LSAC RESEARCH REPORT SERIES

- **Some Practical Issues in Computerized Adaptive Testing With Response Times**

  Bernard P. Veldkamp
  University of Twente, Enschede, the Netherlands

- **Law School Admission Council**
  Research Report 14-06
  October 2014

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation that provides unique, state-of-the-art products and services to ease the admission process for law schools and their applicants worldwide. Currently, 219 law schools in the United States, Canada, and Australia are members of the Council and benefit from LSAC's services. All law schools approved by the American Bar Association are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also members. Accredited law schools outside of the United States and Canada are eligible for membership at the discretion of the LSAC Board of Trustees; Melbourne Law School, the University of Melbourne is the first LSAC-member law school outside of North America. Many nonmember schools also take advantage of LSAC's services. For all users, LSAC strives to provide the highest quality of products, services, and customer service.

Founded in 1947, the Council is best known for administering the Law School Admission Test (LSAT®), with about 100,000 tests administered annually at testing centers worldwide. LSAC also processes academic credentials for an average of 60,000 law school applicants annually, provides essential software and information for admission offices and applicants, conducts educational conferences for law school professionals and prelaw advisors, sponsors and publishes research, funds diversity and other outreach grant programs, and publishes LSAT preparation books and law school guides, among many other services. LSAC electronic applications account for 98 percent of all applications to ABA-approved law schools.

This study is published and distributed by LSAC.

Table of Contents

Executive Summary

Many standardized tests are now administered via computer rather than paper-and-pencil format. The computer-based delivery mode brings with it certain advantages. One advantage is the ability to adapt the difficulty level of the test to the ability level of the test taker in what has been termed computerized adaptive testing, or CAT. A second advantage is the ability to record not only the test taker's response to each item (i.e., question), but also the amount of time the test taker spends considering and answering each item. Combining these two advantages, various methods were explored for utilizing response time data in selecting appropriate items for an individual test taker.

Four strategies for incorporating response time data were evaluated, and the precision of the final test-taker score was assessed by comparing it to a benchmark value that did not take response time information into account. While differences in measurement precision and testing times were expected, results showed that the strategies did not differ much with respect to measurement precision but that there were differences with regard to the total testing time.

Introduction

Computerized adaptive testing (CAT) is increasingly being applied in both educational and psychological testing, because tailoring the test to the performance of the respondent reduces test length, increases the motivation of the test taker to provide responses at his or her cognitive level, and avoids distracting the test taker with questions that are too easy or too difficult (Wainer, Dorans, Flaugher, Green, & Mislevy, 2000). Moreover, computerized adaptive tests (CATs) can be offered at any time, at various locations, or even via the Internet. In general, CATs are administered by applying the following pseudo-algorithm.

1. The performance level of the respondent is initialized: for example, at the mode of the ability distribution, or based on historical data.
2. Based on the estimated performance level, the next item is selected that provides optimal information at the current performance level estimate. Generally, Fisher information is applied for item selection, but other item-selection criteria have been proposed and studied.
3. The selected item is administered to the test taker.
4. After the item is administered, the ability estimate of the test taker is updated based on the response provided.
5. The process of item selection, item administration, and updating the ability estimates continues until a stopping criterion is met. For some applications, the stopping criterion is related to the number of items administered; for other applications, it depends on the precision of the performance level estimate.

In general, in comparison with fixed-length tests, CAT has achieved reductions in test length varying from 22% (Eggen & Straetmans, 2000) to 50% (Hornke, 1999). Many variations, modifications, and extensions of this basic algorithm for CAT have been studied. In the case of CATs developed to measure maximum performance, exposure control measures are added to the algorithm to prevent overexposure of the most informative items and to prevent the security risks involved (e.g., Barrada, Abad, & Veldkamp, 2009). Another extension is related to the question of how to deal with test

specifications. Content specifications, response time constraints, targets for the information function, or other specifications can be imposed to regulate item selection. Taking all of these specifications into account, the item-selection step of the pseudo-algorithm needs to be modified in order to ensure that the resulting CATs meet all the specifications imposed. An efficient way of dealing with specifications is to apply the shadow test approach (STA), first proposed in van der Linden and Reese (1998). In the STA, the item-selection step of CAT is extended. Instead of finding the item that provides the most information at the current ability estimate, a two-step procedure is applied. First, a full-length test that contains all the items already administered and meets all the test specifications—a so-called shadow test—is assembled in every iteration of CAT. Next, the optimal item is selected from this shadow test. Since items are selected from a shadow test that meets all the constraints, it is guaranteed that each CAT meets all the test specifications. So, the STA prevents infeasibility problems and ensures that all individual CATs are comparable with respect to the specifications.

Response times pose a number of specific challenges for CAT. The reason is that different test takers respond to different items in CAT, and some items are more time intensive than others. As a consequence, differential speededness can occur. For some test takers, the CAT might become speeded, whereas for others, the CAT is closer in similarity to a pure power test (e.g. van der Linden, 2006). This phenomenon has to be prevented during CAT assembly. Moreover, Fan, Wang, Chang, and Dougles (2012) noticed that the most informative item at the current ability estimate is not always the most efficient item to be selected for administration, because the most informative item might be very time consuming. Moreover, at the beginning of a CAT, both the ability and the speed estimate might still be off target. As a consequence, it might not be the best strategy to select items that perform optimally at the current ability and speed estimates. Finally, in many CAT applications, the test doesn't necessarily have to be finished as fast as possible. The goal is more that the test takers be able to finish the test within a reasonable amount of time and without too much variation in time across test takers.

The purpose of this report is to study various strategies for dealing with these challenges. In order to study the strategies, a general model for item selection in CAT is presented first.

Item Selection in CAT

Item selection in CAT has been the topic of many studies. Various item-selection criteria, based on, for example, Fisher information, Kullback–Leibler divergence, or alpha-stratification, within either a frequentist or a Bayesian framework, have been proposed. In operational settings, maximum Fisher information is most commonly applied, since (a) it is asymptotically related to the inverse of the measurement precision, (b) it is rather straightforward to compute, and (c) the differences in performance between the various item-selection criteria have turned out to be small. When maximum Fisher information is applied for item selection, the measurement precision of the CAT is optimized.

In many operational CATs, a counterbalancing mechanism is needed to guarantee that specifications (e.g., related to the content of the test) are met. Van der Linden and Reese (1998) proposed the STA for item selection in CAT with test specifications. In their approach, a full-length shadow test is assembled during each iteration of the CAT that contains all previously administered items and optimizes an objective function (e.g., maximizing measurement precision) while test specifications are met.

To model the item selection in CAT, some notation has to be introduced first.

Variables

x_i Whether item i is selected for the test.

Parameters

I Number of items in the pool.

$I_i(\theta)$ Amount of information item i provides at the estimated ability level θ.

$g-1$ Number of items already administered.

n Test length.

R_{g-1} Set of items administered so far.

V_c Subset of items belonging to category c.

b_c Bound for the number of items to be selected for category c.

q_i Amount item i contributes to constraint q.

b_q Bound for constraint q.

V_l Subset of items affected by constraint l.

n_l Bound for constraint l.

Following van der Linden (2005), a 0-1 LP model for selecting the g^{th} item in CAT can be formulated as

$$\max \sum_{i \in I \setminus R_{g-1}} I_i(\theta) \tag{1}$$

subject to

$$\sum_{i \in R_{g-1}} x_i = g-1, \tag{2}$$

$$\sum_{i=1}^{I} x_i = n \tag{3}$$

$$\sum_{i \in V_c} x_i \leq b_c \quad \forall c, \tag{4}$$

$$\sum_{i=1}^{I} q_i x_i \leq b_q \quad \forall q, \tag{5}$$

$$\sum_{i \in V_l} x_i \leq n_l \quad \forall l, \tag{6}$$

$$x_i \in \{0,1\}. \tag{7}$$

The amount of information in the test at the current ability estimate is maximized in (1). In (2) it is imposed that all previously administered items be in the shadow test. The length of the shadow test is defined in (3). Van der Linden (2005) distinguishes among specifications related to categorical attributes, such as content category; quantitative specifications that are related to quantitative attributes, such as word count; and specifications about relationships between items, such as enemy sets. These three categories of specifications result in three types of constraints. Categorical constraints are imposed in (4), quantitative constraints in (5), and logical constraints in (6). In this model, the constraints have been formulated in terms of upper bounds, whereas they can refer to lower bounds as well. The decision variables x_i, denoting whether an item is in the test $(x_i = 1)$ or not in the test $(x_i = 0)$, are defined in (7). Once a shadow test has been assembled, the unadministered item in this shadow test that provides the most information at the current ability estimate is selected for administration.

Response time constraints might be imposed on CAT for practical, psychometric, or security reasons. The testing location might only be available during a fixed time slot (practical), the construct of interest might be the ability to solve the task within a reasonable amount of time (psychometric), or response patterns might be analyzed for cheating (security). For an overview of applications of response times in CAT, see van der Linden (2008). Response time constraints can be formulated at the group or person level. At group level, constraints can be imposed concerning, for example, the percentage of test takers who will be able to finish the whole test (e.g., “85% of test takers should finish all the items”) or the percentage of the total test length each test taker should be able to complete (e.g., “100% of the test takers should complete at least 90% of the items”). When response time constraints are imposed at the individual level, other constraints can be imposed (e.g., “the sum of observed response times must be smaller than the total test time”). The question remains: How do we take these kinds of constraints into account during test assembly?

Various Strategies for Dealing With Response Times in CAT

Two different strategies for taking response times into account during CAT administration have been proposed in the literature. In the first strategy, constraints with respect to response times are imposed on item selection. In the second strategy, the amount of information provided by the items is adjusted for the time needed to respond to the items. In this section, both strategies are described and two new strategies are introduced.

Strategy 1: Constraints With Respect to Response Times

In practice, response time specifications can be modeled as quantitative constraints at the individual level. Van der Linden, Scrams, and Schnipke (1999), for example, proposed imposing an additional constraint on the item-selection procedure that guides the item selection with respect to the response times:

$$\sum_{i \in R_{g-1}} t_{ij} x_i + \sum_{i \in I \setminus R_{g-1}} E\left[t_{ij} \right] x_i \leq t_{tot}. \tag{8}$$

In this constraint, the sum of the times t_{ij} spent on answering the previous items $i \in R_{g-1}$ by person j and the expected times on the remaining items has to be lower than the total amount of time t_{tot} available for the test.

To calculate the expected response times, van der Linden (2007) proposed a hierarchical framework for modeling speed and accuracy. When it comes to response time modeling, a distinction is often made between power tests and speed tests. In the first type of test, there is no time limit and the test taker just has to demonstrate his or her capability of solving items that vary in difficulty. In a speed test, the items are supposed to be equally difficult, and the task is to solve as many items as possible within a given time frame, or to minimize the time needed to complete a fixed number of items. For CAT, we typically have a mix of power and speed testing. The difficulty of the items varies and is adapted to the ability level of the test taker; however, the test taker has to solve the items within a reasonable amount of time.

Within the hierarchical framework, a normal ogive model is proposed for the responses, a log normal model is chosen to deal with the response times, and to model the joint distribution of both person and item parameters, a bivariate normal distribution is assumed. For this log normal response time model, with a time intensity and time discrimination power of item j denoted by β_j and α_j, respectively, the expected response time of a test taker with an estimated working speed $\hat{\tau}$ is equal to

$$E\left(t_{ij} \mid \hat{\tau}_j \right) = \exp\left(\beta_i - \hat{\tau}_j + \frac{1}{2\alpha_i^2} \right) \tag{9}$$

where

$$\hat{\tau}_j = \frac{\sum_{i \in R_{g-1}} \alpha_i^2 \left(\beta_j - \log t_{ij} \right)}{\sum_{i \in R_{g-1}} \alpha_i^2} \tag{10}$$

(Fan et al., 2012; van der Linden, 2006).

Application of Strategy 1 to adaptive testing is described in van der Linden (2008). An alternative approach would be to estimate the expected response time of a test taker based on, for example, the average response time for the whole population.

Strategy 2: Information Per Time Unit

In the Introduction section, it was mentioned that it might happen in practice that some very informative items take a relatively large amount of time to answer. Because of this, it might be a more efficient strategy to administer a larger number of less informative but also less time consuming items within a fixed time slot. With respect to this observation, Fan et al. (2012) proposed selecting items based on maximum information per time unit. The criterion, or objective function in the model, for selecting the g^{th} item, can now be formulated as

$$\max_{i \in I \setminus R_{g-1}} \left\{ \frac{I_i\left(\hat{\theta}_j\right)}{E\left[t_{ij} \mid \hat{\tau}_j\right]} \right\}. \tag{11}$$

In this item-selection criterion, the time required to answer item i is calculated based on the expected response time under the log normal response model, given the time intensity of the item and the estimated speed of the test taker. With this selection criterion, highly informative items will still be chosen, unless they require a relatively large amount of time. At the beginning of the CAT, the estimates of both the speed parameter and the ability parameter will still be rather unstable. But since they are updated during administration, the expected information per time unit will become more and more accurate. An alternative approach would be to use the average time over the whole population as a proxy for the required time.

Strategy 3: Penalized Violations of Maximum Response Time (Deviations)

A third strategy is based on the goal programming or penalized deviations strategy for dealing with test specifications (e.g. Veldkamp, 1999). Goal programming also underlies the normalized weighted absolute deviation heuristic (NWADH algorithm) for CAT (Luecht, 1998) and the weighted deviation method (Stocking & Swanson, 1993). In this strategy, a goal or target is set for specifications (e.g., with respect to the total response time), and the selection of items that result in exceeding this target is penalized. Especially in cases where response time specifications have been formulated at the population level, it is acceptable for a small percentage of test takers to exceed the maximum response time. By imposing a penalty, this percentage is minimized. The criterion, or objective function in the model, for selecting the g^{th} item can now be formulated as

$$\max \sum_{i \in I \setminus R_{g-1}} I_i\left(\hat{\theta}\right) x_i - P^* \max\left\{ \left(\sum_{i \in R_{g-1}} t_{ij} x_i + \sum_{i \in I \setminus R_{g-1}} E\left[t_{ij} \mid \hat{\tau}_j\right] x_i \right) - t_{tot.}, 0 \right\} \tag{12}$$

where P is the penalty imposed on exceeding the maximum total response time t_{tot}.

Strategy 4: Robust CAT

The fourth strategy is based on the observation that whenever specifications are related to person attributes, such as response speed or ability, the definition of the

specifications changes during test administration, because more and more information becomes available about the test taker's attributes as the test progresses, and the attributes can then be estimated more precisely with each response. In the early stages of CAT, hardly any information about speed is available. As a consequence, the speed estimate $\hat{\tau}_j$ might be seriously biased, and the expected response time will be overestimated or underestimated. When the expected response time is underestimated, the shadow tests that were assembled during earlier iterations of the CAT might not be feasible anymore. As a result, the test taker might need more time than expected and might run into time trouble toward the end of the test.

One way of dealing with uncertainty in the speed parameter at the beginning of the CAT is to apply robust item selection methods. These methods take the uncertainty into account by selecting the items based on a conservative estimate of the parameters involved. Bertsimas and Sim (2003) argued that since uncertainty is normally distributed, it only has a large impact on the final solution for a limited number of items. In their approach, uncertainty is assumed to affect the solution for at most a limited number of items. They also demonstrated that a linear optimization problem with uncertainties in the coefficients can be solved as a series of linear optimization problems with fixed coefficients. Veldkamp (2013a) illustrated how this method can be applied to automated test-assembly problems with uncertainties in the item parameters.

Veldkamp (2013b) proposed a pseudo-algorithm that describes the application of the Bertsimas and Sim (2003) method to CAT when uncertainty in the objective function of the item selection model has to be taken into account. The present report addresses uncertainty in expected response times. This uncertainty plays a role in the test specifications, rather than in the objective function. For this purpose, a modified version of the Veldkamp (2013b) pseudo-algorithm is proposed. Let uncertainty play a role in at most Γ of the items:

1. Rank the items such that $I_1\left(\theta_{g-1}\right) \geq I_2\left(\theta_{g-1}\right) \geq \ldots \geq I_n\left(\theta_{g-1}\right)$.

2. Calculate $d_i = E\left[t_{ij} \mid \hat{\tau}_j\right] - E\left[t_{ij} \mid \hat{\tau}_j^{robust}\right]$ for all items.

3. For $l = 1,\ldots,\left(G-\left(g-1\right)\right)+1$, find the item that solves:

$$G^l = \max \sum_{i \in I \setminus R_{g-1}} I_i\left(\theta\right) x_i \tag{13}$$

subject to:

$$\sum_{i \in R_{g-1}} t_{ij} x_i + \sum_{i \in I \setminus R_{g-1}} E\left[t_{ij} \mid \hat{\tau}_j\right] x_i + \left[\sum_{I=1}^{l} \left(d_i - d_l^*\right) x_i + \min\left(G-g,\Gamma\right) d_l^*\right] \leq t_{tot.} \tag{14}$$

$$x_i \in \{0,1\} \quad i = 1,\ldots,I, \tag{15}$$

where $d_l^* = min_{i \le l} \{ d_i \}$.

4. Let $l^* = arg\max G^l$.

5. Item g is the best unadministered item in the solution of G^{l^*}.

In this pseudo-algorithm, the response time constraint is corrected for uncertainty in Γ of the items by adding the term between brackets. This term adds d_l^*, the minimum difference between $E\left[t_{ij} \mid \hat{\tau}_j \right]$ and a robust estimation of the expected response times $E\left[t_{ij} \mid \hat{\tau}_j^{\text{robust}} \right]$, for the l most informative items. If one of the l most informative items is chosen, the penalty is even larger. For a more elaborated discussion and proof of the model, see Bertsimas and Sim (2003). By applying the pseudo-algorithm for item selection in CAT with response time constraints, the uncertainty in the estimated speed parameter is accounted for. The only cost is that a series of l integer programming problems has to be solved in the item selection step instead of just one.

Numerical Example

Item Bank

For this example, an operational item bank for admission testing was used. Using BILOG-MG 3, a three-parameter logistic model (3PLM) was used to calibrate 306 items, yielding a sample of 41,500 test takers. The estimated item parameters ranged from $\hat{a}_i \in [0.22, 1.22]$, $\hat{b}_i \in [-3.14, 2.24]$, and $\hat{c}_i \in [0.0, 0.49]$. Since response times were not available for these items, they were simulated. A log normal model for response times was assumed. The time discrimination parameters α_i and time intensity β_i for the items were simulated. Following Fan et al. (2012), the parameters α_i were simulated from $\alpha_i \sim$ uniform[2,4] and the parameters $\beta_i \sim N(0,1)$. No correlations were assumed between time intensity and difficulty of the items.

Persons

For each ability value θ_j in [−2,−1,5,…,2], 250 persons were simulated. For these persons, the speed parameter τ_j was sampled from a standard normal distribution. In this way, 2,250 persons were simulated.

Settings of the Study

In this study, maximum Fisher information was applied for item selection. The test length was set equal to 25 items. The total time available was set equal to $t_{tot} = 45$. This time limit was based on a small simulation study where the average time for answering an item in the bank turned out to be equal to 1.8 minutes. When the response time

constraint was violated, because the simulated test taker needed more time than the total time available, the CAT simulation continued to administer items, even though infeasibility occurred, and the total response time was still recorded. Maximum likelihood estimation was implemented for estimating the ability parameters. The closed expression in (9) was applied to update the speed parameter estimates. To apply the various strategies, some parameters had to be set. Strategy 1 and 2 do not have any additional parameters. For the penalized deviations strategy (Strategy 3), the penalty P was chosen in such way that both the information component and the penalty component were of equal magnitude in the item selection criterion. For the robust CAT strategy (Strategy 4), the Γ parameter was set equal to 40% of the test length. Finally, a condition that maximized Fisher information without taking response times into account during test assembly was added as a reference.

Software

The open source statistical software package R, version 3.1.0 (CRAN, 2014), was applied to develop a CAT with response times. We used the LPSolve package (CRAN, 2014). This package is an interface called lp_solve version 5.5, and offers a limited number of functions for solving mixed-integer programming problems.

Results

Four different strategies for dealing with response times in CAT were compared. A fifth strategy—one that did not take response times into account during item selection—was added as a benchmark. The mean ability estimates for various strategies are shown in Table 1.

TABLE 1
Mean ability estimate for various ability levels and various item selection strategies

True Ability	Fisher Information	FI Per Time Unit	Penalized Deviation	Constrained CAT	Robust CAT
–2	–2,034	–2,030	–2,028	–2,024	–1,980
–1,5	–1,550	–1,556	–1,534	–1,504	–1,520
–1	–1,026	–1,010	–1,020	–1,004	–1,040
–0,5	–,584	–,592	–,528	–,488	–,540
0	,038	,036	–,010	–,016	,004
0,5	,498	,476	,474	,518	,474
1	,966	,982	,962	,990	,950
1,5	1,502	1,498	1,506	1,532	1,540
2	1,986	2,010	2,078	2,020	2,024

FI = Fisher information.

As can be seen in Table 1, there is hardly any bias. No significant differences between strategies were found. The mean square errors (MSEs) of the ability estimates are shown in Figure 1. The MSEs are also relatively small for all strategies. They vary over the ability points, but no systematic pattern can be found that seems to suggest that one strategy is outperformed by the others.

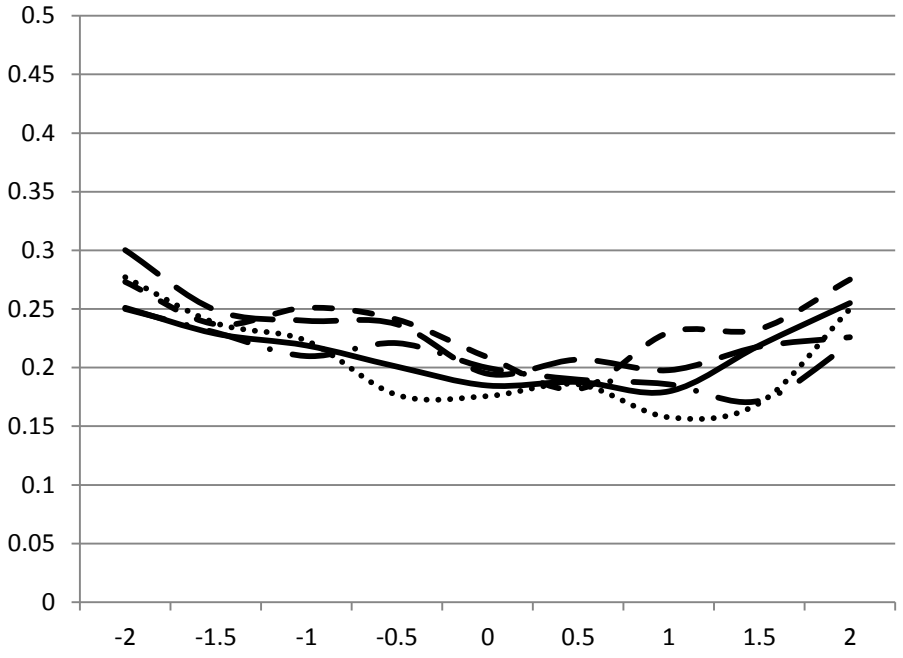

FIGURE 1: *MSE of ability estimator for Fisher information (dash/dotted), Fisher information per time unit (large dashes), Penalized deviation (dotted), Constrained CAT (solid), and Robust CAT (small dashes)*

The average response times for the various strategies are shown in Table 2.

TABLE 2
Average response times for various strategies for item selection

Strategy	Average RT
Constrained CAT	42,6
FI per time unit	38,1
Penalized deviations	46,9
Robust CAT	43,3
FI	46,4

FI = Fisher information; RT = response time.

When response times were not taken into account, the average response time was equal to 46,4. Maximizing Fisher information per time unit resulted in the shortest test. For the constrained CAT strategy, the average response time was smaller than the total time available. The robust CAT strategy needs on average slightly more time, but performs almost as well as the constrained CAT. The penalized deviations strategy resulted in the longest tests. Even though violation of the time constraint is allowed in this strategy, it is surprising that the average testing time exceeds the average testing time of the benchmark strategy. A more detailed picture is revealed in Figure 2.

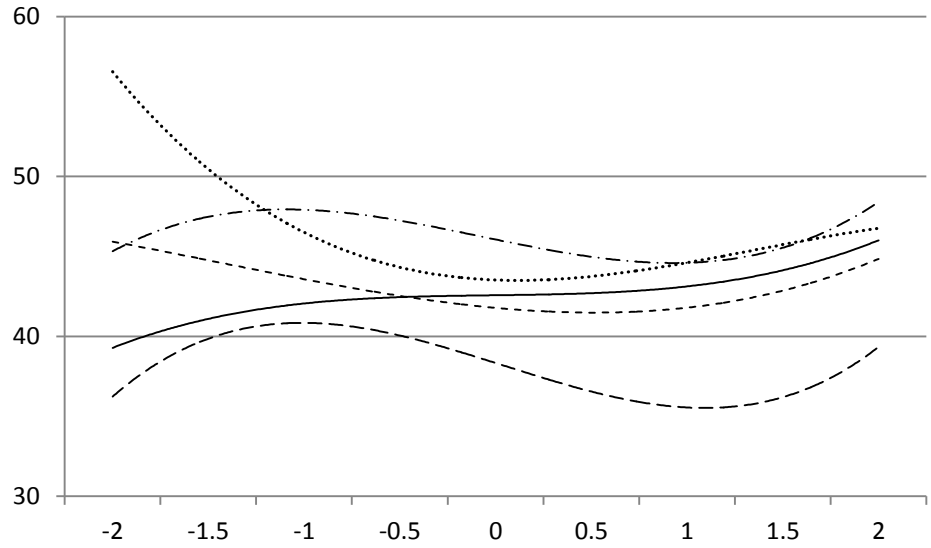

FIGURE 2: *Average response times per ability value for Fisher information (dash/dotted), Fisher information per time unit (large dashes), Penalized deviation (dotted), Constrained CAT (solid), and Robust CAT (small dashes)*

On average, the order of the various strategies is rather stable over the various ability values. For $\theta = -2.0,$ however, some different results were obtained. Especially for the penalized deviations strategy, some outliers were simulated that caused the average response time to increase considerably.

Besides the total response time for the various strategies, the number of violations of the time constraint was also counted. The percentage of violations is denoted in Table 3.

TABLE 3
Percentage of time constraint violations per strategy

Strategy	Violations (%)
Constrained CAT	26
FI per time unit	24
Penalized deviations	28
Robust CAT	24
FI	32

FI = Fisher information; RT = response time.

The percentage of constraint violations is rather high. Apparently, the time limit of $t_{tot} = 45$ turned out to be too restrictive for a large portion of the test takers. Selecting items based on Fisher information resulted in the greatest number of violations. Fisher information per time unit and robust CAT yielded the smallest number of violations.

Discussion

The R package was applied to simulate a CAT with response time constraints. Both the response model and the response time model were implemented and data were simulated using the normal distribution, the uniform distribution, and the log normal distribution functions that are available within the package. In order to deal with the 0-1

linear programming problems, the LPSolve package was used. This package enables the implementation of a variety of test specifications on CAT. The procedures for both the CAT and the response time model were able to be implemented straightforwardly. The R package is very user friendly. The only drawback is that the simulation studies turned out to be rather time consuming, even though only small numbers of test takers were simulated. In the current simulation study, only response time constraints were implemented. The question arises as to whether straightforward implementation of CAT in the R package is feasible for dealing with a large number of test specifications, such as the full set of constraints implemented by the LSAT. More advanced implementations with parallel computation on multiple processors may be necessary to achieve this.

In this study, several strategies for item selection in CAT with response times were compared. From a theoretical point of view, maximizing Fisher information, which was used as a benchmark, was expected to result in the longest response times with the highest measurement precision. Maximizing Fisher information per time unit was expected to result in the shortest response times, with a small loss in precision. Application of constrained CAT was expected to result in response times close to the total testing time, with few violations due to overestimation of the speed of the test taker in the beginning of the CAT. Robust CAT was expected to result in a slightly shorter test due to a more conservative estimation of the speed parameter, with a small loss of information in comparison with constrained CAT. Finally, the penalized deviations strategy was expected to result in a greater number of violations of the response time constraint, with slightly higher measurement precision.

The expected differences in measurement precision were not confirmed in the simulation study. For this study, we also have to conclude that with respect to measurement precision no differences in item selection strategies were found. However, these results are in line with other studies that compare item-selection criteria in CAT where also hardly any differences were reported for test lengths equal to or larger than 25.

With respect to response times, most of the expectations were confirmed. Maximizing Fisher information per time unit resulted in tests that are much shorter than the time limit. Robust CAT resulted in slightly shorter tests than constrained CAT. For the benchmark condition, most violations of the response time constraint were observed. Only with respect to the penalized deviations strategy, our expectations were not confirmed. Somehow, this strategy resulted in much longer tests than expected. A more detailed inspection of the results revealed that some very high response times were observed for a number of simulated test takers with ability level $\theta = -2.0,$ which disturbed the comparison.

One remark should be made with respect to the way the response time constraints were implemented in this study. When the response time constraints were first described, it was mentioned that they can be formulated at the group level and at the individual level. In the simulation study, response time constraints were only implemented at the individual level by fixing the total response time for each test taker. Imagine a test taker working at a very slow pace. The item bank might not contain enough items with a low time intensity at this test taker's level to administer a CAT that can be finished within the total testing time. Formulating response time constraints at the group level (e.g., specifying that 85% of the test takers must be able to finish the whole test) accounts for the fact that the working speed of some test takers might just be too low to finish in time. By adding constraints at the group level, infeasibility problems (e.g., Huitzing, Veldkamp, & Verschoor, 2005) might be prevented.

Finally, this study mainly focused on various strategies for dealing with response times in CAT. Once response times have been implemented, not only can they be applied to ensure that the test takers finish the test in time, but they can also be used to detect aberrant response behavior.

Van der Linden and Guo (2008) successfully applied response times for identifying aberrant response behavior in CAT. Response times are more informative than responses, since they are continuous variables. They may also be harder to simulate by cheating respondents, even though aberrant respondents—or even groups of aberrant respondents—might try to fake realistic response times. Van der Linden and van Krimpen-Stoop (2003) first proposed identifying aberrant behavior using response times. Their method was based on the response time model proposed in van der Linden et al. (1999). Van der Linden and Guo (2008) used the hierarchical model for both responses and response times, and that was also applied in this study. Recently, Marianti, Fox, Avetisyan, and Veldkamp (2014) and Fox (2014) described how response times can be used successfully for identifying various kinds of aberrant behavior. A next step would be to implement these methods in CAT. For example, when aberrant response behavior is detected during CAT administration and item preknowledge is suspected, the remaining items might be selected from a “secret” item pool that had not been applied before, and test length might be increased. In this way, a different kind of adaptation could possibly be applied during CAT administration.

References

Barrada, J. R., Abad, F. J., & Veldkamp, B. P. (2009). METODOLOGÍA: Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema, 21,* 313–320.

Bertsimas, D., & Sim, M. (2003). Robust discrete optimization and network flows. *Mathematical programming, 98*, 49–71.

CRAN (2014). *Statistical software package R.* Retrieved from http://cran.r-project.org/.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological measurement*, *60*, 713–734.

Fan, Z., Wang, C., Chang, H. H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37,* 655–670.

Fox, J.-P. (2014). Modeling differential working speed in assessment testing. *Manuscript submitted for publication.*

Hornke, L. F. (1999). Benefits from computerized adaptive testing as seen in simulation studies. *European Journal of Psychological Assessment*, *15*, 91.

Huitzing, H. A., Veldkamp, B. P., & Verschoor, A. J. (2005). Infeasibility in automated test assembly models: A comparison study of different methods. *Journal of Educational Measurement*, *42*, 223–243.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22,* 224–236.

Marianti, S., Fox, J.-P., Avetisyan, M., & Veldkamp, B.P. (2014). *Testing for aberrant behavior in response time modeling* (LSAC Report Series, RR 14-02). Newtown, PA: Law School Admission Council.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

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

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31,* 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72,* 287–308.

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

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*, 259–270.

van der Linden, W.J., Scrams, D.J., & Schnipke, D.L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195–210.

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*, 251–265.

Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement, 36,* 253–266.

Veldkamp, B. P. (2013a). Application of robust optimization to automated test assembly. *Annals of Operations Research, 206*, 595–610.

Veldkamp, B. P. (2013b). Ensuring the future of CAT. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 137–150). Enschede: RCEC.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer.* Routledge.