proposal is based on an approximation which, even though it was proved to be reliable, may not fit all cases. Among different Bayesian point estimators, maximum a posteriori (MAP) and expected a posteriori (EAP) estimators proposed by Mislevy (1986) and Bock and Mislevy (1988) respectively, are most frequently used in CAT:

$$\hat{\theta}_{k-1}^{MAP} \equiv \arg\max_{\theta}\{g(\theta|y_{j_1}\ldots y_{j_{k-1}}) : \theta \in (-\infty, +\infty)\}, \tag{5}$$

$$\hat{\theta}_{k-1}^{EAP} \equiv \int \theta g(\theta|y_{j_1}\ldots y_{j_{k-1}})d\theta. \tag{6}$$

For a uniform prior, the MAP estimator is equivalent to the ML estimator. Otherwise, the properties of the MAP estimators depend on the shape of the prior distribution which could also be multimodal. In this case, a local maximum can be found by using Eq. (5). Differently, the EAP estimator always exists for a proper prior distribution. From a computational point of view, Owen's approximation does not involve iterative procedures and it is computationally feasible. The MAP estimator requires an iterative procedure, such as Newton–Raphson, while the EAP estimator involves numerical integration. Among Bayesian CAT researchers, Owen's method is very popular and the normal approximation is used also for the item selection in a Bayesian sequential updating of the posterior distribution of θ. However, we should underline that the method is based on an approximation, restricting the potential of a fully Bayesian approach. In the past, the use of Owen's method was justified by its computational feasibiliy. Nowadays, the availability of modern computers has strongly reduced computational limitations related to fully Bayesian estimation and we believe that new approaches should be experienced.

To complete CAT, Steps 2–4 of the algorithm are repeated iteratively until a stopping rule is satisfied (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.

2.2 Empirical information in CAT

During test administrations, 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 estimation within linear testing has been discussed by various authors (see e.g. Zwinderman 1991, 1997). In CAT, the introduction of empirical information has been discussed especially for ability initialization and item selection (Gialluca and Weiss 1979; van der Linden 1999, 2008). In the paper of van der Linden (1999), an empirical initialization of the ability estimator is proposed. To this aim, 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

$$\theta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_P X_{iP} + \epsilon_i, \tag{7}$$

where the error terms are assumed to be independent and normally distributed as $\epsilon_i \sim N(0, \sigma^2)$ 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

$$\theta_i | X_{i1}, \ldots, X_{iP} \sim N(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_P X_{iP}; \sigma^2). \tag{8}$$

Equation (8) represents an informative prior distribution for ability. When regression (7) is estimated with satisfying precision, 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

$$\hat{\theta}_{i0} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_P x_{iP}. \tag{9}$$

As a consequence, a better provisional ability estimate is provided and the first item is selected closer to the true ability of the person. Equation (9) can be used also in item selection based on a point estimate for θ.

3 MCMC CAT with empirical information

In order to improve the efficiency of CAT, we propose a fully Bayesian approach based on MCMC ability estimation with empirical prior information. In the proposed approach, there are mainly two elements of novelty. First of all, empirical information is introduced not only in the ability initialization, but also in the ability estimation, allowing for further improvements in the efficiency of the CAT algorithm. Moreover, unlike the existing literature about the use of empirical information in CAT, the effectiveness of the approach is explored by using comparative simulation studies. Secondly, the Gibbs sampler algorithm is used for ability estimation in CAT. Despite MCMC methods are becoming very popular for the estimation of item and ability parameters in IRT models within linear testing, their implementation in CAT is completely unexplored. A fully Bayesian approach is chosen to overcome the limitations of ML estimation for perfect response patterns. However, Bayesian methods depend on the choice of a prior distribution, which may be misleading especially in the initial ability estimation. To prevent this situation, we propose the use of empirical information in setting the parameters of the prior.

3.1 A joint use of empirical information in CAT

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. Here, we propose a joint use of empirical information in Step 1 and Step 4 of the CAT algorithm. In fact, besides the ability initialization, collateral information may be integrated in the ability estimation phase through the introduction of an empirical prior distribution within a fully Bayesian approach. The research is motivated by the increasing amounts of background information that becomes available about the candidates via all kinds of databases. 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. In medical testing, a patient's

health record could be used. Besides, it often happens that a whole battery of tests is administered to the candidate during an exam, or during a psychological screening.

Following the approach of van der Linden (1999), a linear relationship is assumed between the ability θ and a set of covariates, as described in Eq. (7). As a consequence, Eq. (8) becomes an informative normal prior distribution for θ to be included in the ability estimation phase.

When using collateral information in CAT ability initialization and estimation, two different problems are solved. First of all, the test length is reduced. Additionally, bias due to unforced errors during the beginning of CAT (Guyer 2008) is reduced as well, since the impact of these errors on ability estimation is much smaller due to the use of an informative prior. 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.

3.2 MCMC ability estimation in CAT

In order to proceed with the ability estimation in CAT, a Gibbs sampler scheme is proposed. Recently, 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 1992; 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 et al. 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 computational limitations have been strongly reduced. For unidimensional IRT models, MCMC computational intensiveness is no longer an obstacle.

In the current problem, the algorithm is modified to estimate ability in adaptive testing with the inclusion of an informative empirical prior. First of all, the presence of the binary response variable Y_j is 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

$$Y_j = \begin{cases} 1 & \text{if } Z_j > 0, \\ 0 & \text{if } Z_j \leq 0. \end{cases} \tag{10}$$

According to Eq. (10), 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

$$P(\boldsymbol{Z}, \theta, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = P(\boldsymbol{Z}|\theta, \boldsymbol{\xi}, \boldsymbol{Y}) P(\theta|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}) P(\boldsymbol{\xi}) P(\boldsymbol{\beta}) P(\sigma^2), \tag{11}$$

where $\boldsymbol{\xi}$ 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. $\boldsymbol{Z}|\theta, \boldsymbol{\xi}$
2. $\theta|\boldsymbol{Z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2$
3. $\boldsymbol{\xi}|\theta, \boldsymbol{Z}$
4. $\boldsymbol{\beta}|\boldsymbol{\theta}, \sigma^2$
5. $\sigma^2|\boldsymbol{\theta}, \boldsymbol{\beta}$.

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

$$Z_j|\theta, \boldsymbol{\xi} \sim \begin{cases} N(\eta_j, 1) & \text{with } Z_j > 0 \text{ if } Y_j = 1, \\ N(\eta_j, 1) & \text{with } Z_j \leq 0 \text{ if } Y_j = 0. \end{cases} \tag{12}$$

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

$$Z_j + \delta_j = \alpha_j\theta + \upsilon_j, \tag{13}$$

where υ_j are independent and identically distributed as $N(0, 1)$. Equation (13) 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

$$\theta \sim N(\hat{\theta}; \nu), \tag{14}$$

where $\hat{\theta} = (\alpha_j'\alpha_j)^{-1}\alpha_j'(Z_j + \delta_j)$ is the least square estimate of θ and $\nu = (\alpha_j'\alpha_j)^{-1}$ is the variance. Practically, the variance can be calculated as $\nu = 1/\sum_{j=1}^{J}\alpha_j^2$ 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 (8) and the combination of likelihood and prior leads to a normal posterior distribution, as follows

$$\theta|\boldsymbol{Z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2 \sim N\left(\frac{\hat{\theta}/\nu + \boldsymbol{X\beta}/\sigma^2}{1/\nu + 1/\sigma^2}; \frac{1}{1/\nu + 1/\sigma^2}\right). \tag{15}$$

By introducing the empirical prior within MCMC, the posterior distribution becomes candidate-tailored and more precise ability estimates can be obtained. After the kth 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 (12), 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 (15)
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 (The MathWorks Inc 2005).

4 Simulation studies

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. 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 results of a short study on the algorithm convergence are reported to justify the number of iterations used in the simulations.

In all studies, a normal distribution is chosen as prior for the ability θ. Despite the easiness of implementation of noninformative and improper priors, they may cause instability in the posterior estimates and convergence problems when the Gibbs sampler is used (Ibrahim and Chen 2000). Among informative and proper priors, the

normal distribution is chosen because it allows to work with Bayesian conjugate families of distributions, due to the fact that the measurement model is represented by model (1). Moreover, we would like to compare our approach to existing literature about Bayesian estimation of IRT models, where a standard normal distribution is usually chosen for θ.

4.1 A study 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, whereas 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* whereas 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 $j = 1, \ldots, k$. When the fully empirical approach is adopted, the linear relation $\theta = 0.2 + 0.7X + \epsilon$ with $\epsilon \sim N(0; 0.3)$ 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 $(\theta - 0.2 - \epsilon)/0.7$.

The Gibbs sampler with a chain length of 5,000 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 (SD items). With respect to ability, the expected posterior estimate, bias and standard deviation (SD) are reported. The results of the simulations are shown in Table 1.

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 SD. In fact, the standard solution fails to recover the ability levels when deviating from $\theta = 0$.

Table 1 Final test length and ability parameter recovery for fully empirical and standard solutions

True θ	Fully empirical					Standard				
	Mean n. items	SD items	$\hat{\theta}$	Bias	SD	Mean n. items	SD items	$\hat{\theta}$	Bias	SD
−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

4.2 A study 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 (T) 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.

Table 2 Ability parameter recovery for fully empirical and standard solutions ($T = 10$)

True θ	Fully empirical				Standard			
	$\hat{\theta}$	SD	Bias	RMSE	$\hat{\theta}$	SD	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

Table 2 continued

True θ	Fully empirical				Standard			
	$\hat{\theta}$	SD	Bias	RMSE	$\hat{\theta}$	SD	Bias	RMSE
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			
	$\hat{\theta}$	SD	Bias	RMSE	$\hat{\theta}$	SD	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

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

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

Table 4 Ability parameter recovery for fully empirical and standard solutions ($T = 20$)

True θ	Fully empirical				Standard			
	$\hat{\theta}$	SD	Bias	RMSE	$\hat{\theta}$	SD	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

(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.

4.3 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.

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 Fig. 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.

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.

Table 5 Ability parameter recovery for fully empirical, empirical initialization and standard solutions ($T = 10$)

True θ	Fully empirical				Empirical initialization				Standard			
	$\hat{\theta}$	SD	Bias	RMSE	$\hat{\theta}$	SD	Bias	RMSE	$\hat{\theta}$	SD	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

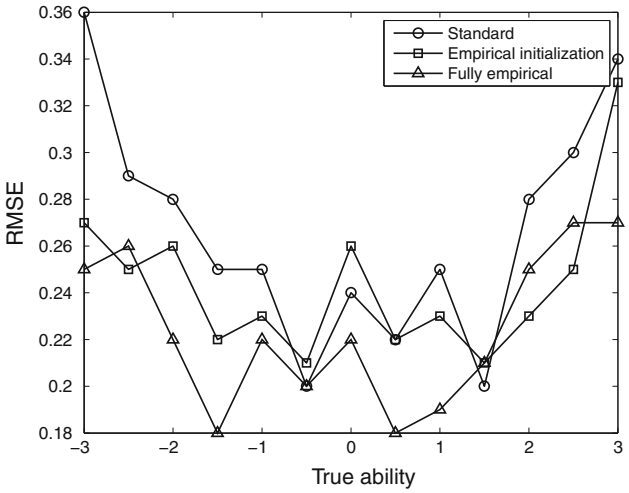

Fig. 1 Root mean square error (RMSE) for the three different approaches (fully empirical, empirical initialization and standard) when the test consists of 10 items

4.4 The choice of the number of iterations

One of the most critical issues in MCMC estimation is assessing the convergence of the algorithm, which is also needed to decide the number of total and to discard iterations. To this aim, diagnostic tools are employed (for a review, see Cowles and Carlin 1996; Gelman et al. 2004). The first intuitive diagnostic tool is the inspection of the

Table 6 Estimated accuracy of simulation across different number of iterations

N. iter	Burn-in	Fully empirical				Standard			
		$\hat{\theta}$	SD	5 % SD	MC error	$\hat{\theta}$	SD	5 % SD	MC error
1,000	100	−0.119	0.393	0.020	0.023	0.070	0.422	0.021	0.025
2,000	200	−0.101	0.391	0.020	0.013	−0.048	0.417	0.021	0.018
5,000	500	0.303	0.411	0.021	0.011	−0.135	0.427	0.021	0.008
10,000	1000	0.048	0.373	0.019	0.006	−0.107	0.410	0.021	0.008

trace plot of the simulated random draws. Even if convergence cannot be ensured by simply looking at the iteration history, a clearly situation of non-convergence can be detected immediately by identifying trends in the samples. Another diagnostic tool is represented by the study of autocorrelation, because patterns of serial correlations in the chain are responsible of slow convergence of the algorithm. After computing the posterior mean and the standard deviation, a measure of the standard error of estimate should be calculated. As suggested in Gelman et al. (2004), 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 diagnostic), as proposed by Geweke (1992) in order to provide an estimate of the asymptotic standard error. The basic idea is that the mean and the variance of the parameter posterior distribution should be equal in the first and in the second half of the chain. As a rule of thumb, the estimated Monte Carlo error should be less than 5 % of the standard deviation. Sensitivity analysis should be conducted to verify that, starting from difference over-dispersed points, the behavior of the chains is the same. To this end, the Gelman-Rubin R diagnostic could be used to compare the variances within and and across the chains.

In order to decide the necessary number of iterations for obtaining an acceptable accuracy, a study has been conducted. In particular, the simulation design of the second study is drawn on in the case of ability $\theta = 0$ and test length $T = 10$, using a different number of iterations (1,000, 2,000, 5,000 and 10,000). The estimate of the Monte Carlo (MC) standard error proposed by Geweke (1992) is considered. Table 6 shows the results both for the fully empirical and the standard approaches.

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 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 s) on a 2.66 GHz Intel Core2 Quad desktop. The simulations conducted by using 1,000 iterations do not satisfy the accuracy condition of MC error less than 5 % of the standard deviation, while the solution with 2,000 iterations slightly satisfies it. On the other hand, running 5,000 or 10,000 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 5,000. For each replication of the simulation studies described in the section, the MC error was assessed to be less than 5 % of standard deviation. Trace plots were inspected, showing random fluctuations of the sample values around the mean, without trends. Absence of autocorrelations was observed at lags greater than 5. Finally, by using multiple chains, the Gelman-Rubin R statistic was computed. Values close to 1 were found, suggesting that stationarity had been reached.

The chosen chain length represents a good compromise between speed of the algorithm 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.

5 Empirical examples

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, two case studies were chosen in the field of intelligence and educational testing.

5.1 An application in intelligence testing

Data regarding a computer adaptive intelligence test for personnel selection, the Connector Ability (Maij-de Meij et al. 2008) were available. The test aims at measuring the general intelligence factor (*G*-factor) by using different types of cognitive, nonverbal items. The complete test consists of three different subscales: Number series (NS), Figure series (FS), and Raven's matrices (RM). For each item, the candidate is required to identify the missing element that completes a pattern. All items measure general mental ability, in particular the candidate's capacity for analyzing and solving problems, abstract reasoning, and the ability to learn. The Connector Ability has been developed for applications in the area of HRM, for example for job selection or for career development.

The item bank for the Number series test consisted of 499 items calibrated with the 2PNO model (Maij-de Meij et al. 2008). Some descriptive statistics on the item parameters included in the item bank are shown in Table 7.

Discrimination parameters vary in the interval [0.180; 1.470], with a mean value around 0.7. The item discrimination parameter reflects the capability of the item to differentiate candidates with different ability. Items with high discrimination parameters are preferred in CAT, because they are also more informative, as demonstrated by the information function in Eq. 3. Items included in the item pool are rather discriminating. Difficulty parameters are included in the range $[-2.290; 2.300]$ with a mean of -0.4. Difficulty parameters are on the same scale of the ability θ, and optimal item pools contain items with different levels of difficulties in order to estimate the ability of different candidates accurately. The pool contains different items with respect to

Table 7 Descriptive statistics on the item parameters included in 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

difficulty. However, we should note that the median and the mean values are close to −0.4, denoting a slight asymmetry in favour of easy items.

Test results, including ability estimates for each of the three subscales, were available for a sample of 660 real examinees. The sample consists of about 61 % females and 39 % males. The majority (56 %) of the candidates comes from an higher vocational education, while the remaining 44 % from a university track. Also, the 40 % of the sample is Dutch native, the 44 % is western immigrant and the remaining part is non western immigrant. The mean ability in the RM subscale was equal to 0.01 (SD = 0.69), while the mean ability in the NS subscale was −0.24(SD = 0.70).

The relation between the RM and NS subscales was estimated, based on the reported abilities estimated for the subscales, resulting in the following empirical prior distribution

$$\theta | X_1 \sim N(-0.243 + 0.394 X_1; 0.414), \tag{16}$$

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 (16) is effective in this case study, an adaptive version of the NS test is simulated using the reported ability estimates for the group of 660 real examinees as true abilities for the candidates.

For each examinee, the adaptive test is replicated 10 times, and the ability estimation is performed by using 5,000 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.

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

Table 8 Results on the mean number of items needed in CAT simulation

Ability range	N. items with difficulty parameter in the range	N. examinees in the range	Fully empirical		Empirical initialization		Standard	
			Mean n. items	SD	Mean n. items	SD	Mean n. items	SD
−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	-	-	-	-	-	-

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.

5.2 An application in educational testing

In this study, data refer to the mathematics test administered by the Italian National Evaluation Institute for the School System (INVALSI) to students at the end of lower secondary school (eight grade). The test consists of 22 items and it is administered in the traditional linear fixed form. The test contains multiple-choice, open and close constructed-response items which have been recoded as binary items. The item parameters of the test items have been estimated according to model (1) by using a random sample of 4,865 students. The sample consists of about 51 % females and 49 % males, and the majority of students are Italian (93,7 %). Descriptive statistics on the estimated item parameters are reported in Table 9.

Item discrimination is moderate while item difficulties are included in the range $[-1.113; 1.252]$, denoting that the items are more informative for abilities that do not diverge too much from zero. Moreover, the test is characterized by rather easy items as can be inferred by the negative mean and median of the difficulty parameters.

Besides the mathematics test, a test on Italian language is administered to the same students. The test consists of 25 items about reading comprehension and grammar. The idea is to use information from the Italian test in the empirical distribution for

Table 9 Descriptive statistics on the item parameters for the mathematics INVALSI test

	Discrimination parameters	Difficulty parameters
Mean	0.600	−0.210
Median	0.569	−0.446
Standard deviation	0.173	0.684
Minimum	0.388	−1.113
Maximum	0.905	1.252

Table 10 Ability estimates for the empirical and standard solution on the mathematics INVALSI test

True θ	Empirical ($T = 5$)			Empirical ($T = 10$)			Standard ($T = 22$)		
	$\hat{\theta}$	SD	Bias	$\hat{\theta}$	SD	Bias	$\hat{\theta}$	SD	Bias
−2	−2.13	0.60	−0.13	−2.09	0.56	−0.09	−1.68	0.41	0.32
−1.5	−1.48	0.51	0.02	−1.57	0.51	−0.07	−1.29	0.47	0.21
−1	−1.02	0.55	−0.02	−1.04	0.46	−0.04	−0.83	0.45	0.17
−0.5	−0.57	0.57	−0.07	−0.47	0.55	0.03	−0.47	0.41	0.03
0	0.09	0.51	0.09	0.03	0.50	0.03	−0.02	0.47	−0.02
0.5	0.48	0.52	−0.02	0.48	0.52	−0.02	0.52	0.50	0.02
1	0.97	0.53	−0.03	1.07	0.48	0.07	0.91	0.44	−0.09
1.5	1.51	0.51	0.01	1.60	0.52	0.10	1.40	0.52	−0.10
2	2.10	0.61	0.10	2.11	0.58	0.11	1.77	0.58	−0.23

the ability in mathematics. Therefore, we estimated the linear relation between the mathematics ability score (θ) and the Italian ability score (X_1) on the random sample of 4,865 students. The results provided the following empirical prior distribution

$$\theta | X_1 \sim N(0.78 X_1; 0.53). \tag{17}$$

The estimated regression coefficient $\hat{\beta}_1 = 0.78$ shows a positive and strong relation between the two ability scores.

In order to demonstrate the effectiveness of our proposal, we treat the mathematics items as a small item pool for adaptive testing and we compare different methods in estimating student abilities in the range $[-2; 2]$ by intervals of width 0.5. As usual, our approach is denoted empirical and consists in the introduction of the information derived from distribution (17) both in the initialization and in the ability estimation. The approach is compared to the standard solution, where a standard normal is used as prior distribution for θ and the ability initialization is fixed in $\theta_0 = 0$. Differently to the previous study, we apply a fixed-length stopping rule where the total number of submitted items is 5 and 10 for the empirical approaches and 22 (the total test length) for the standard solution. For each ability value, 100 replications were used. The results in Table 10 are summarized in terms of estimated ability $\hat{\theta}$, standard deviation among the replications (SD) and Bias.

As can be clearly seen, the results show that the empirical approaches outperform the standard one in terms of bias. In fact, with only 5 items, the introduction of empirical information about the student performance in the Italian test is effective in reproducing the student ability in mathematics precisely. This is especially true for ability levels out of the range $[-0.5; 0.5]$, where the performance of both approaches are comparable. These results are direct consequences of the choice of the prior ability distribution and the ability initialization. In fact, the choice of a standard normal distribution as prior and of the ability initialization at $\theta_0 = 0$ leads to a precise estimation of candidates abilities close to zero, even if a larger number of items ($T = 22$) is used. On the other hand, the empirical approach is based on a variable initialization, improving the

estimation also with a short number of items. Finally, it can be noticed that the role of the prior is much more evident in the shorter test ($T = 5$), when likelihood has a lower weight in the estimation.

6 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 was 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. In particular, we believe that an MCMC approach would be even more useful when the abilities of the respondents do not have a normal distribution, or when the distribution is skewed (see e.g. Woods and Lin 2009). Finally, future research may deal with the introduction of Bayesian item selection methods, as used in van der Linden (1998), Veldkamp (2010).

Acknowledgments The authors would like to thank PiCompany and INVALSI for making data available and for fruitful discussion of the results.

References

Albert JH (1992) Bayesian estimation of normal ogive item response curves using Gibbs sampling. J Educ Stat 17:251–269

Ariel A, Veldkamp BP, van der linden WJ (2004) Cnstructing rotating item pools for contrained adaptive testing. J Educ Meas 41:345–360

Ariel A, van der linden WJ, Veldkamp BP (2006) A strategy for optimizing item pool management. J Educ Meas 43:85–96

Barrada JR, Abad FJ, Veldkamp BP (2009) Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. Psicothema 21:313–320

Bartholomew DJ (1987) Latent variable models and factor analysis. Oxford University Press, New York

Bartholomew DJ, Knott M (1999) Latent variable models and factor analysis. Arnold Publishers, London

Béguin AA, Glas CAW (2001) MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika 66:541–562

Belov DI, Armstrong RD (2009) Direct and inverse problems of item pool design for computerized adaptive testing. Educ Psychol Meas 69:533–547

Birnbaum A (1968) Some latent trait models and their use in inferring an examinee's ability. In: Lord FM, Novick MR (eds) Statistical theories of mental test scores. Addison-Wesley, Reading pp 97–479

Bock RD, Mislevy RJ (1988) Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas 6:431–444

Chang H-H, Ying Z (1996) A global information approach to computerized adaptive testing. Appl Psychol Meas 20:213–229
Cowles MK, Carlin BP (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. J Am Stat Assoc 91:883–904
De Jong MG, Steenkamp JBEM, Veldkamp BP (2009) A model for the construction of country-specific, yet internationally comparable short-form marketing scales. Mark Sci 28:674–689
Fox J-P, Glas CAW (2001) Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika 66:271–288
Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis. 2. Chapman and Hall/CRC, Boca Raton
Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–741
Geweke J (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo JM, Berger J, Dawid AP, Smith AFM (eds) Bayesian statistics 4. Oxford University Press, Oxford pp 169–193
Gialluca KA, Weiss DJ (1979) Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement. Research report 79-6, University of Minnesota, Department of Psychology, Psychometric Methods Program, Minneapolis
Guyer RD (2008) Effect of early misfit in computerized adaptive testing on the recovery of theta. Unpublished doctoral dissertation, University of Minnesota
Ibrahim JG, Chen MH (2000) Power prior distributions for regression models. Stat Sci 15:46–60
Johnson VE, Albert JH (1999) Ordinal data modeling. Springer, New York
Lehman EL, Casella G (1998) Theory of point estimation. 2. Springer, New York
Lord FM (1952) A theory of test scores. Psychometric monograph 7
Lord FM (1970) Some test theory for tailored testing. In: Holtzman WH (ed) Computer-assisted instruction, testing, and guidance. Harper and Row, New York pp 139–183
Lord FM, Novick MR (1968) Statistical theories of mental test scores. Addison-Wesley, Reading
Maij-de Meij AM, Schakel L, Smid N, Verstappen N, Jaganjac A (2008) Connector ability; professional manual. PiCompany B.V., Utrecht
Mislevy RJ (1986) Bayes modal estimation in item response models. Psychometrika 51:177–195
Natesan P, Limbers C, Varni JW (2010) Bayesian estimation of graded response multilevel models using Gibbs sampling: formulation and illustration. Educ Psych Meas 70:420–439
Owen RJ (1969) A Bayesian approach to tailored testing. Research report 69-92, Educational testing service, Princeton, NJ
Owen RJ (1975) A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. J Am Stat Assoc 70:351–356
Patz RJ, Junker BW (1999) A straightforward approach to Markov chain Monte Carlo methods for item response models. J Educ Behav Stat 24:146–178
Reckase MD (2009) Multidimensional item response models. Springer, New York
Segall DO (1996) Multidimensional adaptive testing. Psychometrika 61:331–354
Sheng Y, Wikle CK (2007) Comparing multiunidimensional and unidimensional item response theory models. Educ Psychol Meas 67:899–919
Sheng Y, Wikle CK (2008) Bayesian multidimensional IRT models with a hierarchical structure. Educ Psychol Meas 68:413–430
Sympson JB, Hetter RD (1985) Controlling item-exposure rates in computerized adaptive testing. In: Proceedings of the 27th annual meeting of the military testing association, Navy personnel research and development center, San Diego, CA, pp 973–977
The MathWorks Inc (2005) MATLAB [computer program]. The MathWorks Inc, Natick, MA
van der Linden WJ (1998) Bayesian item selection criteria for adaptive testing. J Educ Behav Stat 22: 203–226
van der Linden WJ (1999) Empirical initialization of the trait estimation in adaptive testing. Appl Psychol Meas 23:21–29
van der Linden WJ (2005) Linear models for optimal test design. Springer, New York
van der Linden WJ (2008) Using response times for item selection in adaptive testing. J Educ Behav Stat 33:5–20
van der Linden WJ, Glas CAW (2000) Computerized adaptive testing: theory and practice. Kluwer, Boston

van der Linden WJ, Glas CAW (2007) Statistical aspects of adaptive testing. In: Rao CR, Sinharay S (eds) Handbook of statistics (vol 27: Psychometrics). Elsevier B.V., Amsterdam pp 801–838
van der Linden WJ, Glas CAW (2010) Elements of adaptive testing. Springer, New York
van der Linden WJ, Pashley PJ (2010) Item selection and ability estimation in adaptive testing. In: van der Linden WJ, Glas CAW (eds) Elements of adaptive testing. Springer, New York pp 3–30
van der Linden WJ, Veldkamp BP (2004) Constraining item exposure rates in computerized adaptive testing with shadow tests. J Educ Behav Stat 29:273–291
van der Linden WJ, Veldkamp BP (2007) Conditional item exposure control in adaptive testing using item-ineligibility probabilities. J Educ Behav Stat 32:398–417
Veerkamp WJJ, Berger MPF (1997) Some new item-selection criteria for adaptive testing. J Educ Behav Stat 22:203–226
Veldkamp BP (2010) Bayesian item selection in constrained adaptive testing using shadow tests. Psicologica 31:149–169
Veldkamp BP (2012) Application of robust optimization to automated test assembly. Ann Oper Res. doi:10.1007/s10479-012-1218-y (online first)
Veldkamp BP, van der Linden WJ (2002) Multidimensional adaptive testing with constraints on test content. Psychometrika 67:575–588
Wainer H, Dorans NJ, Eignor D, Flaugher R, Green BF, Mislevy RJ, Steinberg L (2000) Computerized adaptive testing: a primer. 2. Lawrence Erlbaum Associates, Mahwah
Warm TA (1989) Weighted likelihood estimation of ability in item response theory with tests of finite length. Psychometrika 54:427–450
Woods CM, Lin N (2009) Item response theory with estimation of the latent density using Davidian curves. Appl Psychol Meas 33:102–117
Zwinderman AH (1991) A generalized Rasch model for manifest predictors. Psychometrika 56:589–600
Zwinderman AH (1997) Response models with manifest predictors. In: van der Linden WJ, Hambleton RK (eds) Handbook of modern item response theory. Springer, New York pp 245–256