The use of predicted values for item parameters in item response theory models: an application in intelligence tests

Mariagiulia Matteucci, Stefania Mignani, and Bernard P. Veldkamp

ABSTRACT

In testing, item response theory models are widely used in order to estimate item parameters and individual abilities. However, even unidimensional models require a considerable sample size so that all parameters can be estimated precisely. The introduction of empirical prior information about candidates and items might reduce the number of candidates needed for parameter estimation. Using data for IQ measurement, this work shows how empirical information about items can be used effectively for item calibration and in adaptive testing. First, we propose multivariate regression trees to predict the item parameters based on a set of covariates related to the item solving process. Afterwards, we compare the item parameter estimation when tree fitted values are included in the estimation or when they are ignored. Model estimation is fully Bayesian, and is conducted via Markov chain Monte Carlo methods. The results are two-fold: a) in item calibration, it is shown that the introduction of prior information is effective with short test lengths and small sample sizes, b) in adaptive testing, it is demonstrated that the use of the tree fitted values instead of the estimated parameters leads to a moderate increase in the test length, but provides a considerable saving of resources.

Keywords: item response theory models, Bayesian estimation, multivariate regression trees, item calibration, adaptive testing, intelligence tests.

1. Introduction

In educational and psychological measurement, a test consisting of a number of items is usually administered to candidates in order to infer their proficiency level or psychological profile.

The phase of testing is based on strong methodological theories, among which item response theory (IRT) has a long tradition (Lord and Novick, 1968; Hambleton and Swaminathan, 1985; van der Linden and Hambleton, 1997; de Ayala, 2009). IRT models express the probability of a response to a test item as a non linear function of the item characteristics, called item parameters, and a set of latent variables, representing individual non-observable cognitive abilities or personality traits.

Application of IRT involves two separate steps: a first one to estimate the item parameters and to ensure that they match the desired characteristics in terms of psychometric properties and test requirements (calibration), and a second step to locate examinees into the latent trait scale (scoring). Depending on the complexity of the fitted model and on the measurement context, these phases can be very expensive. In the calibration phase, the precision of the item parameter estimates depends partly on the sample size and several studies have been conducted to investigate the minimum requirements for different combinations of IRT models and test lengths (see e.g. De Ayala, 2009) and the consequences of not meeting these requirements (Veldkamp, Matteucci, and de Jong, submitted).

Within a Bayesian framework, the effects of a small sample size can be compensated by the application of a more informative prior distribution (Gelman, 2002), for example, by the application of an empirical prior distribution based on the relationship between the parameter of interest and a set of covariates. Generally, a normally distributed prior is used in Bayesian item parameter estimation (see e.g. Albert, 1992). When the information derived from auxiliary variables is accurate, the variability of the prior could be decreased considerably, so that it becomes more informative. Besides, the location of the prior might shift, so that it becomes more accurate.

In the literature, most attention has been paid to the relationship between the ability estimates and a set of covariates about the individual, such as demographical variables, socio-economic status, cultural habits, achievements in different tests, as can be found also in large-scale standardized assessments (see e.g. OECD, 2010). Several studies used the specified relationship effectively to improve the ability estimates both in terms of bias and measurement precision (Zwinderman, 1997; van der Linden, 1999; Matteucci, Mignani, and Veldkamp, 2009; Matteucci and Veldkamp, 2011). In IRT models, latent ability is typically viewed as a random variable when marginal maximum likelihood (Bock and Aitkin, 1981) or a fully Bayesian estimation (Albert, 1992) are adopted, and the specification of a prior distribution, even empirical, is immediate.

On the other hand, covariates about items, such as e.g. solving strategy, are collected in surveys less frequently than covariates about candidates. Relatively few studies have been conducted to investigate the impact of explanatory variables on item parameters (Enright and Sheehan, 2001; Wright, 2002), and this relation has hardly been used to fine-tune the prior distribution (Matteucci, Mignani, and Veldkamp, 2011). Within a fully Bayesian approach, the item parameters are viewed as random variables with their own prior distribution, and empirical prior distributions can be used, as well. Finding only few studies on this topic is a little unexpected, since, especially in educational measurement, the use of personal information about candidates may arise the issue of fairness with respect to candidates while, on the other hand, covariates about items do not state the matter of fairness. Test takers may claim that the introduction of information other than their responses in the test is not fair for the assessment, while this is not questioned for item parameters. Besides, a relation between the item parameters and auxiliary information can be found especially helpful in the possibility of predicting the item characteristics in terms of discrimination and difficulty, hence to build a frame of references for item writing. If item writers could know that items with certain features would hold specific psychometric properties, testing efficiency could be improved.

The main aim of this work is to show how efficiency of item parameter estimation can be improved when the sample size is small, or rather, in case the combination of sample size and test length is unfavourable, by using empirical information about the item parameters. In Matteucci, Mignani, and Veldkamp (2011), the performances of different prior distributions for item parameters commonly used in the literature were compared to empirical prior distributions. In particular, the relationship between item parameters and covariates was investigated by specifying two separate regression trees for item discrimination and difficulty, and the tree fitted values were used to set the hyperparameters of prior distributions. The results showed that, for a test length of 20 items, the empirical prior performed better than informative and vague priors commonly used in the literature, especially with small sample sizes. However, the use of two univariate trees appears restrictive, because the item discrimination and difficulty are usually interpreted jointly. Moreover, the approaches should be compared under different calibration conditions, based on several combinations of sample size and test length, and also in a different assessment environment, such as adaptive testing.

To overcome these limitations, in this paper a) multivariate regression trees are proposed to describe the link between the item psychometric properties and a set of available covariates about the item solving process, b) the empirical and non empirical approaches are compared under several conditions based on both sample and test size, c) a simulation study in adaptive testing is also conducted.

The paper is organized as follows. Section 2 describes the intelligence test data used throughout the paper. In Section 3, methods are presented. In particular, multivariate regression trees for predicting item parameters are discussed in Section 3.1 while Bayesian model estimation is described in Section 3.2, where the method is extended to the inclusion of empirical prior information. In Section 4, the main results are shown. They include the prediction of item parameters through regression trees and the advantages of including the fitted values in item calibration and adaptive testing. Finally, conclusions are addressed in Section 5.

2. Data

The data consist of an item bank containing 391 number series items (Matteucci, Mignani, and Veldkamp, 2011), commonly used in intelligence tests The available number series items represent a subscale of a test for IQ measurement, the Connector Ability (Maij- de Meij, Schakel, Smid, Verstappen, and Jaganjac, 2008), created by PiCompany, a Dutch Human Resources company. Connector Ability consists of three different subscales (figure series, number series, and Raven's matrices) and it is used for personnel selection purposes.

A number series item is simply a sequence of numbers and the candidate has to find the subsequent number according to some mathematical rule. An example is the following:

11 14 20 32 ?

Starting from number 11, one has to add 3 (+3) in order to obtain 14, then add 6 ($+3\cdot 2$) to get 20 and finally add 12 ($+6\cdot 2$) in order to get 32. Therefore, the answer to the item is 56. The item can be described by a first level operation (addition), a second level operation (multiplication), the number involved in the first level operation (=3), and the number involved in the second level operation (=2). In the example, a one-series item is presented. However, more difficult items are available, consisting of two series of numbers, where numbers in odd position follow one series while numbers in even position follow another series. Therefore, several covariates about items are available:

- first level operation (Op1) with categories 1=addition, 2=subtraction, 3=multiplication, 4=division;

- second level operation (Op2) with categories 1=addition, 2=subtraction, 3=multiplication, 4=division, 5=none;
- number involved in the first level operation (N1) with positive integer values;
- number involved in the second level operation (N2) with positive integer values;
- number of series (Ns) with values 1 for one series and 2 for two series.

Four alternatives were presented to the candidates, and they had to select the correct one. Item responses were recorded as binary data (1=correct, 0=incorrect). The estimation of the psychometric properties of the items was conducted by using IRT models.

Given a set of items, IRT models express the probability of response to a test item as a mathematical function of item parameters, and a single or multiple latent abilities. Under the assumption of unidimensionality, i.e. the presence of a single underlying variable, PiCompany conducted the item calibration according to the two-parameter logistic (2PL) model (Birnbaum, 1968) which specifies the probability of a correct response as a monotonically increasing function of the underlying trait, as follows

$$P(Y_{ij} = 1 \mid \theta_i, a_j, b_j) = \frac{e^{(a_j \theta_i - b_j)}}{1 + e^{(a_j \theta_i - b_j)}}, \tag{1}$$

where Y_{ij} denotes the response variable of individual i to item j, with i=1,…,n and j=1,…,k, θ_i is the ability of person i, a_j and b_j are the item parameters for item j. The discrimination parameter a_j assesses the power of the item to differentiate candidates of different ability, while the difficulty parameter b_j represents the threshold level of the item. Parameter estimation produced a set of two real-valued parameter estimates for the 391 items in the pool.

The introduction of a scaling constant D=1.702 which multiplies the term ($a_j\theta_i$-b_j) both in the numerator and the denominator, makes model (1) equivalent to the two-parameter normal ogive

(2PNO) model (Lord, 1952; Lord and Novick, 1968) in terms of predicted probabilities (for the proof, see Haley, 1952). The 2PNO model expresses the probability of success to a test item *j* as follows

$$P(Y_{ij}=1\,|\,\theta_i)=\Phi(\eta_{ij})=\int_{-\infty}^{\eta_{ij}}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz, \qquad (2)$$

where Φ is the cumulative normal distribution function, $\eta_{ij}=\alpha_j\theta_i$-$\delta_j$, with α_j and δ_j equivalent in meaning to the 2PL model item discrimination and difficulty parameters respectively, but on a different metric. Thanks to model (1) and (2) probability equivalency, the 2PNO model could be chosen as the measurement model, ensuring an easier treatability within Bayesian estimation.

Figures 1a and 1b show the box-plots related to item discrimination α_j and item difficulty δ_j parameters, respectively, in model (2) metric.

[INSERT FIGURE 1a AT ABOUT HERE] [INSERT FIGURE 1b AT ABOUT HERE]

As regards the discriminations, the plot clearly shows that all values are positive, ensuring that the item characteristic curve is an increasing function of the latent ability. In fact, negative discrimination parameters are not desirable but theoretically possible, and would mean that the probability of giving a correct response to the target item could decrease for increasing abilities. The higher the discrimination parameter is, the higher is the capability of the item to differentiate between the candidates. When the discrimination value is above 0.7, the item becomes more and more discriminating, thanks to an increasing steepness. In the pool, the median discrimination is equal to 0.76 while the mean value is 0.69, with a standard deviation of 0.40. Difficulty parameters usually range from -2 to 2, and the difficulty level of an item increases as the parameter increases.

In the pool, a mean value of -0.34 is observed, with a standard deviation of 0.72, while the median difficulty is -0.30. As a consequence, the item pool is not precisely balanced in terms of difficulty but easy items are predominant. The representation of item parameters through a scatter plot, as shown in Figure 2, can be useful to interpret the parameter values simultaneously.

[INSERT FIGURE 2 AT ABOUT HERE]

What can be seen, is that most items are moderately discriminating and are associated to a medium-low difficulty. The range of most discrimination parameters is from 0.3 to 1, while most difficulties are in the interval [-1.5; 0.5]. This implies that the item bank is more informative for individuals with low abilities. In fact, the amount of information provided by each item can be described by the item information function (Fisher information), which, according to model (2), can be expressed by

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

It can be demonstrated that, when the ability of the candidate is equal to the item difficulty, the item provides maximum information. Moreover, it can be observed that the information function (3) is proportional to the squared discrimination parameter, which means that the most discriminating items are the most informative too.

3. Methods

In order to investigate the relationship between the item parameters and the set of covariates described in Section 2, so that item parameters of new items could be predicted, the use of multivariate regression trees is proposed. When the form of the function relating the response variables and the covariates is not known, we can use a supervised methodology based only on data. Moreover, regression trees can be used to identify homogeneous clusters with respect to the covariates. This aspect is very important in the phase of item writing because, already during the item writing process, an accurate prediction of the item parameters can be made. Besides, for some applications, item parameters might even be assigned based on item characteristics and the expensive step of pre-testing can be skipped.

We want to seek for a single relationship between the bivariate response variable, consisting of the discrimination and the difficulty parameters, and the shared covariates. As the item parameters are usually estimated and interpreted jointly, we used a bivariate tree to reproduce better the link between the dependent and the explanatory variables. Again, this approach would be particularly efficient for item writing: given the covariates, items with the desired combination of discrimination and difficulty could be obtained.

The results of multivariate regression trees are employed to improve item parameter estimation. To this end, a fully Bayesian estimation procedure is implemented, where a prior distribution for item parameters should be specified. The estimation is reviewed and the method is extended to the possibility of including empirical information in the prior distribution for item parameters.

The methods are described in the following.

3.1 Multivariate regression trees

In the approach of classification and regression trees (CART) by Breiman, Friedman, Olshen, and Stone (1984), binary segmentation is used to find a subset of covariates which best predict a single outcome variable, either categorical or quantitative. CART is a nonparametric technique which works by recursively partitioning the complete dataset (root node) by using binary splits in the covariates so that the heterogeneity (impurity) of the resulting nodes at each split is minimized.

When the outcome variable is quantitative, regression trees are used. Deviance is generally taken as impurity measure and each split works by minimizing the deviance within the two nodes resulting from the binary partition or, analogously, maximizing the deviance between the nodes. When a given tree size is reached or the deviance is below a certain threshold, no further partitioning is possible and a maximal tree is specified. The maximal tree typically overfits the data and should be pruned according to some criteria in order to get an optimal tree. Several pruning techniques, such as the 1-SE (one-standard error) rule or techniques based on cost-complexity measure were proposed by Breiman et al. (1984).

Multivariate regression trees (MRT) represent an extension of CART, for continuous outcome variables, to the multivariate case. The first attempt in this direction was made by Segal (1992), who extended the regression tree methodology to longitudinal data. Thereafter, MRT were developed nearly simultaneously by De'ath (2002) and Larsen and Speckman (2004), to describe the relationships between the abundance of co-occurring species and environmental variables.

A multivariate tree should be simultaneously good for estimating the mean response of several dependent variables (Larsen and Speckman, 2004). Given a set of P response variables $w_1,\ldots,w_P$, the impurity of each node should be defined for the multivariate extension. Following the approach proposed by De'ath (2002), the impurity I of a given node N, with N representing a subset of the indices $\{1,\ldots,n\}$ denoting the observational units, is defined as the total sum of squares (SS) around the multivariate mean, as follows:

$$I(N) = \sum_{j \in N} \sum_{p=1}^{P} \left(w_{jp} - \overline{w}_p \right)^2, \tag{4}$$

where w_{jp} is the observed response for variable p and unit j, and $\overline{w}_p$ is the mean of variable w_p at node N. Geometrically, the impurity defined by (4) is the squared Euclidean distance of observations around the node centroid. Other measures of impurity, based on the multivariate sums of absolute deviation around the median or on a different definition of distance, can be used. When the impurity measure (4) is adopted, the multivariate trees are called SS-MRT to recall that the sum of squares is used.

Following the CART approach, the splits are binary and made by a single explanatory variable. Each split is chosen so that the sums of squared distances (SSD) of units from the centroids of their respective nodes are minimized. Once the maximal tree is build, a pruning method should be defined in order to choose the best tree size. A cost-complexity measure can be adopted, taking into account both the tree deviance and the tree complexity through a cost-complexity parameter (Breiman et al., 1984).

Cross validation (CV) is often employed, choosing the tree with the smallest predicted mean square error. In detail, the complete dataset is split into a number of approximately equal subsets (typically 10) that are used for validation. For each subset, the impurity of predictions based on the remaining data is calculated, and the CV error is computed by averaging the results of the validation runs. Different runs will produce slightly different CV errors, because the subsets are randomly selected. Following the approach of SS-MRT, the prediction error E can be defined by

$$E = \sum_{p=1}^{P} \left(w_p^* - \overline{w}_p \right)^2, \tag{5}$$

where w_p^* denotes a new observation.

3.2 Bayesian model estimation

In IRT model estimation, item parameters are usually viewed as fixed and unknown quantities. We decided to use a fully Bayesian estimation, where item parameters are considered random variables

(Albert, 1992), and the information derived from the explanatory variables can be incorporated in prior distributions.

Given the vector of binary responses $\mathbf{Y} = (Y_{11},\ldots,Y_{ij},\ldots,Y_{nk})$, with i=1,…,n individuals and j=1,…,k items, a vector of independent random variables $\mathbf{Z} = (Z_{11},\ldots,Z_{ij},\ldots,Z_{nk})$ is created to represent the continuous underlying responses, so that Y_{ij}=1 when Z_{ij}>0 and Y_{ij}=0 otherwise, and $Z_{ij} \sim N(\alpha_j\theta_i - \delta_j;1)$ when model (2) is chosen as the measurement model. With this approach, the joint posterior distribution of interest is specified as $P(\mathbf{Z}, \boldsymbol{\theta}, \xi \mid \mathbf{Y})$, where $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_i,\ldots,\theta_n)$ is the vector of ability parameters and $\xi = (\xi_1,\ldots, \xi_j,\ldots, \xi_k)$ is the complete vector of item parameters with the generic element $\xi_j=(\alpha_j, \delta_j)$ representing the two item parameters for item j. The joint posterior distribution has an intractable form while all conditional distributions are easy to simulate (Albert, 1992). For this reason, the Gibbs sampler (Geman and Geman, 1984) can be used to reproduce the target distribution by sampling iteratively each single conditional distribution until convergence. The algorithm, which is included in Markov chain Monte Carlo (MCMC) methods, works with the following conditional distributions given the response data **Y**:

(1) $\mathbf{Z} \mid \boldsymbol{\theta}, \xi$,

(2) $\boldsymbol{\theta} \mid \mathbf{Z}, \xi$,

(3) $\xi \mid \boldsymbol{\theta}, \mathbf{Z}$.

As described in Albert (1992), the first conditional distribution is truncated normal, as follows

$$Z_{ij} \mid \boldsymbol{\theta},\xi \sim \begin{cases} N(\eta_{ij};1) & with \quad Z_{ij} > 0 \quad if \quad Y_{ij} = 1, \\ N(\eta_{ij};1) & with \quad Z_{ij} \le 0 \quad if \quad Y_{ij} = 0. \end{cases} \tag{6}$$

Given a standard normal prior distribution for ability, i.e. $\{\theta_i\}$ i.i.d.~ $N(0,1)$, the second conditional distribution can be expressed by

$$\theta_i \mid \mathbf{Z}, \xi \sim \mathrm{N}\left(\frac{\hat{\theta}_i / v}{1/v+1}; \frac{1}{1/v} \right), \tag{7}$$

where $v = 1/\sum_{j=1}^{k} \alpha_j^2$ and $\hat{\theta}_i = v \cdot \sum_{j=1}^{k} \alpha_j (Z_{ij} + \delta_j)$. The third conditional distributions depends on the prior distribution for item parameters. A common choice (see e.g. Albert, 1992) is to assume a vague prior given by and indicator function such as $P(\xi) = \prod_{j=1}^{k} I(\alpha_j > 0)$, so that discrimination parameters are ensured to be positive. Alternatively, a prior covariance matrix for the item parameters can be specified (Béguin and Glas, 2001; Fox and Glas, 2001), as follows

$$\Sigma_0 = \begin{pmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\delta^2 \end{pmatrix}. \tag{8}$$

In Matteucci, Mignani, and Veldkamp (2011), a bivariate normal distribution was assumed as prior distribution for item parameters

$$\xi_j \sim \mathrm{N}(\boldsymbol{\mu}_0; \boldsymbol{\Sigma}_0), \tag{9}$$

where the hyperparameters to be specified are the prior means contained in $\boldsymbol{\mu}_0$ and the prior variances of the prior covariance matrix $\boldsymbol{\Sigma}_0$. Following this last, more general specification, the posterior distribution of the item parameters becomes

$$\xi_j \mid \boldsymbol{\theta}, \mathbf{Z} \sim \mathrm{N}\left(\left(\mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1} \left(\mathbf{X}'\mathbf{Z}_j + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right); \left(\mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1} \right), \tag{10}$$

where **X**=[**θ -1**] is a n×2 matrix containing **θ** in the first column and all elements equal to -1 in the second column. Starting with provisional estimates for abilities and item parameters, the Gibbs sampler is used to iteratively simulate samples from distributions (6), (7), and (10) until convergence is reached. Convergence can be assessed by inspecting the trace plots of simulated Markov chains, and applying convergence diagnostics, as reviewed in Cowles and Carlin (1996).

When using distribution (9) as prior for item parameters, a common choice is to set the prior means equal to zero and the prior variances equal to one. Variances can be increased to make the distribution less informative (Bolt and Lall, 2003). However, empirical information can be included in distribution (9) to set the prior means. This approach will be followed in the application, where fitted values from regression trees will be used as prior means for the prior distribution of item parameters.

4. Results

In this section, our approach has been applied to the intelligence test data. First of all, MRT are used to predict the item parameters of the intelligence test items. Afterwards, the inclusion of tree fitted values is evaluated for item calibration and for adaptive testing. Item calibration is simply the estimation of item parameters based on the responses of a random sample of individuals, drawn from the population of interest. Differently, adaptive testing is the administration of a tailored sequence of items to single individuals, and it is based on the availability of an item pool, where item parameters are known, e.g. estimated with an acceptable accuracy in the calibration phase. Item selection is adaptive as it is based on the provisional ability estimate of the single candidate.

4.1 Prediction of item parameters via MRT

The implementation of SS-MRT in our problem is straightforward. The multivariate response variable is represented by the two item parameters: discrimination α_j and difficulty δ_j (which are here treated as random variables), for j=1,…,k items. In the MRT notation, P=2, w_1=α, w_2=δ, while k is the total number of observed units. As described in Section 2, there are five explanatory variables: Op1, Op2, N1, N2, and Ns.

A multivariate regression tree was fitted by using the R package *mvpart*, which follows the approach of De'ath (2002). Ten-fold cross validation was used. In order to choose the best tree size, the relative error is plotted for different trees with a number of terminal nodes from 1 to 9 (see Figure 3).

[INSERT FIGURE 3 AT ABOUT HERE]

The upper curve represents the cross-validation error while the lower one is the resubstitution error for the sequence of nested trees. For each tree size, a complexity parameter (cp) can be used to define a cost-complexity function which identifies the best pruning. In the figure, the horizontal line (Min + 1 SE) shows the error limit according to the 1-SE (one-standard error) rule, which was adopted to choose the best tree size by cross-validation. According to this rule, the best tree is the one associated to less than one standard error above the minimum value of the impurity measure. In our application, the tree associated to a minimum error consists of 7 leaves, while the optimal tree has 6 leaves. The tree with 6 terminal nodes is judged to best fit the data and it is depicted in Figure 4.

[INSERT FIGURE 4 AT ABOUT HERE]

The tree describes the recursive partitioning process: for each node, a splitting variable is chosen to best separate the items into two groups. Three item covariates are involved in the tree: N2, Ns and Op1. The first splitting variable is N2 (the number involved in the second level operation), which creates a left branch for N2 < 5.5 and a right branch for N2 ≥ 5.5. As described in Section 2, N2 can take any value in the natural set. Therefore, the split can be translated as N2 ≤ 5 and N2 ≥ 6. When the Ns variable is used, the condition Ns < 1.5 means that only one series is present, while Ns ≥ 1.5 means that the item involves two series. Also, the Op1 variable splits the data as Op1=1, 2, 4 (addition, subtraction, and division) and Op1=3 (multiplication). Each terminal node N is characterized by the fitted values for the item parameters, the impurity measure according to (4), and the number of observations (items). The fitted values can be represented by the arithmetic means $\overline{\alpha}_N$ and $\overline{\delta}_N$, or by the median values $\tilde{\alpha}_N$ and $\tilde{\delta}_N$. In Figure 4, at a given node, the bar on the left represents the mean discrimination parameter $\overline{\alpha}_N$ while the bar on the right the fitted difficulty parameter $\overline{\delta}_N$. However, the median can also be chosen to represent the fitted values at each node, due to its robustness and to the property of minimizing the sum of the absolute deviations.

Reading the tree in Figure 4 from left to right, six leaves can be identified and numbered from 1 to 6. Their properties are summarized in Table 1, together with a description of the range of values taken by the covariates. In fact, each node can be viewed as a cluster of items, described by the covariates.

[INSERT TABLE 1 AT ABOUT HERE]

On average, the easiest items are in the terminal nodes 1, with a mean difficulty equal to -1.22, and a fairly good discrimination. The easiness of these items may depend on the first level operation, involving mainly addition and subtraction (in fact, division is not used very often in number series items, and the item pool contains only 14 items involving a division out of 391), the presence of one series only, and a low N2 number. On the other hand, the most difficult group of items can be found in leaf 6, where items are poorly discriminating. The difficulty can be imputed to the concurrent presence of two series in the item and high N2 numbers (N2≥6). Both nodes 2 and 3 consist of easy items, even if items are more discriminating in node 3. Finally, leaves 4 and 5 are characterized by moderately difficult items, with a low and fair mean discrimination, respectively.

4.2 A simulation study in item calibration

We want to evaluate the advantages of including the information derived from auxiliary variables about the item parameters in the item calibration under different conditions. To this aim, model estimation is conducted by setting the parameters of distribution (9) as follows: prior means are specified as the median values fitted through the bivariate regression tree of Section 4.1 and prior standard deviations σ_α and σ_δ are set to 0.5 so that the prior distribution could be fairly informative. Therefore, the prior distributions for item parameters are α_j~N($\tilde{\alpha}_N$,0.5) and δ_j~N($\tilde{\delta}_N$,0.5), where the *j*-th item belongs to the terminal node *N*. We call this approach *empirical* to underline that information derived from data is introduced in the estimation process. The approach is compared to the *classical* one, where prior means and variances for item parameters are set to zero and one, respectively: α_j~N(0,1) and δ_j~N(0,1).

To compare the estimation of item parameters in the empirical and the classical approach, a simulation study was conducted based on different combinations of test length (number of items) and sample size (candidates). Random samples of *k*=10, 20, 30 items were extracted from the pool of intelligence test items described in Section 2 and test submission was simulated for samples of *n*=100, 200, 300, 500 candidates, with $\{\theta_i\}$ i.i.d.~ *N*(0,1). The Gibbs sampler was implemented in the software MATLAB 7.1 (The MathWorks Inc., 2005) to estimate the item parameters. The convergence of the algorithm was assessed inspecting the iteration plot and calculating a time-series estimate of the Monte Carlo error, as proposed by Geweke (1992), which is implemented in the R package BOA (Smith, 2007). A rule of thumb is that the Monte Carlo error should be lower than 5% of the standard deviation. It was checked that this condition was met when 5000 total iterations were used with a burn-in of 500 iterations. For each condition, 100 replications were used. For each run, sampled item parameters from the posterior distribution (10) were recorded and the mean value was computed turning out with the expected a posterior (EAP) estimate for each item parameter. The mean value among replications is our final estimate of the item parameter ($\hat{\alpha}$ or $\hat{\delta}$) and the standard deviation (Sd) is a measure of stability over replications. For each estimate, bias was calculated together with the root mean square error (RMSE) as a measure of accuracy.

Results for *k*=10 items are shown in detail for *n*=100 in Table 2.

[INSERT TABLE 2 AT ABOUT HERE]

By “True *α*” and “True *δ*” we mean respectively the discrimination and difficulty parameters of items extracted from the pool which are taken as true item parameters in the simulations. Besides, the columns denoted by $\tilde{\alpha}_N$ and $\tilde{\delta}_N$ report the median values fitted with the bivariate tree, which are set as the mean value of prior distributions for item parameters.

Comparing the empirical and the classical approach when *n*=100, it can be seen that bias is lower in most cases for the classical approach while standard deviations and root mean square errors are definitely higher. Overall, we can say that the empirical approach is more stable over replications and it is also less variable when comparing the true and the simulated values among the replications. However, estimates are more biased than in the classical approach, with the exception of few cases (items 4, 7, 8 for both discrimination and difficulty). These results derive directly from the different prior parameters used in the prior distribution (9), and in particular from the use of a rather small prior standard deviation, equal to 0.5, for the empirical approach, in combination with a very small sample size of 100 simulees. The results in this particular case are useful to understand how the estimation properties evolves when the sample is increased, as shown in Table 3 for *n*=200.

[INSERT TABLE 3 AT ABOUT HERE]

Here, bias is generally reduced in both approaches with respect to the case of *n*=100. This improvement is more evident in the empirical solution, which is also associated to a lower variability over replications with respect to the classical solution as in the previous case.

Table 4 and Table 5 report the results for *n*=300 and *n*=500, respectively.

[INSERT TABLE 4 AT ABOUT HERE]

[INSERT TABLE 5 AT ABOUT HERE]

With a sample of 300 units, both bias and variability are reduced in the two approaches. Therefore, we can say that the two solutions are comparable, and this is true especially for the case of *n*=500.

Simulations were conducted also for test lengths different from *k*=10. The results for all the combinations of test length and sample size are summarized in Table 6 for discrimination parameters in terms of median bias and median RMSE.

[INSERT TABLE 6 AT ABOUT HERE]

As expected, the median bias decreases as test length and sample size increases. The classical solution outperforms the empirical one in term of bias for short tests (*k*=10) and small sample sizes (*n*=100, 200). In the other cases bias of the two approaches is comparable. On the other hand, the empirical solution presents the best results in terms of median RMSE, particularly for combinations of small test length and small sample size.

The same summarized statistics are presented for difficulty parameters in Table 7.

[INSERT TABLE 7 AT ABOUT HERE]

Again, the best performances of the empirical approach over the classical one can be observed for a very small number of items (*k*=10) and small samples.

It can be noticed that estimates of difficulty parameters are generally more accurate that those of discrimination parameters. The estimates depend on the abilities of the simulated sample, as well as on the specified prior distribution for item parameters. Usually, difficulty estimates are more accurate when abilities are sampled so that the mean ability level is close to the item difficulty. On the other hand, the estimation of discrimination parameters needs well spread abilities to be accurate (see, e.g. de Gruijter and van der Kamp, 2008). This last condition can be more difficult to be reached, and this is why difficulty parameter estimates are more stable and accurate also with small samples while discrimination estimates are not.

4.2 A simulation study in adaptive testing

To evaluate the advantages in the introduction of empirical information, we also considered a case in adaptive testing. In fact, the number series items can be submitted within an automated environment, and the attention is focused on the single examinee respect than on the whole group of test-takers.

Unlike linear testing, computer adaptive testing (CAT) works by submitting a different selection of test items to each candidate, where each subsequent item is adapted to the individual current ability estimate. Here the focus is on the estimation of the examinee ability (scoring phase) with the aim of finding the “optimal” test length.

The administration of an adaptive test starts with ability initialization, in which the initial proficiency level of the candidate is defined. A common choice is typically to set $\theta^{(0)}$=0. 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; Matteucci and Veldkamp, 2011). Afterwards, a criterion for item selection should be defined (for a review, see van der Linden and Pashley, 2010), where a common practice is to use the maximum information criterion (Birnbaum, 1968). According to this criterion, the item providing the highest information is selected from the pool to be administered. Information can be defined as the item information function described by (3). After a specified level of measurement precision is reached or a maximum number of items is submitted, the adaptive algorithm stops and the final ability of the candidate should be estimated.

In this simulation study, an adaptive test is simulated starting from two different items pools. The first one (pool 1) is the number series item bank provided by PiCompany, which contains the estimated item discrimination and difficulty for each of the 391 items. The second pool (pool 2) is built by substituting the estimated item parameters with discrimination and difficulty levels fitted by using regression trees. Clearly, pool 1 contains more heterogeneous items in terms of psychometric properties than pool 2, and items are more easily adapted to the individual candidate.

Adaptive tests were simulated for different candidates with ability from -2 to 2. Ability was estimated by using the Gibbs sampler with known item parameters, where the algorithm works only with the conditional distribution of the underlying response variables (6) and the posterior distribution of the ability (7). Test information above 4.5 was used as a stopping rule and 100 replications were conducted for each ability level.

Table 8 shows the mean number of items needed in order to complete the test and the corresponding standard deviation (Sd) for each ability level.

[INSERT TABLE 8 ABOUT HERE]

As can be easily seen, the number of items needed to get an ability estimate is higher when pool 2 is employed rather than pool 1. This is an expected result, as pool 1 is more variable in terms of item parameters and, as a consequence, provides more information on ability levels than pool 2. However, the crucial point is: how much do we lose by using pool 2 instead of pool 1? Within pool 2, the mean test length is around 9-12 items for low and average abilities which is definitely an acceptable number of items to be submitted in an automated test. Because the item pool is not symmetric with respect to difficulty, the results are not symmetric in the ability scale. In fact, for higher ability levels, the mean test length for a CAT increases rapidly either using pool 1 or pool 2.

Clearly, the use of MRT fitted values as item parameters makes the item pool less optimal from a psychometric point of view because most items are not distinguishable from the others. On the other hand, fitted regression trees could be used to impute item parameters of new items, when covariates are available. As a consequence, the calibration phase would not be needed and a huge saving of resources could be obtained.

5. Concluding remarks

IRT models, used intensively in educational and psychological measurement, may require quite large samples to estimate item parameters with good accuracy. Consequent research questions are how this estimation can be improved, and under which condition it should be improved.

In this paper, we showed how empirical information based on covariates dealing with the item solving process could be included to improve the estimation of item parameters and could be used in an adaptive testing environment. A pool of intelligence test items designed for personnel selection was used.

First of all, we proposed to use multivariate regression trees to predict the item parameters on the basis of a set of covariates. In particular, it was shown that a regression tree, with a bivariate response vector given by the item discrimination and difficulty parameters, was able not only to return fitted values for the response variables but also to identify homogeneous clusters of item parameters with respect to the selected covariates. This result is crucial, in fact information about the item solving process could be used to write items with certain psychometric properties with minor effort.

As a second step, we proposed to use tree fitted values in the prior distribution of item parameters, within a fully Bayesian approach. Simulations based on the real pool of intelligence test items were conducted to compare the item parameter estimation among the proposed empirical approach and the classical one, where a standard normal is specified as prior distribution for item parameters. The results showed that empirical priors improved the parameter estimation especially in terms of efficiency, when a short test was submitted (k=10), combined with a small sample size.

A further study was realized in an automated testing environment, where adaptive tests were simulated for candidates of different ability starting from two item pools. Using the empirical pool, with item parameters fitted by using regression trees, we noticed an increase in the number of items needed to complete the test. However, this was expected because the empirical pool contains more homogeneous items with respect to their psychometric properties and a larger number of items is needed in order to reach the same measurement precision. A test length up to 20 items is common in adaptive testing, and this approach could be used in practice to skip subsequent item pre-testing. MRT based item parameters might be used as initial item parameters, that could be updated on-the-fly (see e.g. Makransky and Glas (2010) for a comparison of methods of online calibration).

Even if we found encouraging results, some aspects need to be deepened. First, the study was conducted using real data on intelligence items, while different item banks may lead to different relationships between the item parameters and the covariates. Also, covariates about items may be difficult to collect, especially when studying psychological traits or in medicine. Finally, much more complicated models should be estimated to verify the efficiency of our proposal.

Further research may evaluate more conditions in the simulations, such as different sample distributions for the simulees and different IRT models. Moreover, the Gibbs sampler allows to estimate item parameters and abilities jointly, so that a joint use of empirical information at item and person level could be assessed. Lastly, in adaptive testing, item pools based on estimated item parameters could be integrated by item parameters predicted on the basis of auxiliary information.

Acknowledgements This work was partially supported by a research grant from the Italian Ministry of University and Research (MIUR), PRIN 2008 on "Latent structure analysis: new boundaries in statistical methods and models". The authors would like to thank PiCompany for the data availability.

References

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

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

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.

Bock, R.D., Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46(4): 443-459.

Bolt, D.M., Lall, V.F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement* 27: 395-414.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth International.

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.

De Ayala, R.J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

De'ath, G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology 83(4): 1105-1117.

De Gruijter, D.N.M., van der Kamp, L.J.T. (2008). *Statistical Test Theory for the Behavioral Sciences*. Boca Raton, FL: Chapman & Hall/CRC.

Enright, M.K., Sheehan, K.M. (2002). Modeling the difficulty of quantitative reasoning items: implication for item generation. In: Irvine, S.H., Kyllonen, P.C., eds. *Item Generation for Test Development*. Lawrence Erlbaum Associates.

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

Gelman, A. (2002). Prior distribution. In: El-Shaarawi, A.H., Piegorsch, W.W., eds., *Encyclopedia of Environmetrics*, vol 3, John Wiley & Sons.

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: Bernardo, J.M., Berger, J., Dawid, A.P., Smith, A.F.M., eds., *Bayesian Statistics 4*. Oxford University Press.

Haley, D.C. (1952). Estimation of the dosage mortality relationship when the dose is subject to error. Technical report N. 15, Stanford University, Applied Mathematics and Statistics Laboratory.

Hambleton, R.K., Swaminathan, H. (1985). Item response theory: principles and applications. Kluwer Nijhoff Publishing, Boston.

Larsen, D.R., Speckman, P.L. (2004). Multivariate regression trees for analysis of abundance data. *Biometrics* 60: 543-549.

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

Lord, F.M., Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

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

Makransky, G., Glas, C. A. W. (2010). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology* 11: 1-29.

Matteucci, M., Veldkamp, B.P. (2011). Including empirical prior information in test administration. In: Fichet, B. et al., eds. *Classification and Multivariate Analysis for Complex Data Structures*. Springer-Verlag .

Matteucci, M., Mignani, S., Veldkamp, B.P. (2009). Issues on item response theory modelling. In Bini, M., Monari, P., Piccolo, D., Salmaso, L., eds. *Statistical Methods for the Evaluation of Educational Services and Quality of Products*. Springer-Verlag.

Matteucci, M., Mignani, S., Veldkamp, B.P (2011). Prior Distributions for Item Parameters in IRT Models. Paper accepted for publication in: *Communications in Statistics – Theory and Methods*.

OECD (2010), PISA 2009 Results: Overcoming Social Background – Equity in Learning, Opportunities and Outcomes (Volume II). http://dx.doi.org/10.1787/9789264091504-en

Segal, M.R. (1992). Tree-Structured Methods for Longitudinal Data. *Journal of the American Statistical Association* 87: 407-418.

Smith, B.J. (2007). Boa: an R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software* 21: 1-37.

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., Hambleton, R.K. (1997). Handbook of Modern Item Response Theory. New York: Springer-Verlag.

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.

Veldkamp, B. P., Matteucci, M., & de Jong, M (submitted). Uncertainties in the item parameter estimates and automated test assembly.

Wright, D. (2002). Scoring tests when items have been generated. In: Irvine, S.H., Kyllonen, P.C., eds. *Item Generation for Test Development*. Lawrence Erlbaum Associates.

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. Properties of the tree leaves (terminal nodes).

Leaf	$\overline{\alpha}_N$	$\tilde{\alpha}_N$	S.d. α	$\overline{\delta}_N$	$\tilde{\delta}_N$	S.d. δ	I(*N*)	N. items	Op1	Op2	N1	N2	Ns
1	0.87	0.77	0.37	-1.22	-1.13	0.52	24.3	60	Add, Sub, Div	Any	Any	≤ 2	1

2	0.63	0.51	0.31	-0.68	-0.68	0.65	13.4	26	Mul	Any	Any	≤ 2	1
3	0.95	0.86	0.41	-0.49	-0.44	0.45	48.5	130	Any	Any	Any	≥ 3	1
4	0.45	0.37	0.28	0.03	0.02	0.52	16.3	47	Any	Any	Any	≤ 5	2
5	0.75	0.67	0.31	0.06	0.09	0.50	33.6	99	Any	Any	Any	≥ 6	1
6	0.41	0.33	0.26	0.58	0.53	0.66	14.4	29	Any	Any	Any	≥ 6	2

Note. Add=addition, Sub=subtraction, Mul=multiplication, Div=division.

Table 2. Estimates of item parameters for *k*=10 items and sample size *n*=100.

			Empirical				Classical						Empirical				Classical			
Item	True α	$\tilde{\alpha}_N$	$\hat{\alpha}$	Bias	Sd	RMSE	$\hat{\alpha}$	Bias	Sd	RMSE	True δ	$\tilde{\delta}_N$	$\hat{\delta}$	Bias	Sd	RMSE	$\hat{\delta}$	Bias	Sd	RMSE
1	0.62	0.77	0.70	0.08	0.21	0.22	0.68	0.06	0.26	0.26	-1.12	-1.13	-1.18	-0.06	0.19	0.20	-1.14	-0.03	0.20	0.20
2	0.78	0.77	0.87	0.09	0.23	0.25	0.81	0.03	0.27	0.27	-0.62	-1.13	-0.71	-0.10	0.16	0.19	-0.62	0.00	0.20	0.20
3	0.36	0.37	0.39	0.03	0.16	0.17	0.38	0.02	0.20	0.20	0.42	0.02	0.40	-0.03	0.13	0.13	0.41	-0.02	0.13	0.13
4	1.07	0.86	1.04	-0.03	0.22	0.22	1.09	0.03	0.25	0.25	-0.43	-0.44	-0.41	0.02	0.16	0.16	-0.40	0.03	0.18	0.18
5	0.53	0.67	0.59	0.06	0.18	0.18	0.58	0.05	0.25	0.25	0.07	0.09	0.08	0.01	0.13	0.13	0.07	0.00	0.16	0.16
6	0.60	0.86	0.71	0.11	0.19	0.22	0.68	0.08	0.27	0.28	-0.82	-0.44	-0.85	-0.03	0.18	0.18	-0.85	-0.04	0.24	0.24
7	0.86	0.67	0.86	0.00	0.19	0.19	0.93	0.06	0.29	0.29	0.38	0.09	0.38	-0.01	0.15	0.15	0.43	0.04	0.18	0.19
8	0.51	0.51	0.53	0.02	0.18	0.18	0.54	0.02	0.26	0.26	-1.15	-0.68	-1.13	0.02	0.16	0.16	-1.18	-0.03	0.20	0.21
9	0.30	0.33	0.23	-0.07	0.15	0.17	0.34	0.04	0.25	0.25	1.71	0.53	1.55	-0.16	0.17	0.23	1.81	0.10	0.27	0.29
10	0.35	0.67	0.42	0.07	0.17	0.18	0.39	0.04	0.18	0.19	0.45	0.09	0.46	0.01	0.12	0.12	0.46	0.01	0.14	0.14

Table 3. Estimates of item parameters for *k*=10 items and sample size *n*=200.

			Empirical				Classical						Empirical				Classical			
Item	True α	$\tilde{\alpha}_N$	$\hat{\alpha}$	Bias	Sd	RMSE	$\hat{\alpha}$	Bias	Sd	RMSE	True δ	$\tilde{\delta}_N$	$\hat{\delta}$	Bias	Sd	RMSE	$\hat{\delta}$	Bias	Sd	RMSE
1	0.62	0.77	0.67	0.05	0.18	0.19	0.63	0.01	0.20	0.20	-1.12	-1.13	-1.17	-0.05	0.13	0.14	-1.13	-0.01	0.16	0.15
2	0.78	0.77	0.83	0.05	0.17	0.18	0.79	0.02	0.18	0.18	-0.62	-1.13	-0.66	-0.04	0.13	0.13	-0.64	-0.02	0.12	0.12
3	0.36	0.37	0.36	0.00	0.13	0.13	0.39	0.03	0.15	0.16	0.42	0.02	0.41	-0.01	0.09	0.09	0.44	0.02	0.10	0.10
4	1.07	0.86	1.09	0.02	0.21	0.21	1.04	-0.03	0.24	0.24	-0.43	-0.44	-0.44	-0.01	0.12	0.12	-0.42	0.00	0.15	0.15
5	0.53	0.67	0.59	0.06	0.16	0.17	0.54	0.02	0.16	0.16	0.07	0.09	0.05	-0.02	0.10	0.10	0.08	0.01	0.10	0.10
6	0.60	0.86	0.64	0.04	0.16	0.16	0.61	0.01	0.18	0.18	-0.82	-0.44	-0.82	0.00	0.11	0.11	-0.83	-0.01	0.12	0.12
7	0.86	0.67	0.87	0.00	0.15	0.15	0.88	0.02	0.18	0.18	0.38	0.09	0.38	0.00	0.12	0.12	0.40	0.02	0.13	0.13
8	0.51	0.51	0.51	0.00	0.16	0.16	0.57	0.05	0.19	0.20	-1.15	-0.68	-1.14	0.00	0.14	0.14	-1.17	-0.02	0.16	0.16
9	0.30	0.33	0.26	-0.04	0.15	0.15	0.31	0.01	0.24	0.24	1.71	0.53	1.67	-0.04	0.15	0.16	1.75	0.05	0.20	0.21
10	0.35	0.67	0.41	0.07	0.13	0.15	0.35	0.00	0.14	0.14	0.45	0.09	0.46	0.01	0.10	0.10	0.45	0.00	0.10	0.10

Table 4. Estimates of item parameters for *k*=10 items and sample size *n*=300.

			Empirical				Classical						Empirical				Classical			
Item	True α	$\tilde{\alpha}_N$	$\hat{\alpha}$	Bias	Sd	RMSE	$\hat{\alpha}$	Bias	Sd	RMSE	True δ	$\tilde{\delta}_N$	$\hat{\delta}$	Bias	Sd	RMSE	$\hat{\delta}$	Bias	Sd	RMSE
1	0.62	0.77	0.66	0.04	0.15	0.16	0.63	0.01	0.15	0.15	-1.12	-1.13	-1.14	-0.02	0.13	0.13	-1.12	-0.01	0.11	0.11
2	0.78	0.77	0.82	0.05	0.15	0.16	0.78	0.01	0.15	0.15	-0.62	-1.13	-0.65	-0.03	0.10	0.11	-0.62	0.00	0.13	0.13
3	0.36	0.37	0.37	0.01	0.09	0.09	0.36	0.00	0.10	0.10	0.42	0.02	0.42	0.00	0.09	0.09	0.42	0.00	0.09	0.09
4	1.07	0.86	1.05	-0.02	0.18	0.18	1.11	0.04	0.21	0.22	-0.43	-0.44	-0.42	0.01	0.11	0.11	-0.44	-0.02	0.10	0.10
5	0.53	0.67	0.58	0.05	0.11	0.12	0.55	0.02	0.12	0.12	0.07	0.09	0.08	0.01	0.09	0.09	0.07	-0.01	0.09	0.09
6	0.60	0.86	0.62	0.03	0.12	0.12	0.59	0.00	0.13	0.13	-0.82	-0.44	-0.82	0.00	0.09	0.09	-0.82	0.00	0.10	0.10
7	0.86	0.67	0.84	-0.02	0.16	0.16	0.89	0.02	0.17	0.17	0.38	0.09	0.36	-0.02	0.09	0.09	0.40	0.01	0.12	0.12
8	0.51	0.51	0.51	0.00	0.13	0.13	0.53	0.02	0.16	0.16	-1.15	-0.68	-1.14	0.01	0.12	0.12	-1.16	-0.01	0.12	0.12
9	0.30	0.33	0.27	-0.03	0.13	0.13	0.32	0.02	0.17	0.17	1.71	0.53	1.67	-0.04	0.11	0.12	1.76	0.06	0.18	0.18
10	0.35	0.67	0.37	0.02	0.11	0.11	0.35	0.01	0.11	0.11	0.45	0.09	0.43	-0.01	0.08	0.08	0.45	0.00	0.09	0.09

Table 5. Estimates of item parameters for *k*=10 items and sample size *n*=500.

			Empirical				Classical						Empirical				Classical			
Item	True α	$\tilde{\alpha}_N$	$\hat{\alpha}$	Bias	Sd	RMSE	$\hat{\alpha}$	Bias	Sd	RMSE	True δ	$\tilde{\delta}_N$	$\hat{\delta}$	Bias	Sd	RMSE	$\hat{\delta}$	Bias	Sd	RMSE
1	0.62	0.77	0.64	0.02	0.10	0.10	0.62	0.00	0.12	0.12	-1.12	-1.13	-1.14	-0.02	0.09	0.09	-1.12	-0.01	0.10	0.10
2	0.78	0.77	0.80	0.02	0.13	0.13	0.79	0.02	0.12	0.12	-0.62	-1.13	-0.64	-0.02	0.08	0.08	-0.63	-0.01	0.08	0.08
3	0.36	0.37	0.37	0.01	0.08	0.08	0.38	0.01	0.08	0.08	0.42	0.02	0.41	-0.01	0.06	0.06	0.43	0.01	0.06	0.06
4	1.07	0.86	1.10	0.03	0.15	0.15	1.08	0.01	0.17	0.17	-0.43	-0.44	-0.44	-0.01	0.08	0.08	-0.45	-0.02	0.09	0.09
5	0.53	0.67	0.54	0.01	0.08	0.08	0.52	-0.01	0.10	0.10	0.07	0.09	0.07	0.00	0.06	0.06	0.07	0.00	0.07	0.07
6	0.60	0.86	0.61	0.01	0.10	0.10	0.58	-0.02	0.10	0.10	-0.82	-0.44	-0.81	0.01	0.07	0.07	-0.81	0.01	0.07	0.07
7	0.86	0.67	0.86	-0.01	0.12	0.12	0.89	0.03	0.14	0.14	0.38	0.09	0.37	-0.01	0.08	0.08	0.39	0.00	0.08	0.08
8	0.51	0.51	0.51	0.00	0.11	0.11	0.51	0.00	0.10	0.10	-1.15	-0.68	-1.16	-0.01	0.09	0.09	-1.15	0.00	0.09	0.09
9	0.30	0.33	0.26	-0.04	0.11	0.12	0.30	0.00	0.13	0.13	1.71	0.53	1.68	-0.03	0.09	0.09	1.74	0.03	0.13	0.13
10	0.35	0.67	0.37	0.03	0.09	0.10	0.35	0.01	0.09	0.09	0.45	0.09	0.45	0.01	0.07	0.07	0.44	0.00	0.07	0.07

Table 6. Median bias and median RMSE for discrimination parameters under different conditions.

		Median bias		Median RMSE	
k	*n*	Empirical	Classical	Empirical	Classical
10	100	0.07	0.04	0.19	0.25
	200	0.04	0.02	0.16	0.18
	300	0.02	0.01	0.13	0.15
	500	0.02	0.01	0.11	0.11
20	100	0.04	0.03	0.19	0.22
	200	0.04	0.02	0.15	0.16
	300	0.02	0.01	0.12	0.13
	500	0.01	0.01	0.10	0.09
30	100	0.03	0.02	0.19	0.21
	200	0.03	0.02	0.15	0.15
	300	0.01	0.01	0.12	0.12
	500	0.01	0.01	0.09	0.09

Table 7. Median bias and median RMSE for difficulty parameters under different conditions.

		Median bias		Median RMSE	
k	*n*	Empirical	Classical	Empirical	Classical
10	100	0.02	0.03	0.16	0.19
	200	0.01	0.01	0.12	0.12
	300	0.01	0.01	0.10	0.11
	500	0.01	0.01	0.08	0.08
20	100	0.03	0.01	0.18	0.18
	200	0.01	0.01	0.12	0.13
	300	0.02	0.01	0.10	0.10
	500	0.01	0.01	0.07	0.08
30	100	0.02	0.01	0.16	0.18
	200	0.01	0.01	0.12	0.13
	300	0.01	0.01	0.09	0.10
	500	0.01	0.01	0.08	0.08

Table 8. Number of items needed to complete CAT in two different pools.

		Pool 1			Pool 2	
True θ		Mean	Sd		Mean	Sd
-2		8.39	3.2		11.09	2.29
-1.8		6.9	2.9		10.24	1.60
-1.6		5.7	2.3		9.59	0.95
-1.4		4.8	1.7		9.37	0.74
-1.2		4.21	1.3		9.11	0.66
-1		3.56	1.2		9.03	0.48
-0.8		3.02	0.8		8.88	0.53
-0.6		2.86	0.7		9.10	0.82
-0.4		2.74	0.8		9.37	1.32
-0.2		2.77	0.8		9.57	1.34
0		3.21	1		10.39	2.13
0.2		3.56	1.3		11.52	2.61
0.4		4.48	1.7		12.30	3.22
0.6		5.05	2.1		13.89	3.56
0.8		5.98	2.7		16.05	4.99
1		7.41	3.5		17.93	5.11
1.2		8.24	4.2		19.85	6.75
1.4		10.27	4.6		24.28	10.30
1.6		14.29	6.8		26.25	9.49
1.8		15.92	7.3		30.24	11.89
2		19.39	8.5		37.06	12.88

Figure 1a. Box-plot for the item discrimination parameter.

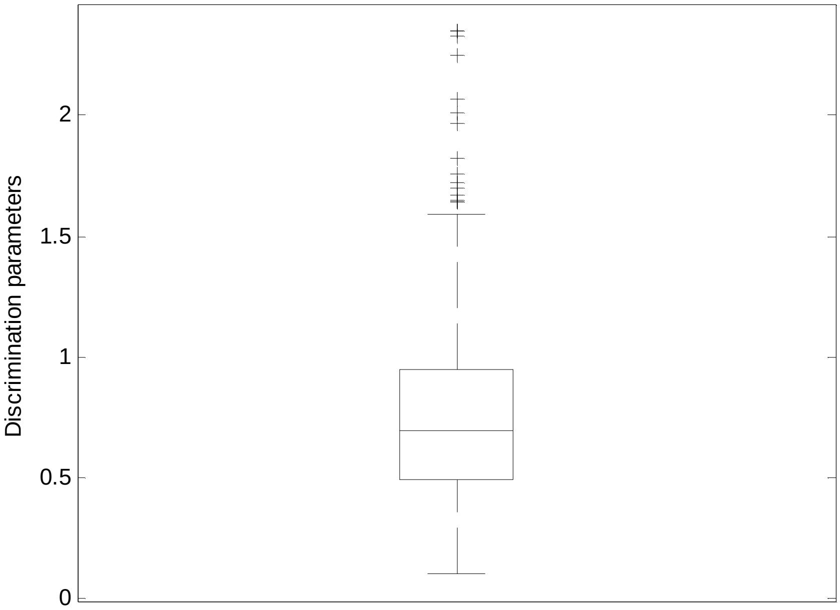


Figure 1b. Box-plot for the item difficulty parameter.

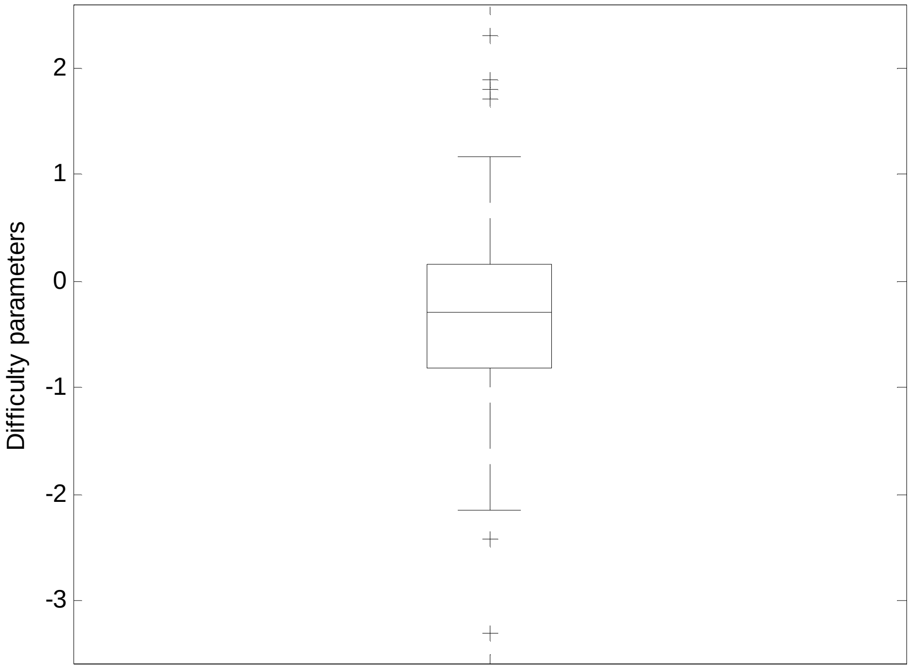


Figure 2. Scatterplot of item parameters.

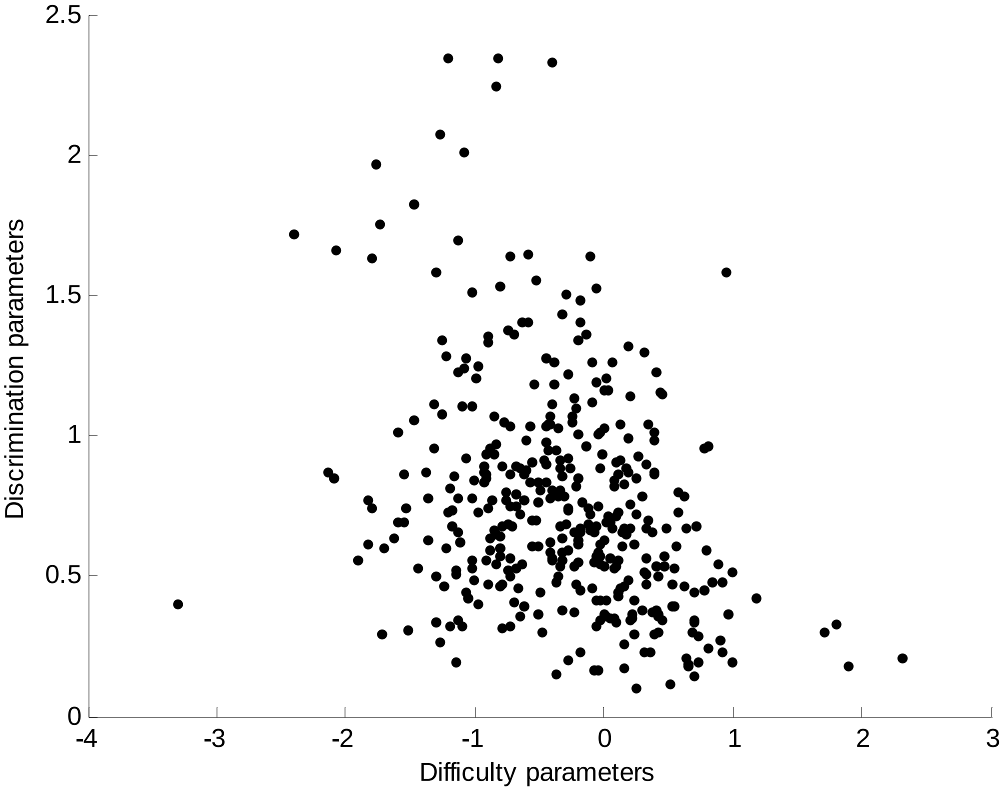


Figure 3. Tree relative error for different tree sizes and complexity parameters (cp).

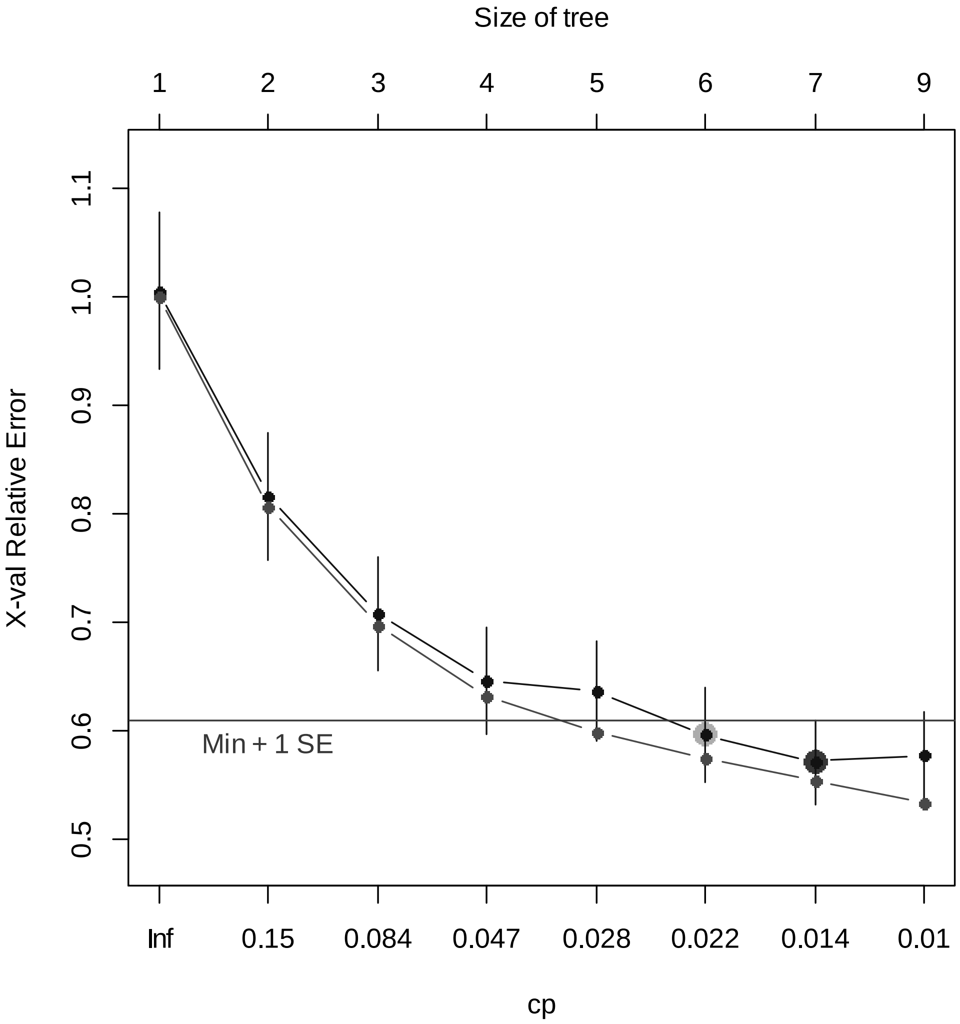
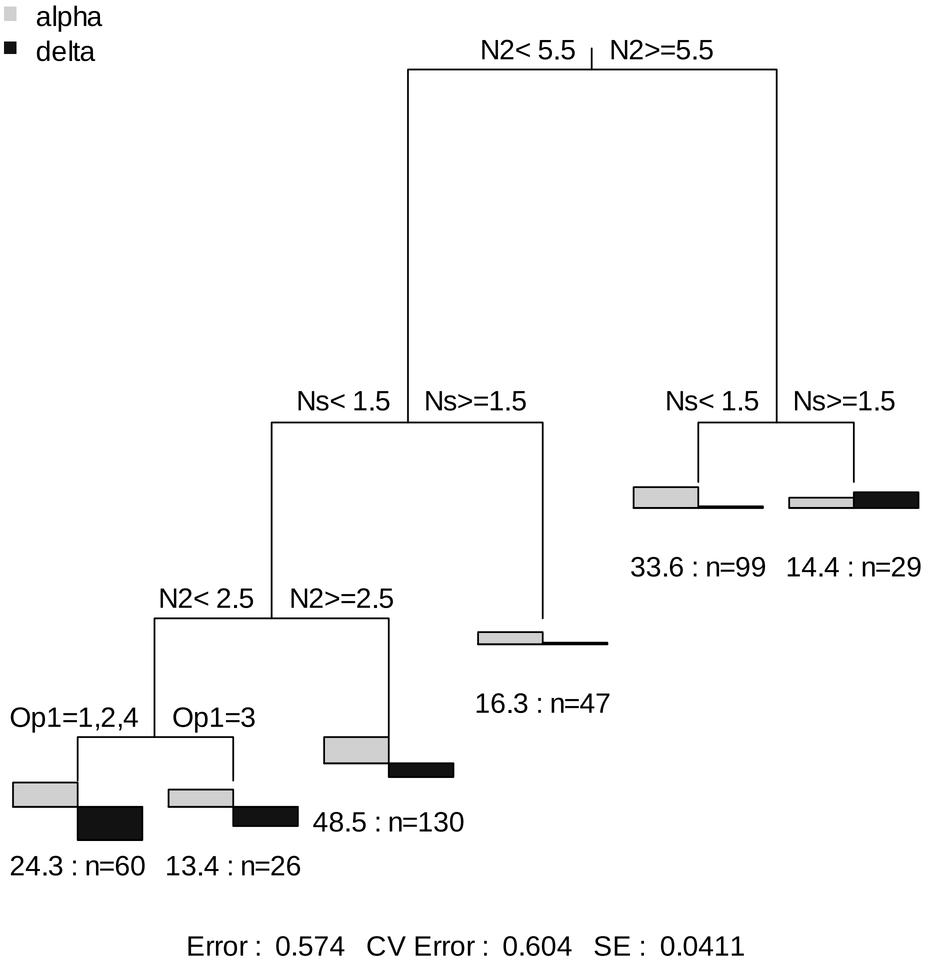


Figure 4. Fitted tree for the item parameters.


■ alpha
■ delta
N2< 5.5
N2>=5.5
Ns< 1.5
Ns>=1.5
Ns< 1.5
Ns>=1.5
33.6 : n=99
14.4 : n=29
N2< 2.5
N2>=2.5
16.3 : n=47
Op1=1,2,4
Op1=3
48.5 : n=130
24.3 : n=60
13.4 : n=26
Error : 0.574 CV Error : 0.604 SE : 0.0411