Ensaio: Avaliação e Políticas Públicas em Educação
ISSN: 0104-4036
ensaio@cesgranrio.org.br
Fundação Cesgranrio
Brasil

Veldkamp, Bernard P.; Matteucci, Mariagiulia
Bayesian Computerized Adaptive Testing
Ensaio: Avaliação e Políticas Públicas em Educação, vol. 21, núm. 78, enero-marzo, 2013, pp. 57-81
Fundação Cesgranrio
Rio de Janeiro, Brasil

Bayesian Computerized Adaptive Testing

Bernard P. Veldkamp*
Mariagiulia Matteucci**

Abstract

Computerized adaptive testing (CAT) comes with many advantages. Unfortunately, it still is quite expensive to develop and maintain an operational CAT. In this paper, various steps involved in developing an operational CAT are described and literature on these topics is reviewed. Bayesian CAT is introduced as an alternative, and the use of empirical priors is proposed for estimating item and person parameters to reduce the costs of CAT. Methods to elicit empirical priors are presented and a two small examples are presented that illustrate the advantages of Bayesian CAT. Implications of the use of empirical priors are discussed, limitations are mentioned and some suggestions for further research are formulated.
Keywords: Bayesian IRT modeling. Computerized Adaptive Testing. Eliciting priors. Item Response Theory. Item selection. Parameter estimation.

Teste Adaptativo Computadorizado Bayesiano Resumo

O teste adaptativo Computadorizado (CAT) chega com muitas vantagens. Infelizmente, ainda é bastante caro para desenvolver e manter um CAT operacional. Neste artigo, descreve-se várias etapas envolvidas no desenvolvimento de um CAT operacional e faz-se uma revisão da literatura nesse tópico. O CAT Bayesiano é introduzido como uma alternativa, e propõe-se o uso de prioris empíricas para estimar parâmetros de itens e de indivíduos com o objetivo de reduzir os custos de CAT. Apresenta-se métodos para obtenção de prioris empíricas e dois pequenos exemplos para ilustrar a vantagem do CAT Bayesiano. Discute-se algumas implicações no uso de prioris empíricas, menciona-se limitações e formula-se algumas sugestões para novas pesquisas.
Palavras-chave: *Modelagem Bayesiana da TRI. Teste Adaptativo Computadorizado. Obtenção de priors. Teoria de Resposta ao Item. Seleção de itens. Estimação de parâmetros.*

* Director of the Research Center for Examination and Certification (RCEC) /University of Twente, The Netherlands. *E-mail:* b.p.veldkamp@gw.utwente.nl
** Department of Statistical Sciences, University of Bologna, Italy. *E-mail:* m.matteucci@unibo.it

Test Adaptativo Computadorizado Bayesiano

Resumen

El test adaptativo Computadorizado (CAT) tiene muchas ventajas. Aunque, infelizmente, es bastante caro desarrollar y mantener un CAT operacional. En este artículo se describen varias etapas de su desarrollo y se hace una revisión de literatura del tópico. El CAT Bayesiano aparece como una alternativa, y se propone el uso de prioris empíricas para estimar parámetros de ítems y de individuos con el objeto de reducir sus costos. Se presentan métodos para obtener prioris empíricas y dos pequeños ejemplos que ilustran la ventaja del CAT Bayesiano. Se discuten algunas implicaciones en el uso de prioris empíricas, se mencionan limitaciones y se formulan sugerencias para nuevas investigaciones.

Palabras clave: *Modelo Bayesiano de la TRI. Test Adaptativo Computadorizado. Obtención de prioris. Teoría de Respuesta al Ítem. Selección de ítems. Estimación de parámetros.*

Introduction

In computerized adaptive testing (CAT) the difficulty of the items is adapted to the performance level of the candidate. In this way, more information can be obtained about the level of the candidate by administering fewer items. Reductions in test length up to 40% can be obtained without any loss of measurement precision. Besides, candidates will not get bored by items that are too easy or too hard, and in this digital age, candidates often like computer-based testing and prefer CAT above a paper and pencil testing. Weiss (1973) presented one of the first CATs. Later on many large scale tests, like for example the Armed Services Vocational Aptitude Battery (ASVAB);(SANDS;WATERS;MCBRIDE,1997), the Graduate Management Admission Test (GMAT);(RUDNER, 2010), or the MathCAT (VERSCHOOR; STRAETMANS, 2010), were administered adaptively.

Despite the obvious advantages of adaptive testing, there were some misconceptions and some drawbacks when CAT was first implemented. First of all, developing a CAT turned out to be rather expensive, because of the amount of items that had to be written and pre-tested. An item bank needed to consist of enough items with various difficulty levels to enable the algorithm to select items with difficulty level close to the estimated ability level of the candidate. This implied considerable costs of item development, pre-testing and item calibration. Besides, in the early years of CAT, it was believed that CAT enabled continuous testing. Candidates could log on and do the test whenever they felt ready. Since the content of the test was adapted to the ability level of the candidate, chances of answer copying were assumed to be negligible. Besides, the probability of two candidates having the same test was controlled by the application of exposure control methods (SYMPSON; HETTER, 1985). Unfortunately, organized attempts were conducted to

crack the item bank by memorizing the items and publishing them on the web. For several high-stakes tests the content of the items was revealed within days after the test became operational.

As a result, the popularity of CAT for educational measurement decreased. Even thought it was realized that problems related to item bank compromise, were problems of continuous testing rather than of CAT, the problem of expensive item banks still existed. These costs could either be reduced by decreasing the test length or by reducing the costs of item development. Reducing test length would result in less informative tests and in larger measurement errors. Decreasing expenses on item writing would imply less money for writing, pre-testing, and calibrating the new items. This would result in more uncertainty in the quality of the items. Both effects are unwanted.

In this paper, we propose the use of empirical priors to decrease costs involved in CAT. We introduce methods for eliciting empirical priors based on covariates about the candidates and the items to increase the efficiency of CAT. First, a general framework for CAT is presented and the various steps of CAT are introduced more into detail. After that, a Bayesian model for CAT and procedures for eliciting and implementing empirical priors in CAT are presented. Some results are presented. Finally, implications of applying empirical priors in CAT are discussed, and some topics for further research are mentioned.

Computerized Adaptive Testing

Before a CAT can be administered, a whole system of testing has to be designed. Sands, Waters e Mcbride (1997), Wainer et al. (2000), and Van der Linden ; Glas (2010) describe various aspects involved in developing and implementing CAT.

As a start, test specifications have to be formulated. During this process, many questions have to be answered. In this paragraph, we just mention a few of them. What is the purpose of the test? Is it for classification, for mastery decisions or for proficiency estimation? Is it a fixed- or a variable length test? Are there any content- or other type of specifications that have to be met? Will the test be administered via the web, or at specific testing locations? Will the test be administered continuously or only during specific time slots? How many candidates will do the test? What types of items will be used? Are there any groups of candidates with specific needs that have to be accounted for? It generally takes quite some time to get a complete picture of the operational test. However, it really pays off to invest time and resources in designing the framework.

The next step is to choose a measurement framework. In CAT, item response theory (IRT) models (LORD, 1980), are applied to formulate the relation between the

observed responses and the underlying abilities of the candidate. IRT distinguishes the item parameters from the person parameters, which is a very convenient property. The item parameters can be estimated separately during item pre-testing. The calibrated items can be stored in an item bank, and during test administration some of them can be selected adaptively to estimate the person parameters. Logistic IRT models are most commonly applied for scoring dichotomous items. The 3-parameter logistic model (3PLM) can be formulated as:

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \qquad (1)$$

where θ_j denotes the ability of person *j*, and (a_i, b_i, c_i) denote the discrimination, difficulty, and guessing parameter of item *i*. In the 2-parameter logistic model (2PLM), the guessing parameter is assumed to be zero. Finally, in the Rasch model or the 1-parameter logistic model (1PLM), the discrimination parameters of all items are assumed to be equal to each other as well. Besides, many models for scoring polytomous items (OSTINI; NERING, 2006) have been presented in the literature. When several abilities account for the response behavior, multidimensional IRT models can be applied (SEGALL, 1996; VELDKAMP; VAN DER LINDEN, 2002; RECKASE, 2009).

Once test specifications have been written and an IRT model is selected, the item bank can be developed. Sometimes, the item bank is developed from scratch and a blueprint can be developed first (VELDKAMP; VAN DER LINDEN, 2000) to guide the item writing process. In other applications, a previous set of items might be available that could be used to develop an item bank (VAN DER LINDEN; ARIEL; VELDKAMP, 2006). For many testing programs a distinction is made between a master pool and operational item pools. This master pool contains all the items that are available for the testing program. New items are added regularly, and old items (temporarily) retire when they have been exposed too often or when their content is not up to date anymore. Operational item banks are selected from this master pool (ARIEL;VAN DER LINDEN; VELDKAMP, 2006) and in some applications several parallel operation item banks can be selected to rotate over locations and time (WAY; STEFFEN; ANDERSON, 1998, ARIEL; VELDKAMP; VAN DER LINDEN, 2004). As a rule of thumb, Stocking (1994) suggested the number of items in the bank to be roughly equal to twelve times the test length. Once the items are available they can be pre-tested and calibrated. During pre-testing the items are administered and their item parameters are estimated, either with commercial software packages, like for example BILOG-MG3 (ZIMOWSKI et al., 2003), or with one of the non-commercial software packages that are available for estimating IRT models. The size of the pre-testing sample depends on the IRT model. The more parameters in the model, the larger the size of the sample has to be. For a 2-parameter IRT model, a sample size of 500 or more is often

suggested, while for a 3-parameter model it is generally recommended to have a sample size of at least 1000 candidates per item. To pre-test a large item bank of hundreds of items, it is often too demanding to administer all items to each candidate. Instead a linked design (SCHEERENS; GLAS; THOMAS, 2003, Chap. 8) might be applied where all candidates respond to a subset of items and the various subsets of items overlap. To calibrate the whole item bank, a pre-test sample of thousands of candidates might be needed. Model fit statistics are available to check the psychometric quality of the items and to decide whether the items can be added to the item bank, or whether they have to be revised first (GLAS, 1988). It often happens that only half of the items show desirable psychometric properties, while the other half has to be rejected. This number has to be taken into account when the item bank is developed. The result of the item bank development step is an operational bank with a well-balanced content and a distribution of item difficulties such that there is always at least one item available at the performance level of the candidate.

1. The actual administration of CAT consists of five basic steps:
2. Initiation of ability estimate
3. Selection of subsequent item
4. Administration of the item
5. Updating the ability estimate
6. Checking whether the stopping criterion has been met

They are being dealt with in more detail in the next section.

A final issue is related to the platform that is applied to administer the CAT. Various commercial software packages are available (for an overview, see www.iacat.org, Resources). besides, many testing agencies develop their own tailor made CAT software, to be able to meet their own specific needs.

The whole process of developing an operational CAT generally takes several years. Even though the theoretical framework is there and software for administering the test is available, there are still many decisions that have to be made. The item bank has to be developed with care, since the quality of the bank determines the quality of the CATs. Besides, thorough field testing is recommended not to run into nasty surprises once the test is operational.

Five basic steps of CAT

The previous paragraph already introduced the five basic steps of CAT. They are being dealt with more into detail in this section of the paper to provide more insight into the specificities of CAT. Some steps are rather straightforward to implement, but especially the step of item selection entails many issues.

Step 1. Initiation

In Step 1, an initial estimate is made of the proficiency level of the candidate. Generally, the ability level is initialized at the mean of the proficiency distribution of the population (THISSEN; MISLEVY, 2000). An alternative is to randomly draw it from the ability distribution. Besides, initialization based on previously known information about the candidate might be applied as well (VAN DER LINDEN, 1999).

Step 2. Item selection

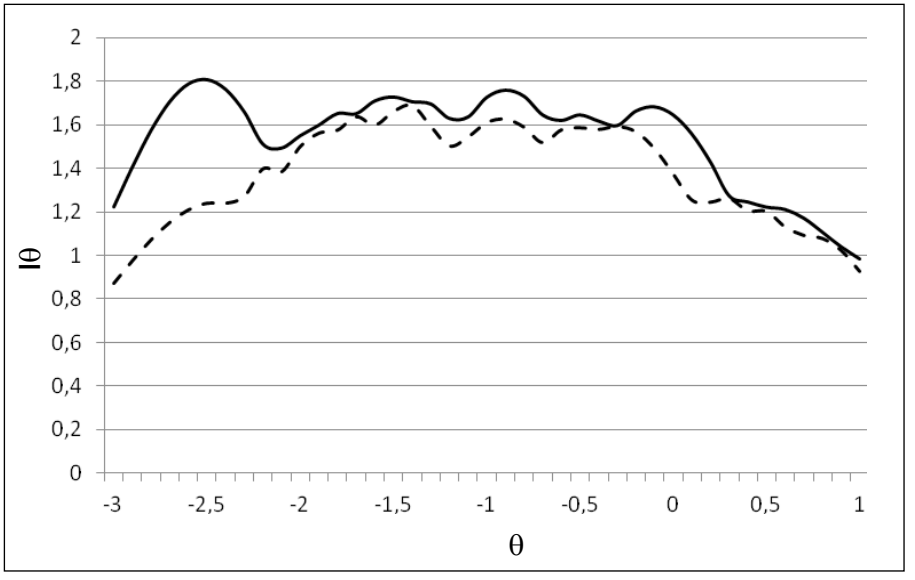
Many item selection rules have been proposed for CAT. Maximum Fisher information (BIRNBAUM, 1968) is most commonly applied, but Fisher interval information (VEERKAMP; BERGER,1997), Kullback-Leibler information (CHANG; YING, 1996; VELDKAMP; VAN DER LINDEN, 2002), or mutual information (WEISSMAN, 2007) might be applied as well. All these item selection rules have in common that they try to maximize information obtained about the candidate in order to minimize the error of estimation. Chang e Ying (1999), on the other hand, observed that during the early stages of CAT administration, the ability estimate is not very precise yet. They reasoned that selection of very informative items at an uncertain ability estimates might not be optimal in practice. As an alternative, they proposed alpha-stratified CAT, where the item bank is stratified with respect to the discrimination parameter. Items with lower discrimination parameters have flatter item information curves. By selecting items with lower discrimination parameters during the early stages of CAT, the items will provide a comparable amount of information irrespective of the true ability of the candidate. When the estimated ability level is more precise, items from higher discrimination strata can be selected. Over the past ten years, many comparison studies have been carried out to find the best item selection rule. No overall winner has been found. Most of the item selection rules perform rather well when twenty or more items are being selected for the test.

During the second step, test specifications have to be taken into account as well. These specifications can be related to the content of the test, they can be about time limits, or about the distribution of answer keys. They can also be about the word count or about items excluding each other from the same test when one item contains clues to the other. Specifications can be about the psychometric properties of the test, or about technical issues, like a minimum number of items to be selected for a text passage or a graph. For an overview of various types of specifications, see Van der Linden (2005, chap. 2). Kingsbury e Zara (1998) proposed to stratify the item bank with respect to, for example, content classifications, and to rotate item selection over the various strata. When a limited number of specifications have to be met, this approach might

work well. For testing programs where large numbers of specifications have be formulated, this approach can become intractable. Stocking e Swanson (1993) introduced a Weighted Deviation Model, where targets were set for various specifications and the weighted deviation from these targets was minimized, and Luecht (1998) developed a Normalized Weighted Absolute Deviation Heuristic. However, both of these methods cannot guarantee that the final CAT will meet all specifications. As an alternative, Van der Linden e Reese (1998) proposed the shadow test approach, a 2-stage procedure for item selection where 0-1 linear programming techniques are applied to make sure that all specifications will be met. During the first stage, a full-length test is constructed (the shadow test) that performs optimal with respect to the item selection rule at the current ability estimate and meets all the specifications. During the 2nd stage, the best unadministered item is selected from the shadow test to be presented to the candidate. For an extensive description of the shadow test approach, see also Van der Linden (2005, Chap. 9).

Another issue that has to be mentioned is exposure control. When maximum Fisher information is applied for selecting the next item, only those items that perform optimally with respect to this criterion will be selected. Typically, 20 percent of the items in the bank are selected for administration, while 80 percent are not selected at all. The same pattern can be found when any of the other selection rules that maximize some kind of information criterion is applied. The best items in the pool are over exposed, while the other items are hardly exposed during test administration. Van der Linden and Veldkamp (2007) studied this phenomenon more into detail and they found that for an operational item bank, only a few items will be maximally informative over the whole ability range. We repeated their analyses for an operational bank of 499 items. The items bank was part of an intelligence testing battery. To calibrate these items, they had been administered to a pre-testing sample of 3000 candidates. These candidates represented the Dutch labor force. Each of the candidates answered a subset of items in the bank. Bilog MG (ZIMOWSKI et al.,1996) had been applied to calibrate the items with the 2PLM. The curves of the most informative and second most informative items for every ability level are shown in Figure 1.

Figure 1 - Curve of most informative (solid line) and second most informative (dashed line) items



Source: Authors (2012).

In our analyses, only 12 out of 499 items were maximally informative at any ability level. These are the items that will be selected when Fisher information is optimized during item selection. It is obvious that their exposure rates will be high. The same phenomenon can be observed for the group of items that are 2nd most informative or 3rd most informative. Since only 26 items had to be selected for CAT, it demonstrates why only 20 percent of the items in the bank are actually administered. This has some undesired consequences. The items with highest exposure rates might become known to the candidates, which implies a test security problem, and it might compromise the testing results. Besides this risk, there is also the loss of investments. Considerable efforts and money have been put in writing and pre-testing items that are not selected for CAT. To deal with both the problems of overexposure and underexposure of items in the bank, exposure control methods have been proposed. The most famous exposure control method has been proposed by Sympson e Hetter (1985). In their method, they conduct a probability experiment after an item has been selected. In fact, they add this as Step 2b to the pseudo algorithm for CAT. In this experiment, the probability of being administered after being selected depends on the popularity of the item. In an extensive simulations study, the probabilities are set in such a way that for all items the expected exposure rate is smaller than the maximum exposure rate (generally set at $r_{\max} = 0.20$ or $r_{\max} = 0.25$) allowed. The more popular items have a small

probability of passing this hurdle and the less popular items have a probability of being administered close to 1. Whenever a selected item does not pass the hurdle, a next item (Step 2) is selected until an item passes the Sympson Hetter probability experiment and it can be administered to the candidate. Many modifications of the Sympson Hetter method have been proposed. For example, Stocking e Lewis (1998) observed that within certain ability groups, the same items were administered even though the Sympson Hetter method was applied. So, even though the overall exposure rate was below, within certain ranges of ability values, it was high. Therefore they proposed to modify the Sympson Hetter approach and to make it conditional on the ability estimates. In this way, the problem could be dealt with at the costs of even more extensive simulation studies. Van der Linden and Veldkamp (2004, 2007) proposed an entirely different approach. They developed an exposure control method that did not need any simulations, but that was based on observed exposure rates instead. In their item eligibility method, a probability experiment is carried out for every item in the bank with exposure rate higher than the maximum exposure rate imposed. In this probability experiment it is determined whether the item is eligible for administration, that is, whether it is included in the sub item bank from which the subsequent item is selected or not. The probability of being eligible depends on the ratio of the observed exposure rate and the maximum exposure rate. Barrada, Abad e Veldkamp (2009) compared both methods and found a slightly better performance for the eligibility method. Besides, Barrada, Olea e Veldkamp (2009) observed that it would pay off to randomize item selection at the beginning of CAT and to save the more informative items towards the end when the estimated ability is close to the true ability of the candidates. Besides, towards the end of the adaptive test, the candidates are spread over the whole ability continuum and over exposure is less of an issue. Because of this, multiple maximum exposure rates can be applied, depending on the position of the item in the test. They were able to demonstrate that a more balanced usage of the item bank could be obtained this way, with hardly any decrease in measurement precision. All of these methods focus on the control of over exposure, and it is generally assumed that underexposure problems will reduce when the exposure of the most informative items is limited. In practice, exposure control methods only increased the use of a subset of the under exposed item. Revuelta e Ponsoda (1998) addressed this issue. They proposed the progressive or restricted method for item selection where item selection is based both on the information provided by the item and on a randomized component. During the early stages of CAT the randomized component is more important. Towards the end, the information component is more important. By partly randomizing item selection, a more evenly distributed exposure was obtained. The alphas-stratified method (CHANG; YING, 1999) also deals with

underexposure. During the early stages of CAT, items are selected from strata with lower discrimination indices. These items are less informative, and tend to have low exposure rates. Limiting item selection to these strata increases the exposure rates of these items and reduces the problems of underexposure. The best results with exposure control problems have been obtained when methods for dealing with over exposure were combined with methods that deal with underexposure. Veldkamp, Verschoor e Eggen (2010) therefore proposed a method that combined both approaches.

Item selection in CAT has been an important research topic for many years, also because the adaptive item selection process is what makes CAT different from administering linear test forms. This paragraph only covers a small part of the literature on it. Most of the existing papers are about CAT with dichotomously scored items that are calibrated with a unidimensional IRT model. Nowadays, the focus is shifting more towards developing methods for CAT with more complex item types that are calibrated with more complicated, often multidimensional, IRT models. Even though impressive results have been obtained, there are still many more areas that have to be studied.

Step 3. Administration

In the third step the item is presented to the candidate. The mode of presentation has to be robust against the use of various computer platforms and various types of monitors. It has to be guaranteed that every candidate can respond to the item based on the same amount of information without any distraction. After presenting the item, a candidate either has limited or unlimited time to answer the item. When response times are limited, one might consider the use of response time models (VAN DER LINDEN, 2007) to correct for speededness of the test. In some CATs, several items are presented on the same page, for example, when they are all related to the same stimulus. In most CATs however, only one item is presented at a time. It is important to realize that as a result of selecting each subsequent item based on information obtained in previous items, in most CATs it is not allowed to review earlier responses. Allowing item review would reduce measurement efficiency and would make CAT vulnerable to test taking strategies (WAINER, 1993), besides the assumption of local independence would be violated, that states that the observed responses are independent of each other given an individual's score on the latent trait. Bowles e Pommerich (2001) studied the impact of item review and found only limited effects on bias and root mean squared error of the ability estimates.

Step 4. Ability estimation

Ability estimation in CAT is very much comparable to ability estimation in paper and pencil testing. For every response pattern $(u_1, u_2, ..., u_g)$ where u_i denotes

whether item i is answered correctly $(u_i = 1)$ or not $(u_i = 0)$ and g is the number of items administered, a likelihood can be defined as:

$$L(\mathrm{u} \mid \theta) = \prod_{i=1}^{g} P_i(\theta)^{u_i} (1 - P(\theta))^{1-u_i}. \qquad (2)$$

Since the items have been calibrated, the item parameters are assumed to be known. A Gaussian quadrature procedure can be applied to obtain maximum likelihood estimates of the ability parameter (ABRAMOWITZ; STENGUN, 1964). It should be noted that maximum likelihood estimates stay undetermined until a mixed response pattern is observed. As an alternative the Warm estimator (WARM, 1989) can be applied, where a weighted likelihood is maximized.

Step 5. Stopping rules

In CAT, the composition of the test is adapted to the performance of the candidate. Because of this, high and low performing candidates will answer different sets of items, and measurement precision might vary over candidates. To compensate, a CAT could be terminated when a pre-defined level of measurement precision is reached. In this way it is guaranteed that all candidates are measured with the same level of precision, even though some candidates might have to answer more questions than others. For some applications a variable length CAT might not be feasible, either because the content of the test is specified into detail, or because candidates might perceive a variable test length as unfair. For these applications, a CAT could be terminated after a fixed number of items. A third stopping rule, sometimes combined with either variable length or fixed length CAT, is to set a time limit for the whole test. For practical reasons this is very convenient, but one should be aware of the risks of making the test speeded, which might threat test validity.

A Bayesian framework for CAT

Both the item parameters and the ability parameters are estimated based on response patterns obtained during pre-testing (item parameter estimates) or during operational testing (ability estimates). No other information about the composition of the items or background information about the candidates is taken into account. Efficiency of CAT might increase when additional sources of information would be taken into account, and a Bayesian IRT framework can be applied to include additional information in CAT.

Bayesian IRT models

In educational and psychological measurement, we are generally interested in the distribution of the item and the person parameters given the observed response patterns. When Bayes Theorem (1763) is applied, the conditional probability of the

item and person parameters given the data can be modeled as a combination of prior beliefs about them and a parametric model about what the data should look like, conditional on the item and person parameter values:

$$p((\xi,\theta)\,|\,u) \propto p(\xi,\theta)p(u\,|\,\xi,\theta) \qquad (3)$$

where ξ denotes the item parameters (a_i,b_i,c_i) and θ the person or ability parameter. IRT models can be used to model the relationship between the observed data and the item and the person parameters in this framework. However, the interesting addition of Bayesian models is that information available about the items and the persons can be applied to elicit informative priors.

Instead of the logistic IRT models in Equation (1), normal ogive IRT models (LORD, 1952) are commonly applied within a Bayesian framework. The 3-parameter Normal Ogive (3PNO) model can be formulated as:

$$P(u_i = 1\,|\,a_i,b_i,c_i,\theta) = c_i + (1-c_i)\Phi(\alpha_i\theta - b_i) \qquad (4)$$

where

$$\Phi(a_i\theta - b_i) = \int_{-\infty}^{a_i\theta - b_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz. \qquad (5)$$

Both the 3PLM and the 3PNO define item characteristic curves that are identical up to a scaling constant $d = 1.702$. To estimate the item and ability parameters in this model a Gibbs sampler (GEMAN; GEMAN, 1984) might be applied. Among others, Albert (1992), Béguin e Glas (2001), Fox e Glas (2001), and Matteucci, Mignane e Veldkamp (2009) developed Gibbs samplers for various IRT models. Standardized software packages, like MATLAB or the R-package, can be applied for Bayesian parameter estimation. Albert (1992) provides MATLAB code for the 2PNO model and Fox (2010) provides R codes for several more complicated IRT models.

Eliciting and including priors for item parameters

Several methods have been proposed in the literature to predict item parameters based on item features. In item cloning, families of clones are derived from a parent item, by varying those attributes that are assumed not to be related to the item difficulty (BEJAR, 1993; GLAS; VAN DER LINDEN, 2003). Luecht (2009) proposed the assessment engineering approach, where items are generated based on construct maps that describe performance expectations at various levels of the scale. Sheehan (1997) introduced the application of Classification and Regression Trees (CART) (BREIMAN et al., 1984), to model the relationship between skills needed to solve the items and item difficulty. All of these methods have in common that they result

in an initial prediction of the item parameters with a certain level of uncertainty. Incorporating these predicted values in the parameter estimation process, might offer serious reductions in the costs of item bank development.

Including collateral information about persons in CAT

Various kinds of background variables about the abilities of the candidates might be available during testing. They might include socio-economic or demographical variables, but they might also be scores resulting from earlier tests. Several authors discussed the issue of including collateral information in CAT (ZWINDERMAN, 1991, 1997; VAN DER LINDEN, 1999, VAN DER LINDEN; PASHLEY, 2010; MATTEUCCI; MIGNANI; VELDKAMP, 2009). First of all, collateral information can be used to find an initial estimate of the ability of the candidate. When the initial ability estimate is close to the true ability of the candidate, only items that are informative will be selected, item exposure will be less of a problem, and the CAT will converge faster, which results in shorter, less expensive tests. Guyes (2008) motivated the use of more accurate initial ability estimates in CAT by demonstrating how a poor initial estimate might cause a very slow convergence of CAT. Besides, an informative prior could be used to formulate a posterior distribution of the ability parameter and expected a posteriori (EAP) estimation can be applied to obtain ability estimators. Bayesian item selection criteria have been developed and compared (OWEN, 1975; VAN DER LINDEN, 1998; VAN DER LINDEN; PASHLEY, 2010; VELDKAMP, 2010). Matteucci e Veldkamp (2012) even proposed to use a Gibbs sampler for ability estimation, for example, to handle applications where the abilities of the population are not normally distributed.

Empirical example

To illustrate the use of empirical information about both the items and the persons, data was analyzed from a computerized adaptive IQ test. The Connector Ability (MAIJ-DE MEIJ et al., 2008) is developed to measure IQ for application in the field of Human Resource Development, either during job applications or for career development. It consists of several subtests: Number Series, Figure Series, and Raven's Matrices. All of the items can be solved by applying a certain set of rules. For example, for the Number Series subtest, each item consists of a range of numbers and the candidates have to select the correct next number from a set of alternatives. Each item can be described by:

1. *Initial number in the interval [-10,10]*
2. *Operator at level 1 (addition, subtraction, multiplication, division)*
3. *Operand at level 1*
4. *Operator at level 2 (addition, subtraction, multiplication, division, none)*
5. *Operand at level 2*

To solve the item correctly, a candidate has to identify the features of the item and to apply them correctly to infer the next number in the series. For example, the item

$$1 \quad 4 \quad 10 \quad 22 \quad 46 \quad ?$$

can be decomposed as

$$1 \quad +3 \quad +6 \quad +12 \quad +24 \quad ? \quad \Leftrightarrow$$
$$1 \quad +3 \quad 2*(+3) \quad 2*(2*(+3)) \quad 2*(2*(2*(+3))) \quad ?$$

The initial number equals 1, the level 1 operator is addition, the level 1 operand equals 3, the level 2 operator is multiplication, and the level 2 operand equals 2. As a consequence, the next step would be $2*(2*(2*(2*(+3)))) = +48$, and the correct answer would be 94. In the Connector Ability, the difficulty of the Number Series items is increased even further by adding items that are a mix of two series. The odd positions in the range of these items belong to the first series, while the even positions belong to the second series. For all subtests it holds that the complete set of rules that has to be applied to solve the items is presented to the candidate in the introduction section of the test. Some operators are more difficult than others, and the size of the operands also influences the complexity of the items. In this example we use the features of the items as background information to elicit priors for the item parameters. The scores on one subtest were used as background information about the person, to elicit information priors for the person parameters.

Example 1. Item parameter estimation

Matteucci, Mignani,Veldkamp (2012) studied the use of informative priors in CAT. A pool of 391 Number Series items, that were calibrated with a 2PNO model, was available. Discrimination parameters were in the range of [0.10,2.35], with the median equal to 0.69. The difficulty parameters were in the range of [-3.30,2.30]. Regression trees were used to build a model that predicted the psychometric item parameters based on the item features using MATLAB 7.1 (MATHWORKS, 2005). The minimum node size was set equal to 10 items, the 1-SE rule was applied to choose the best tree, and 10-fold cross validation was applied. The resulting regression trees were used as informative priors in the process of item parameter estimation.

To simulate a real item calibration process, 20 items were randomly selected from the 391 item bank. Response patterns of 100 candidates were simulated as a calibration sample. In order to evaluate the parameter recovery by using different prior distributions for item parameters, a vague prior distribution for the item parameters, represented by the product of an indicator function ensuring positive discrimination parameters, i.e. $p(\xi) = \prod_{i=1}^{n} I(a_i > 0)$ (see ALBERT, 1992; FOX; GLAS, 2001; BAKER; KIM, 2004), was compared to the empirical prior distributions elicited using regression trees.

A Gibbs sampler (MATTEUCCI; MIGNANI; VELDKAMP, 2009) was used to estimate the parameters. For each simulation, 5000 iterations were used with a burn-in of 500 and 100 replications were conducted. The convergence of the algorithm was checked by calculating the Monte Carlo error as implemented in the R package BOA (SMITH, 2007). A rule of thumb is that the Monte Carlo error should be smaller than 5% of the standard deviation. All simulations were implemented in the software MATLAB 7.1 (MATHWORKS, 2005). 20 items were randomly chosen from a calibrated item bank. Based on their discrimination parameters they were classified as low discriminative (α<0.60), medium discriminative ($a \in [0.60, 1.00]$), and high discriminative ($a > 1.00$) items, and based on their difficulties as very easy ($b < -1.00$), easy ($b \in [-1.00, 0.00]$), moderate ($b \in (0.00, 1.00]$), and difficult ($b > 1.00$) items. These categorizations were used to compare the results of re-estimating the item parameters based on empirical priors based on regression trees on the one hand and the vague prior on the other hand. The results are shown in Tables 1 and 2.

Table 1 - Item discrimination recovery using different priors

	Vague prior			Empirical prior		
a_{TRUE}	$\hat{a}$	bias	RMSE	$\hat{a}$	bias	RMSE
Low	0.53	0.06	0.22	0.52	0.04	0.18
Medium	0.92	0.15	0.44	0.80	0.03	0.20
High	1.74	0.42	1.06	1.22	-0.11	0.25

Source: Authors (2012).

Table 2 - Item difficulty recovery using different priors

	Vague prior			Empirical prior		
b_{TRUE}	$\hat{b}$	bias	RMSE	$\hat{b}$	bias	RMSE
Very easy	-1.51	-0.21	0.63	-1.31	-0.02	0.20
Easy	-0.62	-0.06	0.30	-0.56	0.00	0.18
Moderate	0.17	0.02	0.16	0.12	-0.03	0.14
Difficult	0.89	0.07	0.20	0.81	-0.00	0.15

Source: Authors (2012).

Informative priors results in a more accurate item parameter recovery for both the parameters. Especially for medium ($a \in [0.60, 1.00]$) and high ($a > 1.00$) discriminating items, the use of empirical priors increased measurement precision

considerably. The average Root Mean Squared Error (RMSE) was reduced by more than fifty percent. This is especially important for CAT, because high discriminating items are selected more often to be administered. For the difficulty parameter, the effects were less drastic. Only for the very easy items $(b < -1.00)$, the use of empirical priors reduced RMSE more than fifty percent.

The lesson learned from this example is that one might reduce the sample size considerably when empirical priors are used, without any loss in measurement precision. This is one way to reduce the costs of CAT.

Example 2. Person parameter estimation

To study the impact of background variables in CAT, the scores of the Raven Matrices (RM) subscale of the Connector Ability were used as collateral information for the Number Series (NS) subtest. The method in Matteucci, Mignani e Veldkamp (2009) was applied to find the relationship between both subscales. It could be formulated as

$$\theta_{NS} \sim N(-0.243 + 0.394 \cdot \theta_{RM}, 0.414) \qquad (6)$$

where θ_{NS} and θ_{RM} represent the latent scores on the NS and RM subscales. Matteucci e Veldkamp (2012) studied the effects of the use of background information for person parameter in CAT. The NS CAT was simulated for 660 real candidates whose NS and RM scores were available. The NS item bank in this example consisted of 499 items calibrated with the 2PNO model. The Connector Ability is a variable length CAT where Standard Error (SE)<0.32 is used as a stopping rule for each subtest.

In a simulation study, the average test length for CAT where the empirical prior $\theta_{NS} \sim N(-0.243 + 0.394 \cdot \theta_{RM}, 0.414)$, based on Equation 6, for both initialization and ability estimation was compared to CAT with a non-informative prior $\theta_{NS} \sim N(0,1)$. For every candidate, the CAT was replicated 10 times to get reliable results. Based on the known abilities of 660 real candidates, more or less evenly distributed over the ability ranges {<-0.9;[-0.9,-0.6];(-0.6,-0.3];(-0.3,0.0];(0.0,0.3];>0.3}, answer patterns to the variable length CAT were simulated and the person parameters were re-estimated. The CAT was terminated when the SE < 0.32. For various groups of candidates the resulting test lengths are reported in Table 3.

Table 3 - Person parameter estimation using different priors

			Empirical prior	Non-Informative prior
Ability range	Number of items with difficulty in this range	Number of candidates in this range	Average number of items	Average number of items
< -0.9	136	106	10.27	11.02
[-0.9, -0.6]	64	97	9.14	9.48
[-0.6, -0.3]	78	132	9.09	9.20
[-0.3, 0.0]	78	123	9.32	9.50
[0.0, 0.3]	61	86	9.92	10.31
>0.3	82	116	13.78	15.38

Source: Authors (2012).

For those candidates with a true ability close to zero, the informative prior resulted in slightly shorter tests. For the candidates in the lowest and highest ability category however, considerable reduction in test length was obtained.

Empirical priors therefore, can be used successfully to reduce test length without any loss of measurement precision.

Discussion

In this paper, various aspects of CAT were introduced and reviewed. It was argued that even though CAT has some important advantages, the cost of development and maintenance are high. Much higher in general, then the costs of linear testing. To reduce the costs, Bayesian CAT was introduced. In Bayesian CAT, prior beliefs and observed data are combined to estimate both item and person parameters. It was demonstrated that, both in the item parameter estimation and in the person parameter estimation phase, considerable gains can be made by eliciting empirical priors for both the person and the item parameters, and implementing them in CAT. Bayesian CAT might therefore be an important future direction of CAT.

The quality of the information is of course very important. If the predictive power of the model is low, hardly any gains will be made. Moreover, as was also illustrated by Guyer (2008), inaccurate initialization of CAT will even result in longer and less informative tests. Another issue is related to the ethical implications of the use of empirical priors. When they are applied, each candidate is not only scored based on his/her responses, but background information is taken into account as

well. In medical applications it would be no problem to use available information about the patient to obtain more precise results of testing. But in high-stakes educational measurement, it would not be accepted. For those applications, it could be considered to use empirical information during CAT administration, but to report final scores based on response patterns only.

With respect to the item parameter estimation, the use of empirical information is less controversial. Recently the interest in automated item generation has grown considerably. Based on cloning models, CART models or Assessment Engineering models, psychometric item parameters can be predicted, and the time consuming expensive pre-testing phase might be skipped completely. Initially, predicted item parameters can be used to administer the test, and these parameters can be updated on-the-fly (MAKRANSKY, 2009). Of course, the uncertainty in the item parameters is considerable initially, which might result in over estimation of the information in the test (HAMBLETON; JONES, 1994). But recently, some papers have been written about taking the uncertainty in the item parameters into account during test assembly (VELDKAMP, 2012).

Finally, this paper dealt mainly with increasing the efficiency, just one of the fascinating aspects of CAT. Another important topic of research would be the use of more complex item types. At this moment, almost all operational CATs use either dichotomously or polytomously scored multiple-choice items. But since CATs are administered on a computer, more complex item types with constructed responses and dependencies between items might be developed. New IRT models will be needed to account for then. Besides, almost all operational CATs have been developed for assessment of learning. The first initiatives (EGGEN, 2011) have been taken to develop CATs for assessment for learning, where an adaptive algorithm is used to optimize the learning process instead of the measurement of learning outcomes. More research and practical work will be needed to explore all possibilities of CAT.

References

ABRAMOWITZ, M.; STEGUN, I. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Washington: Dover Publications, 1964.

ALBERT, J. H. Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics,* Washington, DC, v. 17, p. 251-269, 1992. DOI: http://dx.doi.org/10.3102/10769986017003251.

ARIEL, A.; VAN DER LINDEN, W.J.; VELDKAMP, B.P. A strategy for optimizing item pool management. *Journal of Educational Measurement,* Washington, DC, *v. 43, n. 2, p.* 85-96, 2006. DOI: http://dx.doi.org/10.1111/j.1745-3984.2006.00006.x.

ARIEL, A.; VELDKAMP, B.P.; VAN DER LINDEN, W. J. Constructing rotating item pools for constrained adaptive testing. *Journal of Educational Measurement,* Washington, DC, v. *41, p.* 345-360, 2004. DOI: http://dx.doi.org/10.1111/j.1745-3984.2004.tb01170.x.

BAKER, F. B.; KIM, S. H. *Item Response Theory Parameter Estimation Techniques.* New York: Marcel Dekker, 2004.

BARRADA, J. R.; ABAD, F.; VELDKAMP, B. P. Comparision of methods for controlling maximum esposure rates in computerized adaptive testing. *Psicothema,* Oviedo (Spain), v. *21,* p. 313-320, 2009.

BARRADA, J. R.; VELDKAMP, B. P.; OLEA, J. Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement*, Thousand Oaks, CA, v. 33, p. 58-73, 2009. DOI: http://dx.doi.org/10.1177/0146621608315329.

BAYES, T. An essay towards solving a problem in the doctrine of chances, communicated by M. Price in a letter to John Canton. *Phil. Trans. Royal Society London,* [S.l.], v. 53, p. 269-271, 1763.

BÉGUIN, A. A.; GLAS, C. A. W. MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, New York, v. 66, p. 541-562, 2001. DOI: http://dx.doi.org/10.1007/BF02296195.

BEJAR, I. I. A generative approach to psychological and educational measurement. In: FREDERIKSEN, N.; MISLEVY, R. J.; BEJAR, I. I. (Ed.). *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum, 1993. p. 323-357.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. In: LORD, F. M.; NOVICK, M.R. (Ed.). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968. p. 397-479.

BOWLES, R.; POMMERICH, M. *An Examination of Item Review on a CAT Using the Specific Information Item Selection Algorithm.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).

BREIMAN, L. et al. *Classification and Regression Trees*. Belmont, CA: Wadsworth International, 1984.

CHANG, H.; YING, Z. Global information approach to computerized adaptive testing. *Applied Psychological Measurement,* Thousand Oaks, CA, v. *20*, p. 213-229, 1996. DOI: http://dx.doi.org/10.1177/014662169602000303.

CHANG, H.; YING, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, Thousand Oaks, CA, v. 23, p. 211-222. DOI: http://dx.doi.org/10.1177/01466219922031338.

EGGEN, T. J. H. M. *What is the purpose of CAT.* Presidential address at the 2nd Conference of the International Association for Computerized Adaptive Testing, Pacific Grove, CA, October 4th, 2011.

FOX, J. P.; GLAS, C. A.W. Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, New York, v. 66, p. 271-288, 2001. DOI: http://dx.doi.org/10.1007/BF02294839.

FOX, J. P. *Bayesian item response modeling:* theory and applications. New York: Springer, 2010. DOI: http://dx.doi.org/10.1007/978-1-4419-0742-4.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, New York, v. 6, p. 721-741, 1984.

GLAS, C. A. W. The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, New York, v. 53, p. 525-546, 1988.

GLAS, C.A.W.; VAN DER LINDEN, W. J. Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, Thousand Oaks, CA, v. 27, p. 247-261, 2003. DOI: http://dx.doi.org/10.1177/0146621603254291.

GUYER, R. D. (2008). *Effects of early misfit in computerized adaptive testing on the recovery of theta.* 2008. Unpublished doctoral dissertation of the University of Minnesota (MN).

HAMBLETON, R. K.; JONES, R. W. Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education*, Lincoln, NE, v. 7, n. 3, p. 171-186, 1994.

KINGSBURY, G. G.; ZARA, A. R. Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, Lincoln, NE, v. 2, p. 359-375, 1989. DOI: http://dx.doi.org/10.1207/s15324818ame0204_6.

LORD, F. A. *Theory of Test Scores*. Richmond, VA: Psychometric Corporation, 1952. (Psychometric Monograph, n. 7).

LORD, F. M. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

LUECHT, R. M. Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement,* Thousand Oaks, CA, v. 22, p. 224-236, 1998. DOI: http://dx.doi.org/10.1177/01466216980223003.

LUECHT, R.M. Adaptive computer-based tasks under an assessment engineering paradigm. In: WEISS, D. J. (Ed.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing, 2009.* Available from: <www.psych.umn.edu/psylabs/CATCentral/>. Accessed: 05th April 2012.

MAIJ-DE MEIJ, A. M. et al. *Connector Ability; Professional Manual.* Utrecht, The Netherlands: PiCompany B. V., 2008.

MAKRANSKY, G. An automatic online calibration design in adaptive testing. In: WEISS, D. J. (Ed.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing,* 2009. Available from: <www.psych.umn.edu/psylabs/CATCentral/ >. Accessed: 05th May 2012.

MATHWORKS. *MATLAB 7.1:* user manual. Natick, MA: Math Works Inc., 2005.

MATTEUCCI, M.; MIGNANI, S.; VELDKAMP, B. P. Issues on item response theory modeling. In: BINI, M.; MONARI, P., PICCOLO, D.; SALMASO, L. (Ed.). *Statistical models for the evaluation of educational services and quality of products.* Berlin Heidelberg: Springer Verlag, 2009. p. 29-45.

MATTEUCCI, M.; MIGNANI, S.; VELDKAMP, B. P. Prior distributions for item parameters in IRT models. *Communications in Statistics, Theory and Methods,* Philadelphia, PA, v. 41, p. 2944-2958, 2012.

MATTEUCCI, M.; VELDKAMP, B. P. The use of MCMC CAT with empirical prior information to improve the efficiency of CAT. *Statistical Methods and Applications,* Heidelberg, *2012.* In press.

OSTINI, R.; NERING, M. L. *Polytomous Item Response Theory Models.* Thousand Oaks, CA: Sage Publications, 2006.

OWEN, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive testing. *Journal of the American Statistical Association,* [S. l.], v. 70, p. 351-356, 1975.

RECKASE, M. D. *Multidimensional item response theory.* New York: Springer, 2009. DOI: http://dx.doi.org/10.1007/978-0-387-89976-3.

REVUELTA, J.; PONSODA, V. A comparison of item-exposure control methods in computerized adaptive testing. *Journal of Educational Measurement,* Washington, DC, v. 38, p. 311-327, 1998. DOI: http://dx.doi.org/10.1111/j.1745-3984.1998.tb00541.x.

RUDNER, L. M. Implemention the Graduate Management Admission Test Computerized Adaptive Test. In: VAN DER LINDEN, W. J.; GLAS, C. A.W. (Ed.). *Elements of adaptive testing.* New York: Springer, 2010. p. 151-165. DOI: http://dx.doi.org/10.1007/978-0-387-85461-8_8.

SANDS, W. A.; WATERS, B. K.; MCBRIDE, J. R. *Computerized adaptive testing: from inquiry to operation*. Washington, DC: American Psychological Association, 1997. DOI: http://dx.doi.org/10.1037/10244-000

SCHEERENS, J.; GLAS, C. A.W.; THOMAS, S.A. *Educational evaluation, assessment, and monitoring:* a systemic approach. Lisse, The Netherlands: Swets and Zeitlinger, 2003.

SEGALL, D. O. Multidimensional adaptive testing. *Psychometrika,* New York, v. 61, p. 331-354, 1996. DOI: http://dx.doi.org/10.1007/BF02294343.

SHEEHAN, K. M. A tree based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement,* Washington, DC, v. 34, p. 333-352, 1997. DOI: http://dx.doi.org/10.1111/j.1745-3984.1997.tb00522.x.

SMITH, B. J. Boa: an R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software*, Los Angeles, CA, v. 21, p. 1-37, 2007.

STOCKING, M. L. *Three practical issues for modern adaptive item pools.* Princeton: Educational Testing Service, 1994. (Research Report 94-5).

STOCKING, M. L.; LEWIS, C. Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics,* Washington, DC, v. 23, p. 57-75, 1998. DOI: http://dx.doi.org/10.3102/10769986023001057.

STOCKING, M. L.; SWANSON, L. A method for severely constraint item selection in adaptive testing. *Applied Psychological Measurement,* Thousand Oaks, CA, v. 17, p. 277-292, 1993. DOI: http://dx.doi.org/10.1177/014662169301700308.

SYMPSON, J. B.; HETTER, R. D. Controlling item-exposure rates in computerized adaptive testing. In: *Proceedings of the 27th Annual Meeting of the Military Testing Association*. San Diego, CA: Navy Personnel Research and Development Center, 1985. p. 973–977.

THISSEN, D.; MISLEVY, R. J. Testing algorithms. In: WAINER et al. (Ed.). *Computerized Adaptive Testing:* a primer. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

VAN DER LINDEN, W. J. Empirical initialization of the trait estimation in adaptive testing. *Applied Psychological Measurement,* Thousand Oaks, CA, v. 23, p. 21-29, 1999. DOI: http://dx.doi.org/10.1177/01466219922031149.

VAN DER LINDEN, W. J. Bayesian item selection criteria for adaptive testing. *Psychometrika,* New York, v. 63, p. 201-216, 1998. DOI: http://dx.doi.org/10.1007/BF02294775.

VAN DER LINDEN, W. J. *Linear Models for Optimal Test Design.* New York: Springer, 2005.

VAN DER LINDEN, W. J. A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika,* New York, v. 72, p. 287-308, 2007. DOI: http://dx.doi.org/10.1007/s11336-006-1478-z .

VAN DER LINDEN; W. J., ARIEL; A.;VELDKAMP, B. P. Assembling a CAT pool as a set of linear tests. *Journal of Educational and Behavioral Statistics,* Washington, DC, v. 31, p. 81-100, 2006. DOI: http://dx.doi.org/10.3102/10769986031001081.

VAN DER LINDEN, W. J., & GLAS, C. A. W. *Elements of adaptive testing.* New York: Springer, 2010. DOI: http://dx.doi.org/10.1007/978-0-387-85461-8.

VAN DER LINDEN, W.J.; PASHLEY, P. J. Item selection and ability estimation in adaptive testing. In: VAN DER LINDEN, W. J.; GLAS, C. A. W. (Ed.). *Elements of adaptive testing* (pp. 3-30). New York: Springer, 2010. DOI: http://dx.doi.org/10.1007/978-0-387-85461-8_1.

VAN DER LINDEN, W. J.; REESE, L. M. A model for optimal constrained adaptive testing. *Applied Psychological Measurement,* Thousand Oaks, CA, v. 22, p. 259-270, 1998. DOI: http://dx.doi.org/10.1177/01466216980223006.

VAN DER LINDEN, W. J.; VELDKAMP, B. P. Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics,* Washington, DC, v. 29, p. 273-291, 2004. DOI: http://dx.doi.org/10.3102/10769986029003273.

VAN DER LINDEN, W. J.; VELDKAMP, B. P. Conditional item exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics,* Washington, DC, v. 32, p. 398-417, 2007. DOI: http://dx.doi.org/10.3102/1076998606298044.

VEERKAMP, W. J. J.; BERGER, M. P. F. Some new item selection criteria for computerized adaptive testing. *Journal of Educational and Behavioral Statistics,* Washington, DC, v. 22, p. 203-226, 1997. DOI: http://dx.doi.org/10.3102/10769986022002203.

VELDKAMP, B. P. Bayesian item selection in constrained adaptive testing using shadow tests. *Psicologica,* v. 31, p. 149-169, 2010.

VELDKAMP, B. P. Application of robust optimization to automated test assembly. *Annals of Operations Research,* New York, 2012. DOI: http://dx.doi.org/10.1007/s10479-012-1218-y.

VELDKAMP, B. P.; VAN DER LINDEN, W. J. Designing item pools for Computerized Adaptive Testing. In: VAN DER LINDEN, W. J.; GLAS, C. A. W. (Ed.). *Computerized Adaptive Testing:* theory and practice. Boston, MA: Kluwer Academic Publishers, 2000. p. 149-162.

VELDKAMP, B. P.; VAN DER LINDEN, W. J. Multidimensional constrained adaptive testing. *Psychometrika,* New York, v. 67, p. 575-588, 2002. DOI: http://dx.doi.org/10.1007/BF02295132.

VELDKAMP, B. P.; VERSCHOOR, A. J.; EGGEN, T. J. J. M. A multiple objective test assembly approach for exposure control problems in Computerized Adaptive Testing. *Psicologica,* Valencia, v. 31, p. 335-355, 2010.

VERSCHOOR, A. J.; STRAETMANS, G. J. J. M. MathCAT: a flexible testing system in Mathematics Education for Adults. In: VAN DER LINDEN, W.J.; GLAS, C.A.W. (Ed.). *Elements of adaptive testing.* New York: Springer, 2010. p. 137-156. DOI: http://dx.doi.org/10.1007/978-0-387-85461-8_7.

WARM, T. A. Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika,* New York, v. 54, p. 427-450, 1989. DOI: http://dx.doi.org/10.1007/BF02294627.

WAINER, H. Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice,* Hoboken, NJ, v. 12, p. 15-20, 1993. DOI: http://dx.doi.org/10.1111/j.1745-3992.1993.tb00519.x.

WAINER, H. et al. *Computerized adaptive testing:* a primer. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

WAY, W. D.; STEFFEN, M.; ANDERSON, G. S. (1998). Developing, maintaining, and renewing the item inventory to support computer-based testing. In: MILLS, C. N. (Ed.) et al. C*omputer-based testing: building the foundation for future assessments.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1998. p. 89-102.

WEISS, D. J. *The stratified adaptive computerized ability test*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1973. (Research Report 73-3).

WEISSMAN, A. Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement,* Thousand Oaks, CA, v. 67, p. 41-58, 2007. DOI: http://dx.doi.org/10.1177/0013164406288164.

ZIMOWSKI, M. F. et al. *BILOG-MG:* multiple-group IRT analysis and test maintenance for binary items. Chicago: Scientific Software, 1996.

ZIMOWSKI, M. F. et al. *BILOG-MG3 User guide*. Chicago: SSI Central, 2003.

ZWINDERMAN, A. H. A generalized Rasch model for manifest predictors. *Psychometrika,* New York, v. 56, p. 589-600, 1991. DOI: http://dx.doi.org/10.1007/BF02294492.

ZWINDERMAN, A. H. Response models with manifest predictors. In: VAN DER LINDEN, W. J.; HAMBLETON, R. K. (Ed.). *Handbook of modern item response theory*. New York: Springer-Verlag, 1997. p. 245-256.

Recebido em: 19/12/2012
Aceito para publicação em: 18/04/2013