Article:

Delgado-Gómez, D., Laria, J.C. and Ruiz-Hernandez, D. orcid.org/0000-0001-5538-6930 (2019) Computerized adaptive test and decision trees: A unifying approach. Expert Systems with Applications, 117. pp. 358-366. ISSN 0957-4174

https://doi.org/10.1016/j.eswa.2018.09.052

Computerized adaptive test and decision trees: a unifying approach

David Delgado-Gómez[a,*], Juan C. Laria[a], Diego Ruiz-Hernández[b]

[a] *Universidad Carlos III de Madrid, Department of Statistics, Leganés, Spain*
[b] *University of Sheffield Management School, Sheffield, UK*

Abstract

In the last few years, several articles have proposed decision trees (DTs) as an alternative to computerized adapted tests (CATs). These works have focused on showing the differences between the two methods with the aim of identifying the advantages of each of them and thus determining when it is preferable to use one method or another. In this article, Tree-CAT, a new technique for building CATs is presented. Unlike the existing work, Tree-CAT exploits the similarities between CATs and DTs. This technique allows the creation of CATs that minimise the mean square error in the estimation of the examinee's ability level, and controls the item's exposure rate. The decision tree is sequentially built by means of an innovative algorithmic procedure that selects the items associated with each of the tree branches by solving a linear program. In addition, our work presents further advantages over alternative item selection techniques with exposure control, such as instant item selection or simultaneous administration of the test to an unlimited number of participants. These advantages allow accurate on-line CATs to be implemented even when the item selection method is computationally costly.

Keywords: Decision trees, linear programming, computerized adaptive tests

1. Introduction

Computerized Adaptive Tests (CATs) are sophisticated tests capable of improving the accuracy of conventional tests while administering a much smaller number of items (Weiss, 2004). They are based on the Item Response Theory (IRT) that emerged as an alternative to the traditional pencil and paper tests with the goal of obtaining comparable estimates of the participants' abilities when these are obtained with different test designed for measuring the same trait (van der Linden and Glas, 2000). These characteristics have lead to multiple applications of CATs as clinical and academical assessments (Fliege et al., 2005; Tseng, 2016); or personnel recruitment (Chapman and Webster, 2003), among others.

In a standard CAT, each examinee receives a tailored test whose integrating items are aimed at attaining the best fit to the participant's actual level of the trait, avoiding the presentation of non-informative items to the examinee. With this aim, each of the items presented to the participant is selected from an item bank taking into consideration the responses to all previously presented items, as well as their characteristics (difficulty, discriminating capacity, etc.)

*Corresponding author.
Email addresses: ddelgado@est-econ.uc3m.es (David Delgado-Gómez), jlaria@est-econ.uc3m.es (Juan C. Laria), D.Ruiz-Hernandez@sheffield.ac.uk (Diego Ruiz-Hernández)

and those of the items that have not yet been presented. Because of this, one of the core components of a CAT is the item selection criterion.

In this regard, the most widely used criterion is *Fisher Maximum Information* (Lord, 1980; Weiss, 1982). However, despite its widespread use, several weaknesses have been pointed out. These include item selection bias, large estimation errors at the beginning of the test, high item exposure rates, and content imbalance problems (Lu et al., 2012, among others). Various alternatives have been proposed as attempts for addressing these problems; e.g. the minimum Expected Posterior Variance (EPV) (van der Linden and Pashley, 2009), Maximum Likelihood Weighted Information (MLWI) (Veerkamp and Berger, 1997), Kullback-Leibler information (KL) (Chang and Ying, 1996) or mutual information (MI) (Weissman, 2007). Notwithstanding these item selection techniques have solved many of the mentioned weaknesses, the computational cost of some of them limits their application in practice, in particular because of the need of numerical integration (Ueno and Songmuang, 2010).

Another well known weakness of information-based item selection methods is the overexposure of items. This is a consequence of the fact that that only a few items from the test bank are maximally informative over the ability range (van der Linden and Veldkamp, 2007). Indeed, Veldkamp and Matteucci (2013) observed that only 12 out of a 499 items bank were maximum-informative to any skill level. Among the exposure control methods that have appeared in literature (Georgiadou et al., 2007) we can mention the *randomesque* method (Kingsbury and Zara, 1989; Shin, 2017); the Sympson-Hetter procedure (Sympson and Hetter, 1985); the elegibility method (van der Linden, 2003); the shadow test (van der Linden and Veldkamp, 2005); the restricted procedure (Revuelta and Ponsoda, 1998); the adaptive tests method (Armstrong and Edmonds, 2004); and the progressive-restricted method (Revuelta and Ponsoda, 1998). Unfortunately the additional procedures introduced by these techniques add computational time to the already heavy item-selection methods. Moreover some of the above mentioned techniques require the recalculation of some parameters every time a participant completes the test, preventing the simultaneous application of the test to more than one participant.

In recent years, Decision Trees (DTs) have been proposed as an alternative to CATs. One of the main advantages of the DTs is that the complete test can be designed in advance (using a tree structure) and applied to the examinee without delay, avoiding the item selection step and the associated computational cost. In addition, some researchers have underlined some theoretical benefits of the DTs. Ueno and Songmuang (2010) developed a DT to predict the standardised total raw test score of the respondents. Their proposal has the advantages of not having to satisfy the local independence condition of traditional CATs, and being capable of obtaining accurate estimates of the standardised scores whilst using of a smaller number of items than CATs. Despite these benefits, there are two main drawbacks to this work. The most important one is that, when using total scores, the comparability property of the IRTs is lost. i.e. their approach suffers from the same problem that existed in the classical test theory. The second limitation is that, for the construction of the DT, a large amount of data must be available for guaranteeing that each of the subsequent subsets, created during the construction of the tree, has sufficient information about the distribution of the latent variable. Earlier, Yan et al. (2004) had proposed a related method where nodes with similar scores are merged for keeping the number of nodes within reasonable limits. Notwithstanding this solves the second limitation, the most important problem, the lack of comparability between tests, which hinders the use of DTs as an alternative to CATs, remains unresolved.

From an applied point of view, healthcare has probably been the field where the most intense and fruitful debate has appeared regarding the use of CATs and DTs. For example, in clinical psychology and psychiatry, several papers have been published using CATs for diagnosing mental disorders. Among them, Gardner et al. (2004) developed a CAT to identify individuals with major depressive episodes based on the Beck Depression Inventory scale; Moore et al. (2018) developed a CAT to identify individuals with psychotic spectrum disorder. In a different medical area, Leung et al. (2016) pointed out the PROMIS CAT as an excellent instrument for predicting clinically significant fatigue, sleep disturbance, and sleep impairment among patients who attended to a cancer research centre. Despite these good results, some researchers have argued that CATs are not suitable for diagnostic classification tasks. For example, Gibbons et al. (2016) argued that CATs are ideal for measuring severity but not for diagnosis screening, distinguishing between CATs and Computerized Adaptive Diagnosis (CADs). and developed a DT based CAD for detecting major depression disorder. Recently, Delgado-Gomez et al. (2016) compared the performance of a DT and a CAT for identifying suicidal behaviour using the personality and life events scale (Blasco-Fontecilla et al., 2012). Their results showed that a DT required fewer items than a CAT for obtaining a similar classification rate. Those works reinforce the idea that DTs, a supervised technique, are more suitable for diagnostic classification, while CATs, being unsupervised, are more suitable for quantifying severity.

As the discussion above suggests, the existing literature has mainly focused on emphasising the differences between CATs and DTs. This article addresses the study of these two techniques from the opposite perspective: it seeks to identifying and exploiting their similarities. First, we show that a CAT can be represented by a tree structure. This allows pre-computing, storing and lately administering a CAT without incurring any item selection time, regardless of the item selection criterion used. Second, we prove that building a DT that minimises the mean square error (MSE) is equivalent to designing a CAT using the minimum EPV as item selection criterion. This result provides a better understanding to the EPV criterion and establishes a bridge between the DTs and the CATs, providing a new perspective to the aforementioned debate on the use of these techniques. Finally, we show that a CAT with exposure control can be seen as a forest of DTs. This allows the development of an optimization algorithm for the simultaneous construction of the trees that make up this forest. The above results together enable the construction of a CAT with minimum MSE and exposure control.

The rest of the article is structured as follows. In Section 2, we show that an unconstrained CAT can be represented in a tree structure. In Section 3 we show that, using DTs, it is possible to construct an unconstrained CAT that minimises the MSE. In this section we also discuss some computational aspects of the proposed technique. Finally, it is proved that the constructed tree is equivalent to a CAT that uses minimum EPV as item selection technique. In

Section 4, we adapt the proposed technique for controlling the item exposure rate. With this aim, we first show that a CAT with controlled exposure rate can be seen as the simultaneous construction of several decision trees. Section 5 shows the results of a study aimed at comparing our methodology with other methods for creating CATs with item exposure control using simulated data. Results of the application of the proposed technique on real data are discussed in Section 6. Finally, the article concludes in Section 7 with a discussion of the results obtained and their implications.

2. Representing an Unconstrained CAT in a Tree Structure

In this section we show that a CAT without exposure control can be represented in a tree structure. This representation enables a fast selection (in the order of milliseconds) of the items presented to the examinee. It also facilitates the development of the models introduced in the following sections. The notation introduced herein will be used throughout the rest of the article and is summarised in the Appendix.

Consider a test composed of I items that will be administered to J individuals for assessing certain trait θ. For the sake of simplicity, and without loss of generality, we assume that all items have R possible answers. When the test is to be administered to participant j, the only information available is the distribution of θ in the population, given by the density function $f(\theta)$. Before any item has been administered, it is frequent to assume that the value of this trait for a particular examinee is given by the maximum of $f(\theta)$. This value is denoted by $\hat{\theta}_{\emptyset}$.

The first item that is administered to this participant, i_1^j, is the one that reaches the maximum value of a pre-established item selection criteria (FMI, MEPV, KL, etc.) given $\hat{\theta}_{\emptyset}$. We note that, when item exposure control is not taken into account, the first item to be administered to all participants is the same, i_1^j, since $\hat{\theta}_{\emptyset}$ is identical for all participants. Once the examinee responds to this item, providing the answer $r(i_1^j) \in \{1, ..., R\}$, his trait is re-assessed to a new value $\hat{\theta}_{u_1^j}$, where $u_1^j = r(i_1^j)$ indicates the first item given to examinee j and the answer provided.

This newly estimated value of the trait, $\hat{\theta}_{u_1^j}$, is then used to select the next item to be presented to the examinee, i_2^j. It is important noticing that all participants who provide the same answer to the first item will get the same estimate $\hat{\theta}_{u_1^j}$, and will therefore be given the same second item. Once the examinee has answered to the new item, the estimated value of the trait is updated to $\hat{\theta}_{u_2^j}$ where $u_2^j = \{r(i_1^j), r(i_2^j)\}$.

This way, subsequent items are administered iteratively until a given criterion is reached. Briefly, when examinee j has responded to the first n items by obtaining the response pattern $u_n^j = \{r(i_1^j), \ldots, r(i_n^j)\}$, a new estimate of the trait, $\hat{\theta}_{u_n^j}$, is calculated and the next item is selected based on this value. All those examinees who share the same response pattern u_n^j to the first n items will be given the same item $n + 1$. Based on this discussion, a CAT can be represented in a tree structure as shown in figure 1.

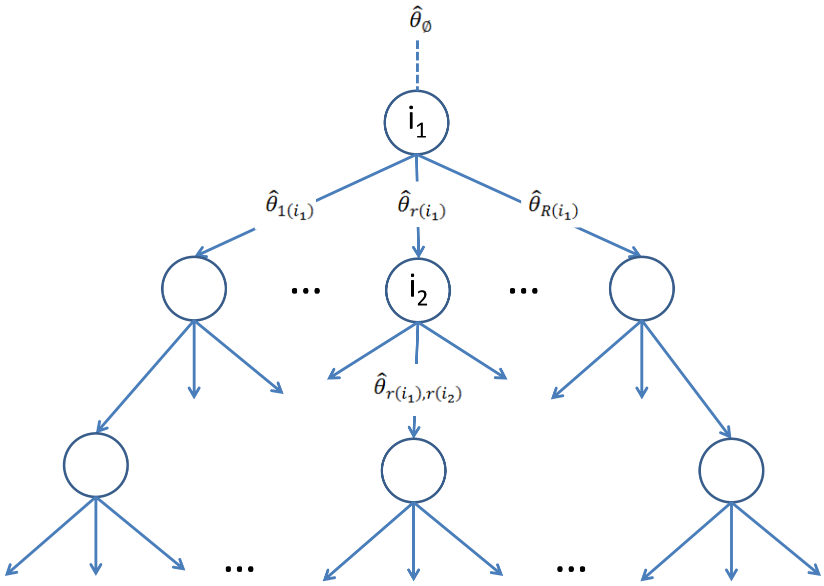


Figure 1: Tree Representation of a CAT.

3. Building a CAT with Minimum MSE

DTs are supervised methods built by minimising the square error in the estimation of an explanatory variable (Rokach and Maimon, 2014). As mentioned above, the available research work using the DT methodology as an alternative for CATs, use either the total test's score (Yan et al., 2004; Ueno and Songmuang, 2010) or an external criterion as dependent variable (Delgado-Gomez et al., 2016; Riley et al., 2011). In this section we present a methodology for building a DT that minimises the MSE in the trait's estimation (instead of the test score used in the aforementioned works). The MSE in the estimation of the trait is the most frequently used criterion for building DTs and for assessing the accuracy of a CAT.

In the design of this CAT, we start by building the root of the tree. Take an item i from the test battery. Let θ be the actual trait of a person, j, who answers this item; $p_i(r|\theta)$, the probability that this person will give the answer $r \in \{1, ..., R\}$; and $\hat{\theta}_r$, the value of the trait estimated for each of the possible answers. The MSE of this item for this person is

$$E_i(\theta|\emptyset) = \sum_{k=1}^{R} (\theta - \hat{\theta}^j_{v_1^k})^2 p_i(k|\theta) \tag{1}$$

where the empty set in the expectation emphasises the fact that no item has yet been administered; and $v_1^k = \{r(i) = k\}$. The MSE that will be obtained if item i is administered to the population is, consequently, given by

$$E_i = \int E_i(\theta|\emptyset) f(\theta) d(\theta) \tag{2}$$

The starting item, i_1, which constitutes the root of the tree, will be the one for which the value E_i is minimal.

Once the tree root has been defined, the R items corresponding to its children will be added as follows: if item $i \neq i_1$ is administered after an examinee with real trait θ chose the r-th answer to item i_1, the MSE of this person will be given by

$$E_i(\theta|u_1) = \sum_{k=1}^{R}(\theta - \hat{\theta}_{v_2^k})^2 p_i(k|\theta) \tag{3}$$

where $v_2^k = \{u_1, r(i) = k\}$ and $\hat{\theta}_{v_2^k}$ is the estimated trait considering pattern v_2^k. Therefore, the MSE of the group that gave answer r to item i_1 is given by

$$E_i = \int E_i(\theta|u_1) f(\theta|u_1) d\theta \tag{4}$$

where

$$f(\theta|u_1) = \frac{p(u_1|\theta)f(\theta)}{p(u_1)} = \frac{p(r(i_1)|\theta)f(\theta)}{\int p(r(i_1)|\theta)d\theta} \tag{5}$$

In general, given an individual with trait θ and response pattern $u_n = \{r(i_1), ..., r(i_n)\}$, the MSE obtained if unused item i is administered next can be written as

$$E_i(\theta|u_n) = \sum_{k=1}^{R}(\theta - \hat{\theta}_{v_{n+1}^k})^2 p_i(k|\theta) \tag{6}$$

where $v_{n+1}^k = \{u_n, r(i) = k\}$. Then, the MSE of a group of participants that has followed pattern u_n becomes

$$E_i = \int E_i(\theta|u_n) f(\theta|u_n) d\theta \tag{7}$$

where

$$f(\theta|u_n) = \frac{p(u_n|\theta)f(\theta)}{p(u_n)} = \frac{\prod_{j=1}^{n} p(r(i_j)|\theta)f(\theta)}{\int \prod_{j=1}^{n} p(r(i_j)|\theta)d\theta} \tag{8}$$

3.1. Computational Issues

An important aspect that needs to be addressed is how to efficiently build the tree, as the number of nodes grows exponentially when the tree expands. Below we discuss three strategies aimed, the first two, at speeding-up the construction; and, the last one, at keeping the number of nodes within reasonable limits.

Parallel programming. Nodes within the same level are constructed independently. Therefore, the items that constitute these nodes can be determined using parallel programming. For example, if a tree developed in a personal computer with four cores was programmed in parallel, the time required to build it would be reduced to 25 percent of the time required time in a single core. Currently, most universities and research centres have small clusters with a few thousand cores available, making the development of the proposed methodology easily attainable.

Passing information from parent to child nodes. As seen in formula (8), to calculate the posterior probability of the ability level, it is necessary to calculate a product of n probabilities. However, given that $n - 1$ of them have already been calculated in the parent node, if this information is stored, only one multiplication is required for each child node and item pair.

Merging branches. One way for limiting the growth in the number of nodes is joining together those branches that lead to similar estimates of ability level. As an example, if an accuracy of 0.001 is set –which is a quite sensible bound-, and assume that the ability takes values between -4 and 4, the maximum number of nodes in each of the tree's levels will be only 8000, which is a more manageable number than the R^ℓ nodes that may potentially appear at level ℓ.

An alternative method, frequently used in DT design, for controlling the size of the tree is pruning some branches. In our case this will imply stopping the growth of the tree in nodes associated to improbable answer patterns. However, this may in practice give raise to situations where one of these nodes is actually visited, implying that an on-line selection of the remaining items in the CAT will need to be conducted. This would considerably increase the duration of the test if the item selection criteria used is among the most computationally expensive ones. For this reason we do not consider this practice a good alternative to branch merging.

3.2. Equivalence of Minimum MSE and Minimum EPV

In this section we establish an interesting result: building a DT minimising the MSE is mathematically equivalent to building a CAT where the item selection criterion is the minimum EPV.

As discussed around equations (6) to (8), the MSE can be written as

$$MSE = \int p(\theta|u_{j-1}) \sum_{r=1}^{R} p_i(r|\theta)(\theta - \hat{\theta}_{u_j})^2 d\theta \tag{9}$$

which becomes

$$= \int \sum_{r=1}^{R} p(\theta|u_{j-1}) p_i(r|\theta)(\theta - \hat{\theta}_{u_j})^2 d\theta \tag{10}$$

and using Bayes theorem

$$= \int \sum_{r=1}^{R} \frac{p(u_{j-1}|\theta)p(\theta)}{p(u_{j-1})} p_i(r|\theta)(\theta - \hat{\theta}_{u_j})^2 d\theta \tag{11}$$

using the local independence condition this equation can be simplified to

$$= \int \sum_{r=1}^{R} \frac{p(u_j|\theta)p(\theta)}{p(u_{j-1})} (\theta - \hat{\theta}_{u_j})^2 d\theta \tag{12}$$

after multiplying and dividing by $p_i(r|u_{j-1})$ we get

$$= \int \sum_{r=1}^{R} \frac{p(u_j|\theta)p(\theta)p_i(r|u_{j-1})}{p(u_{j-1})p_i(r|u_{j-1})} (\theta - \hat{\theta}_{u_j})^2 d\theta \tag{13}$$

which, after using conditional probability, becomes

$$= \int \sum_{r=1}^{R} \frac{p(u_j|\theta)p(\theta)p_i(r|u_{j-1})}{p(u_j)} (\theta - \hat{\theta}_{u_j})^2 d\theta \tag{14}$$

using Bayes agaoin, this expression can be further simplified to

$$= \int \sum_{r=1}^{R} p(\theta|u_j)p_i(r|u_{j-1})(\theta - \hat{\theta}_{u_j})^2 d\theta \tag{15}$$

finally, after reordering terms we get

$$= \sum_{r=1}^{R} p_i(r|u_{j-1}) \int p(\theta|u_j)(\theta - \hat{\theta}_{u_j})^2 d\theta = \sum_{r=1}^{R} p_i(r|u_{j-1}) Var(\theta|u_j) \tag{16}$$

which is precisely the EPV criterion.

Consequently, notwithstanding the works discussed in the introduction treat CATs and DTs as disjoint methods, in this section we have established the equivalence between them. In practical terms, this implies that building a CAT with minimal EPV is equivalent to constructing a DT minimising its standard MSE criterion. This result suggests that when the objective of the CAT is minimising the MSE, the most appropriate item selection criterion would be EPV.

4. Tree-CAT: A CAT with Controlled Item Exposure Rate and Minimum MSE

In this section, we propose a method for building a CAT that minimises the MSE with controlled maximum exposure rate (proportion of the individuals taking the test that receive a particular item) by building several decision trees simultaneously.

The underlying idea stems from the so-called randomesque method. At each level, this method randomly selects the next item among the K items with the best selection criteria values, given the current estimated ability $\hat{\theta}$. For each participant, randomesque starts selecting one of the K items attaining maximal values for the selection criteria at the initial trait $\hat{\theta}_0$. Each of these items can be seen as constituting the root of one of K trees. From each root will stem R branches, corresponding to the R possible answers, each of them spanning K nodes. This process is repeated at each level, ℓ, of the tree. Therefore, the randomesque method can be visualised as a forest of K trees. This is represented as a DTs forest in Figure 2 for $R = 2$ and $K = 3$. In this figure white items represent the selected items and the black dots the corresponding trait estimates.

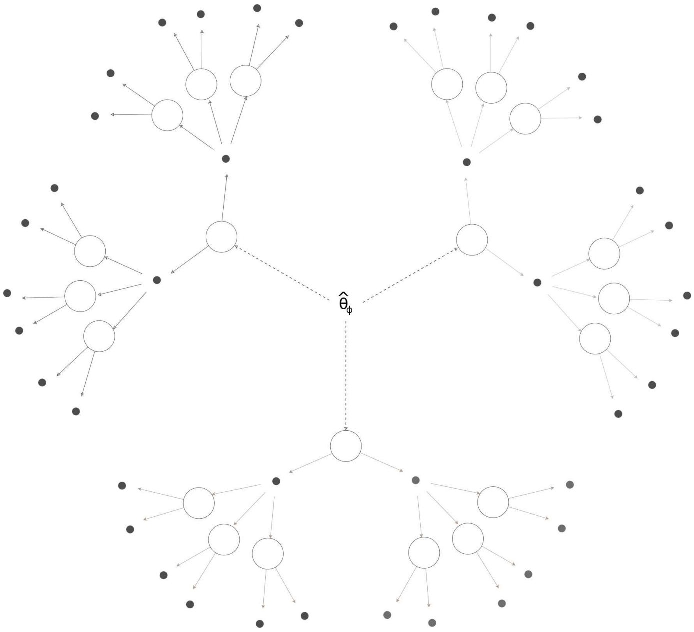


Figure 2: Representation of randomesque method as a DTs forest.

Although this method reduces the item's exposure, it does not prevent an item from exceeding the maximum exposure rate. To address this problem, in the following lines we present the Tree-CAT method. This method builds on ramdomesque for generalising the method developed in the previous section. Tree-CAT imposes a probabilistic bound to the maximum rate of item exposure when creating the forest of trees.

Tree-CAT starts by selecting the K initial nodes. Let E be the vector containing the items' MSEs as computed by equation (2); D, a vector indicating the items' availability; P, a vector containing the probability of each item to be administered as first item in the test; and r_{max}, the maximum item exposure rate. Initially, each of the elements in D is set equal to the maximal exposure rate. Given that 100% of the participants has to be assigned an item at the beginning of the test, the algorithm utilises a capacity variable c to represent the proportion of individuals that remain uncovered after each item is included. $\mathcal{L}$ is a very large number. The selection of the nodes and determination of their number, K, is conducted as indicated in Algorithm 1.

The algorithm starts by selecting the item i with least MSE and associates to this item the minimal value among its current availability, D_i, and the unassigned capacity, c. This value, P_i, is then subtracted from both, the item's availability and the capacity variable. For guaranteeing that this item will not be selected again, its value in vector E is replaced by a very large number $\mathcal{L}$. This procedure is then repeated until c is equal to zero. The algorithm returns the set of $K = |\mathcal{F}|$ initial nodes, and the administration probabilities and updated availability vectors.

Once the K roots have been chosen, the trees spanned by each root will grow jointly in an iteratively fashion. For the sake of clarity in the exposition, we start by describing the procedure generating the second level of the trees.

Algorithm 1 RootSpan

Require: E, D

$c := 1$
$P := \mathbf{0}_{(I\times 1)}$
$\mathcal{F} := \emptyset$
while $c > 0$ **do**
    $i := \text{argmin}\{E\}$
    $P_i := \min\{c, D_i\}$
    $c := c - P_i$
    $D_i := D_i - P_i$
    $\mathcal{F} := \mathcal{F} \cup i$
    $E_i := \mathcal{L}$
end while

Ensure: $\mathcal{F}, D, P$

Let $\mathbf{E}$ be a matrix whose element E_{ij} is the MSE incurred if item i was added to branch j, where each j is given by a different root/answer combination, i.e. $j = R \times (k-1) + r$ for $k = 1,\ldots,K; r = 1,\ldots,R$. Let $\mathbf{C}$ be a vector containing the proportion of participants associated with branch j, where $C_j = P_k \int P(r|\theta, i_k) f(\theta) d\theta$ and $\sum_j C_j = 1$. Let $\mathbf{D}$ be the available capacity vector returned by Algorithm 1. Then, the choice of the items associated with each of the branches is done by means of the following linear program:

$$\begin{aligned} \min \quad & \sum_i \sum_j X_{ij} E_{ij} \\ s.t. \quad & \sum_i X_{ij} \leq D_i \\ & \sum_j X_{ij} = C_j \end{aligned} \tag{17}$$

This simple model minimises the MSE subject to the constraints that not item will exceed its availability; and that all participants must be given a second item during the test. Further levels of the trees are obtained by successive applications of this procedure, with system (17) solved over the matrix $\mathbf{E}$ obtained for the corresponding item/response combination (henceforth referred to as branch); the last update of vector D; and a newly obtained vector $\mathbf{C}$ where $C_j = P_k \int P(r|\theta, u_{k-1}) f(\theta) d\theta$.

Unfortunately, the number of constraints grows exponentially on the number of levels, making the linear program computationally intractable. A computationally efficient heuristic, illustrated in Algorithm 2, has been developed for addressing this problem.

Algorithm 2 can be seen as a bi-dimensional extension of Algorithm 1. Working with inherited vector D and matrices E and $\mathcal{C}$ as inputs, the Algorithm returns an array $\mathcal{F}$ of sets of items for all possible branches stemming from the previous level. It also returns a matrix P containing the relative probability for each item to be administered to an individual in a given branch, and a vector D with the updated items' availability.

It is important noticing that at any givel level ℓ of the tree, nodes may be

Algorithm 2 Growing the tree

Require: $E, D, \mathcal{C}$

$c := 1$
$P := (0)_{I \times RK}$
$\mathcal{F} := \{\mathcal{F}_1, \ldots, \mathcal{F}_{RK}\}$, $\mathcal{F}_h := \emptyset \; \forall h = 1, \ldots, RK$
while $c > 0$ **do**
    for $j \leq I$ **do**
        if $D_j == 0$ **then**
            $E_{j\bullet} := \mathcal{L}$
        end if
    end for
    $(i, j) := \text{argmin}\{E\}$
    $P_{ij} := \min\{C_j, D_i\}$
    $D_i := D_i - P_{ij}$
    $c := c - P_{ij}$
    $\mathcal{F}_j := \mathcal{F}_j \cup i$
    $E_{i,j} := \mathcal{L}$
end while

Ensure: $\mathcal{F}, D, P$

assigned more than one item. The reason for this is that the best item for a given node may not have the required capacity (i.e. $D_j < C_j$).

5. Numerical Experiments: Simulated Data

In this section we present the results of an experimental assessment of the performance of the Tree-CAT method. The experiment compares our method with three other available methods designed for controlling item exposure, namely, restrictive (disallows the use of items that exceed the maximum rate), item eligibility (restricts the likelihood of administering an item to a given exposure rate), and randomesque methods (randomly selects the next item from a subset of the most informative items). In order to achieve a fair comparison between the four methods, MEPV is used in all of them as the item selection criteria. This choice is due to the fact that, as shown in Section 3.2, this criterion minimises the MSE.

5.1. Data and experimental set-up

The experiment set-up is similar to the one used by other authors when comparing item exposure control techniques in CATs (Pastor et al., 2002). In detail, the item bank consists of 100 items with randomly generated parameters according to Samejima's graded response model (Samejima, 2016). Each item's discrimination parameter was generated following a log-normal distribution with zero mean and standard deviation equal to 0.1225. The difficulty parameters were generated following a standard normal distribution (Magis and Raîche, 2011). The maximum exposure rate was set to 0.3 with test length equal 10. This length is considered to be enough for comparing the different methods and it is similar to the one appearing in recent works. For example, CATs developed by De Beurs et al. (2014); Stucky et al. (2014); and Hsueh et al.

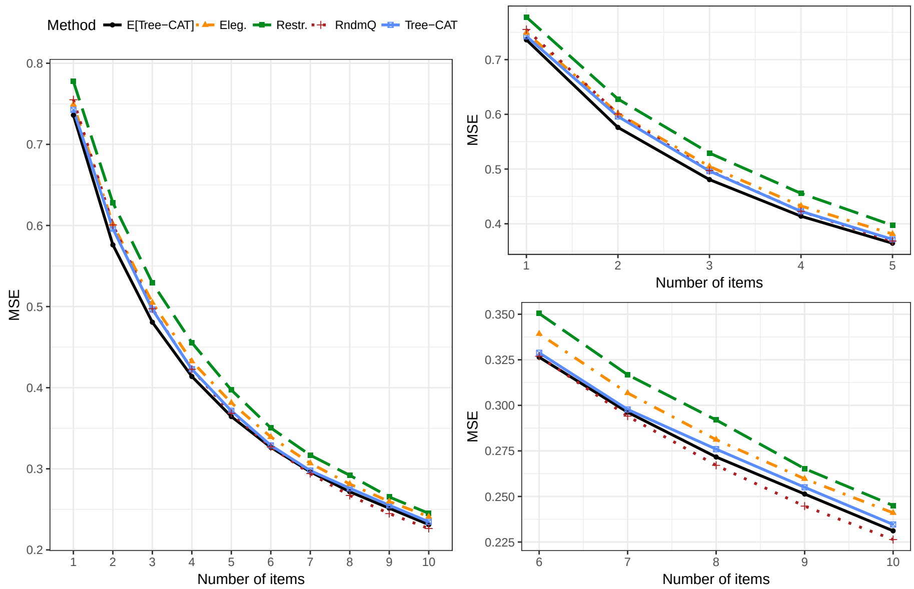
(2016), for assessing different clinical conditions, used averages of 4, 5.3 and 6 items, respectively. Regarding the randomesque method, the number of random alternatives available for each node at each level of the tree is set to six.

The performance of the CATs was evaluated by means of the answers of 500 randomly generated examinees (Magis et al., 2012). Given the random nature of the item selection of three of the used procedures (randomesque, item eligibility and ours), and to avoid path dependence in the results, the test was repeated 25 times for each examinee and means were taken. In order to improve the significance of the results, this scenario was repeated 10 times.

5.2. Results

Figure 3 shows the evolution of MSE attained by each of the techniques during the test execution. The large panel shows the entire execution, with the two small panels being zoomed-in versions of the performance over the first and last five items, respectively. The dot-dash yellow line represents the eligibility method; the dash green line, the restrictive method; the dotted line, the randomesque method; and the the solid blue line, the Tree-CAT method. An extra line, solid black, shows the theoretical expected MSE corresponding to the Tree-CAT method.

Figure 3: Average MSEs for the Alternative Techniques



The figure shows that the Tree-CAT method obtains more precise estimates than the eligibility and the restricted methods in terms of MSE. This graph also shows that the Tree-CAT attains a performance close to the theoretically expected one. Finally, the randomesque method shows a slightly better performance than the Tree-CAT from the seventh item administered on. This can be explained by looking at the overlap rate, which is a common measure of test security defined as the percentage of common items for any two randomly selected examinees (Barrada et al., 2007). In our experiment, the computed overlap rates are 0.268 for restrictive; 0.275 for eligibility; 0.283 for Tree-Cat; whereas it reaches 0.538 for randomesque.

Regarding the computation time, Table 1 shows the time needed to create the DT as well as the minimum time required by each of the methods to select the 10 items for the 500 participants. It is important to note here that in both, item eligibility and restricted methods, participants receive the test sequentially. That is, in order to recalculate the parameters, the current participant must have finished the test before the next one receives it. In contrast, randomesque and Tree-CAT methods are able to administer the test simultaneously. Moreover, whereas the tree alternative methods select the next item on-line, Tree-CAT generates the whole tree at once, which means that the time required for generating the next item is, indeed, zero. The experiment was conducted using 128 cores of a cluster with a Xeon 2630 processor and 32 GB of RAM.

Table 1: Training and Execution Times

Method	Training Time	Test Time serial
Tree-CAT	$\approx$ 7 days	0 secs
Randomesque	0 secs	$\approx$ 16.8 hours (120 secs$\times$500)
Eligibility	0 secs	$\approx$ 23.6 hours (170 secs$\times$500)
Restricted	0 secs	$\approx$ 16.8 hours (120 secs$\times$500)

According to the table, the randomesque, restricted and eligibility methods take 2 minutes for selecting the items. In practical terms this means that the examinee will need to wait 12 seconds in average before the next item is provided. These long execution times are explained, firstly, by the use of MEPV, which has a high computational cost. More economical item selection methods such as FMI could render better results in terms of computational times, at the cost of incurring the problems highlighted in the introduction to this paper. Secondly, those long times can also be attributed to the use of the implementation catR (Magis and Raîche, 2011), which does not use any of the two speeding-up strategies described in Section 3.1. It should be said that, even if those strategies were implemented, the eligibility and restrictive method still suffer from the sequential application burden, which imposes a serious penalty in the execution time (23.6 and 16.8 hours for 500 administrations of the test).

It is also important to mention that the cost in computational time incurred by the three alternative methods discussed in this section is paid every time the test is conducted. With the Tree-CAT method, in contrast, once the trees are built and all the alternative sequences stored, the time between the answer and the selection of the next item is –to all practical extent- zero, regardless the number of participants. This feature enables the simultaneous on-line application of the test to an unlimited number of participants, something that is not possible with the other methods. Hypothetically, this could be attained with randomesque, but in this case the simultaneous application of the test to a large number of people will require the availability of a server with as many nodes as participants.

6. Numerical Experiments: Real Data

This section evaluates the proposed methodology using actual data. These data have been obtained from a previous study (Rubio et al., 2007), in which a psychometric scale for measuring emotional adjustment was developed. Before presenting the experimental results, in the following section we describe both the data set and the design of the experiment.

6.1. Data and experimental set-up

The data in this study contain the answers provided by 792 psychology students to the 28 items of the Emotional Adjustment Bank (Rubio et al., 2007). For our experiments, it was considered that the item responses have three levels (”disagree”, ”neutral” and ”agree”). For testing the unidimensionality of the scale, a factor analysis in conjunction with a parallel analysis (Hayton et al., 2004) showed that only one factor is retained. This confirms the unidimensionality and justifies the use of a graded response model.

In order to compare the performance of the Tree-CAT method against the chosen exposure control methods (Restrictive, Eligibility, Randomesque) under conditions similar to the real ones, the hold-out validation method was used. Specifically, the data set was randomly divided into two disjoint subsets of equal size: the training set and the test set. The training set was used to estimate the different items' parameters and to build the DT for the Tree-CAT method, whereas, the test set was used for the comparisons. It was assumed that the traits θ of the participants were those obtained when the 28 items of the bank were administered to them. The test length was set to 7 items. The remaining parameters that define the experiment have been set to the same values as those of the simulation study in Section 5. Namely, the MEPV was chosen as item selection criterion; the maximum exposure rate was fixed at 0.3; and the number of random alternatives for the Randomesque method was set to 6. As before, in order to avoid path dependence, the test was repeated 25 times for each examinee, and means were taken for the Tree-CAT, Elegibility and Randomesque methods. In addition, to achieve more reliable results, this scenario was simulated 10 times.

6.2. Results

Figure 4 shows the MSE obtained by the different techniques as a function of the number of items administered to the subjects. It can be noticed that, except for the Randomesque method in the last levels, Tree-CAT is the one achieving the best performance (based on the MSE). As explained in the discussion to our simulated experiments, the reason why Randomesque outperforms the other three methods at the last levels of the test is that it exceeds the maximum exposure rate. The overlap rates of Tree-CAT, Restrictive, Eligibility and Randomesque methods are 0.28, 0.28, 0.29 and 0.58, respectively.

Table 2 depicts the computational time used to construct the decision tree for the Tree-CAT method, and the time needed to select the next item for each of the four techniques. These numbers are similar to those obtained in Table 1 of the previous experiment on a smaller scale, as the item bank used in this study is 28% the size of the previous one, and the length of the test is 7 items instead of 10.

Figure 4: Average MSEs for the Alternative Techniques

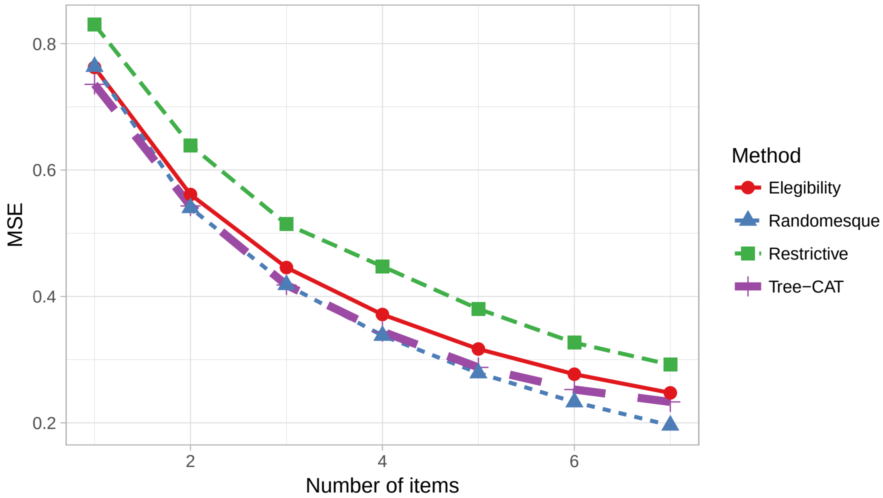


Table 2: Training and Execution Times

Method	Training Time	Test Time serial
Tree-CAT	≈ 36 min.	0 secs
Randomesque	0 secs	≈ 103 min. (15.6 secs×396)
Eligibility	0 secs	≈ 117 min. (17.7 secs×396)
Restricted	0 secs	≈ 103 min. (15.6 secs×396)

7. Conclusion

In this article, we present a new method for building CATs, referred to as Tree-CAT, based on the DTs methodology. The proposed method creates and stores a representation of the CAT in a tree structure that allows items to be selected in milliseconds. This property is especially valuable when the chosen item selection method involves the calculation of integrals (e.g. when a CAT uses minimal EPV for item selection). In this regard, it is demonstrated that building a CAT that minimises the EPV is equivalent to building a DT that minimises the MSE.

In the article we also show that creating a CAT with item exposure controls can be understood as the simultaneous construction of several trees, and propose an algorithm for performing this task. This algorithm allows the use of different strategies that accelerate its construction. First, it is possible to use parallel programming to calculate the MSE matrix required by the algorithm. Second, the calculation of MSEs can be simplified using information obtained at the previous level nodes. Finally, it seems possible to merge branches that produce similar estimates of the trait level, allowing the tree to be kept within reasonable dimensions. In this article we have conducted experiments taking advantage of the first two strategies.

Tree-CAT presents several advantages with respect to other existing methods. Firstly, the results obtained experimentally show that Tree-CAT is the method with the lowest MSE among those with the lowest overlap rate. Another advantage is that it can potentially be administered simultaneously to an unlimited number of participants. In contrast to existing methods, which calculate in real time each of the items to be presented based on previous answers, the Tree-CAT selects the next item to be presented from a previously stored structure. This allows, for practical purposes, to eliminate the time required for item selection. This is especially useful when item selection criteria are computationally expensive. These two properties, namely, simultaneous application and zero time in the selection of items, make Tree-CAT an ideal candidate for the simultaneous administration of on-line tests to a large number of participants.

One weakness of the method is the need of a small computer cluster for building the tree within reasonable time. For example, in the experiment developed in this article, 128 nodes of a cluster were used. However, the availability of a larger cluster could reduce the construction time of the tree from one week –as in our case- to a few hours. The importance of this limitation is further reduced by the fact that, once the tree has been built, the test can be administered from any personal computer.

Regarding this limitation, an appealing future research line consists of finding a mechanism for optimally merging the branches of the trees in order to limit the size of the trees. Additional research could also be developed for addressing issues like content balance, variable test length, or multidimensional-trait assessment.

We conclude the article by stating our conviction, supported by the experimental and analytical results obtained, that the DTs approach for building CATs is a promising research line that opens up several lines of research and combines the knowledge of the areas of Psychology, Statistics, Operational Research and Computer Science.

Acknowledgments

Numerical experiments were conducted in Uranus, a supercomputer cluster located at Universidad Carlos III de Madrid and jointly funded by EU-FEDER funds and by the Spanish Government via the National Projects No. UNC313-4E-2361, No. ENE2009-12213- C03-03, No. ENE2012-33219, No. ENE2012-31753 and No. ENE2015-68265-P.

References

Armstrong, R. and Edmonds, J. (2004). A study of multiple stage adaptive test designs. In *annual meeting of National Council of Measurement in Education,(NCME), San Diego, CA*.

Barrada, J. R., Olea, J., and Ponsoda, V. (2007). Methods for restricting maximum exposure rate in computerized adaptative testing. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 3(1):14. doi:.

Blasco-Fontecilla, H., Delgado-Gomez, D., Ruiz-Hernandez, D., Aguado, D., Baca-Garcia, E., and Lopez-Castroman, J. (2012). Combining scales to assess suicide risk. *Journal of psychiatric research*, 46(10):1272–1277. doi:10.1016/j.jpsychires.2012.06.013.

Chang, H.-H. and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3):213–229. doi:10.1177/014662169602000303.

Chapman, D. S. and Webster, J. (2003). The use of technologies in the recruiting, screening, and selection processes for job candidates. *International journal of selection and assessment*, 11(2-3):113–120.

De Beurs, D. P., de Vries, A. L., de Groot, M. H., de Keijser, J., and Kerkhof, A. J. (2014). Applying computer adaptive testing to optimize online assessment of suicidal behavior: a simulation study. *Journal of medical Internet research*, 16(9). doi:10.2196/jmir.3511.

Delgado-Gomez, D., Baca-Garcia, E., Aguado, D., Courtet, P., and Lopez-Castroman, J. (2016). Computerized adaptive test vs. decision trees: development of a support decision system to identify suicidal behavior. *Journal of affective disorders*, 206:204–209. doi:10.1016/j.jad.2016.07.032.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., and Rose, M. (2005). Development of a computer-adaptive test for depression (d-cat). *Quality of life Research*, 14(10):2277.

Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., and Frank, E. (2004). Computerized adaptive measurement of depression: a simulation study. *BMC psychiatry*, 4(1):13. doi:10.1186/1471-244X-4-13.

Georgiadou, E. G., Triantafillou, E., and Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, 5(8).

Gibbons, R. D., Weiss, D. J., Frank, E., and Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual review of clinical psychology*, 12. doi:10.1146/annurev-clinpsy-021815-093634.

Hayton, J. C., Allen, D. G., and Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, 7(2):191–205.

Hsueh, I.-P., Chen, J.-H., Wang, C.-H., Chen, C.-T., Sheu, C.-F., Wang, W.-C., Hou, W.-H., and Hsieh, C.-L. (2016). Development of a computerized adaptive test for assessing balance function in patients with stroke. *Physical therapy*, 90(9):1336–1344. doi:10.2522/ptj.20090395.

Kingsbury, G. G. and Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied measurement in education*, 2(4):359–375. doi:10.1207/s15324818ame0204_6.

Leung, Y. W., Brown, C., Cosio, A. P., Dobriyal, A., Malik, N., Pat, V., Irwin, M., Tomasini, P., Liu, G., and Howell, D. (2016). Feasibility and diagnostic accuracy of the patient-reported outcomes measurement information system (PROMIS) item banks for routine surveillance of sleep and fatigue problems in ambulatory cancer care. *Cancer*, 122(18):2906–2917. doi:10.1002/cncr.30134.

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

Lu, P., Zhou, D., Qin, S., Cong, X., and Zhong, S. (2012). The study of item selection method in cat. In *Computational Intelligence and Intelligent Systems*, pages 403–415. Springer. doi:10.1007/978-3-642-34289-9_45.

Magis, D. and Raîche, G. (2011). catR: An R package for computerized adaptive testing. *Applied Psychological Measurement*, 35(7):576–577. doi:10.1177/0146621611407482.

Magis, D., Raîche, G., et al. (2012). Random generation of response patterns under computerized adaptive testing with the r package catR. *Journal of Statistical Software*, 48(8):1–31. doi:10.18637/jss.v048.i08.

Moore, T. M., Calkins, M. E., Reise, S. P., Gur, R. C., and Gur, R. E. (2018). Development and public release of a computerized adaptive (CAT) version of the schizotypal personality questionnaire. *Psychiatry research*. doi:10.1016/j.psychres.2018.02.022.

Pastor, D. A., Dodd, B. G., and Chang, H.-H. (2002). A comparison of item selection techniques and exposure control mechanisms in cats using the generalized partial credit model. *Applied Psychological Measurement*, 26(2):147–163. doi:10.1177/01421602026002003.

Revuelta, J. and Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 35(4):311–327. doi:10.1111/j.1745-3984.1998.tb00541.x.

Riley, B., Funk, R., Dennis, M.and Lennox, R., and Finkelman, M. (2011). The use of decision trees for adaptive item selection and score estimation. In *Annual Conference of the International Association for Computerized Adaptive Testing*.

Rokach, L. and Maimon, O. (2014). *Data mining with decision trees: theory and applications*. World scientific.

Rubio, V. J., Aguado, D., Hontangas, P. M., and Hernández, J. M. (2007). Psychometric properties of an emotional adjustment measure: An application of the graded response model. *European Journal of Psychological Assessment*, 23(1):39–46.

Samejima, F. (2016). Graded response models. In *Handbook of Item Response Theory, Volume One*, pages 123–136. Chapman and Hall/CRC.

Shin, C. D. (2017). Conditional randomesque method for item exposure control in cat. *International Journal of Intelligent Technologies & Applied Statistics*, 10(3). doi:10.6148/IJITAS.2017.1003.02.

Stucky, B. D., Edelen, M. O., Sherbourne, C. D., Eberhart, N. K., and Lara, M. (2014). Developing an item bank and short forms that assess the impact of asthma on quality of life. *Respiratory medicine*, 108(2):252–263. doi:10.1016/j.rmed.2013.12.008.

Sympson, J. and Hetter, R. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association*, pages 973–977.

Tseng, W.-T. (2016). Measuring english vocabulary size via computerized adaptive testing. *Computers & Education*, 97:69–85.

Ueno, M. and Songmuang, P. (2010). Computerized adaptive testing based on decision tree. In *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on*, pages 191–193. IEEE. doi:10.1109/ICALT.2010.58.

van der Linden, W. J. (2003). Some alternatives to sympson-hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 28(3):249–265. doi:10.3102/10769986028003249.

van der Linden, W. J. and Glas, C. A. (2000). *Computerized adaptive testing: Theory and practice*. Springer.

van der Linden, W. J. and Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In *Elements of adaptive testing*, pages 3–30. Springer. doi:10.1007/978-0-387-85461-8_1.

van der Linden, W. J. and Veldkamp, B. P. (2005). *Constraining item exposure in computerized adaptive testing with shadow tests*, volume 2. Law School Admission Council.

van der Linden, W. J. and Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics*, 32(4):398–418. doi:10.3102/1076998606298044.

Veerkamp, W. J. and Berger, M. P. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, 22(2):203–226. doi:10.3102/10769986022002203.

Veldkamp, B. P. and Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio: Avaliação e Políticas Públicas em Educação*, 21(78):57–82. doi:10.1590/S0104-40362013005000001.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied psychological measurement*, 6(4):473–492. doi:10.1177/014662168200600408.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 37(2):70–84.

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, 67(1):41–58. doi:10.1177/0013164406288164.

Yan, D., Lewis, C., and Stocking, M. (2004). Adaptive testing with regression trees in the presence of multidimensionality. *Journal of Educational and Behavioral Statistics*, 29(3):293–316. doi:10.3102/10769986029003293.

Appendix A. Notation

<u>Section 2</u>

$\mathcal{J}$**:** set of participants;

$\mathcal{I}$**:** item bank;

i_n^j**:** n–th item $i \in \mathcal{I}$ to be administered to participant $j \in \mathcal{J}$;

R**:** number of possible answers to an item;

$r(i_n^j)$**:** answer of individual $j \in \mathcal{J}$ to item i_n^j, $i = 1, \ldots, R$.

θ**:** real-valued random variable describing a trait;

$f : \mathbb{R} \to \mathbb{R}^+$ density function of θ;

$\hat{\theta}_\emptyset$**:** $\text{argmax}_{\theta \in \mathbb{R}} f(\theta)$;

u_n^j**:** sequence of items and responses of individual j, with $u_n^j = \{r(i_k^j)\}_{k=0,\ldots,n}$ and $u_0^j = \emptyset$;

$\hat{\theta}_{u_n^j}$**:** estimated θ given pattern u_n^j;

<u>Section 3</u>

$p_i(u_n)$**:** probability of observing sequence u_n in a participant;

$p_i(r|\theta)$**:** probability that a participant with trait θ will answer $r \in \{1 \ldots R\}$ to item $i \in \mathcal{I}$;

$p(u_n|\theta)$**:** probability that a participant with trait θ will show response sequence u_n up to the n-th item shown;

$p(\theta|u_n)$**:** posterior probability of trait θ given a response sequence u_n;

v_n^k**:** sequence of items and responses if an individual with sequence u_{n-1} chooses answer $k \in \{1, 2 \ldots R\}$ to the n–th item.

$\hat{\theta}_{v_n^j}$**:** estimated θ given pattern v_n^j.

<u>Section 4</u>

$X_i j$**:** capacity of item i assigned to branch j;

$E_i j$**:** MSE incurred if item i is added to branch j;

D_i**:** capacity availability vector for item i;

C_j**:** proportion of participants associated to branch j.