ED 481 158 TM 035 240

AUTHOR	van der Linden, Wim J.; Scrams, David J.; Schnipke, Deborah L.
TITLE	Using Response-Time Constraints in Item Selection To Control for Differential Speededness in Computerized Adaptive Testing. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-RR-98-03
PUB DATE	2003-09-00
NOTE	16p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; Algorithms; *Computer Assisted Testing; Linear Programming; Responses; *Selection; *Test Items; *Timed Tests
IDENTIFIERS	*Constraints

ABSTRACT

This paper proposes an item selection algorithm that can be used to neutralize the effect of time limits in computer adaptive testing. The method is based on a statistical model for the response-time distributions of the test takers on the items in the pool that is updated each time a new item has been administered. Predictions from the model are used as constraints in a 0-1 linear programming model for constrained adaptive testing that maximizes the accuracy of the ability estimator. The method is demonstrated empirically using an item pool from an operational, large-scale computer adaptive test. (Contains 4 figures and 14 references.) (Author/SLD)

LSAC RESEARCH REPORT SERIES

■ Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing

Wim J. van der Linden
University of Twente

David J. Scrams
Virtual Psychometrics

Deborah L. Schnipke
Virtual Psychometrics

■ Law School Admission Council
Computerized Testing Report 98-03
September 2003

LSAC®

A Publication of the Law School Admission Council

LSAC RESEARCH REPORT SERIES

■ Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing

Wim J. van der Linden
University of Twente

David J. Scrams
Virtual Psychometrics

Deborah L. Schnipke
Virtual Psychometrics

■ Law School Admission Council
Computerized Testing Report 98-03
September 2003

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 201 law schools in the United States and Canada.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

Test takers tend to differ from one another in the amount of time required to respond to items. This is true even among test takers of the same ability level. Although this finding is not surprising, it may lead to a serious scoring problem. If some test takers do not complete all test items, the test-scoring procedure must include a provision for unreached items. Such items could be treated as incorrect (e.g., a test taker's final score could be influenced by the number of unreached items) or unreached items could be ignored (i.e., treated as missing). This decision should be made according to beliefs about the independence and relative importance of response speed and response accuracy in the context of the test.

If speed and accuracy are independent and the test is designed to measure accuracy, test taker ability should be based on accuracy alone, and test takers should not be penalized for unreached items. If speed and accuracy are related or if both are important in the test context, response speed may be included in the scoring rubric, and unreached items would count against a test taker. In the latter case, estimates of test taker ability would reflect both response speed and response accuracy. Realistic scoring models that combine measures of speed and accuracy are not yet available, but the scant empirical research concerning the relationship between response speed and response accuracy in large-scale testing suggest that speed and accuracy are independent factors in power tests (i.e., tests that measure accuracy alone).

The best solution to the problem of unreached items may be to design the test in such a way that they do not occur or are minimized. This could be accomplished with very generous time limits (a costly solution). Computer adaptive testing (CAT), however, offers an attractive alternative. Test taker speed can be assessed along with test taker ability (measured in terms of response accuracy), and the estimated test taker speed can be included in the item-selection algorithm. Thus, items are selected for a test taker that are appropriate for the test taker's ability, but are unlikely to be so time-consuming that the test taker fails to complete all test items. This solution requires a model of response speed and an item-selection algorithm that accommodates response-speed constraints. Both aspects are addressed by the current work.

A model of response speed is used as the basis for predicting a test taker's response time for each item in the item pool. Items are selected according to an algorithm for constrained CAT. The item-selection algorithm constrains item selection so that the test taker is likely to have sufficient time to answer all items while simultaneously insuring that test specifications are met and all test takers receive items that are tailored to their ability level. Response-time predictions are modified according to the time taken by the test taker to respond to items already administered. Analyses of operational data from a large-scale standardized test support the use of the response-speed model, and simulations of the item-selection algorithm demonstrate that response-time constraints could be included in item selection while maintaining test quality.

The present approach to adaptive item selection is a solution to the scoring problems introduced by differences in response speed across test takers. This solution may be preferable to the obvious alternatives of reduced test length (with a reduction in measurement precision) or increased time limits (with added administration costs). The preliminary results reported here demonstrate the reasonableness of the response-speed model and the feasibility of including response-time constraints in item selection.

Abstract

Test takers with the same ability differ in the amount of time they need to complete a test item. Therefore, some test takers may be affected unfavorably by the presence of a time limit on a test. This paper proposes an item-selection algorithm that can be used to neutralize the effect of time limits in computer adaptive testing. The method is based on a statistical model for the response-time distributions of the test takers on the items in the pool that is updated each time a new item has been administered. Predictions from the model are used as constraints in a 0–1 LP model for constrained adaptive testing that maximizes the accuracy of the ability estimator. The method is demonstrated empirically using an item pool from an operational, large-scale computer adaptive test.

Introduction

Test takers generally need different amounts of time to complete the same item in an educational or psychological test even if they have the same ability. As a consequence, some test takers may finish the test in time whereas others do not reach the items at the end of the test. However, unreached items involve a

This paper was written while the first author was on a sabbatical leave at the Law School Admission Council (LSAC), the second author was a summer intern at Educational Testing Service, and the third author was a Research Scientist with LSAC. The authors are indebted to Wim M. M. Tielen for adapting the software used in the study.

serious scoring problem. First of all, in a conventional paper-and-pencil test it is impossible to discriminate between unreached items and reached items that were left unanswered because their answers were unknown. But even if it were exactly known which items were not reached (as is possible in computerized testing), scoring would remain a complicated issue. If the test is designed such that speed and ability are independent factors, responses to items not reached can be viewed as missing at random and should be ignored when the test is scored. For the notion of data that are missing at random, see Gelman, Carlin, Stern, and Rubin (1995, sect. 7.4). However, if speed and ability are dependent, the correct way to score the test is under a model that represents their joint effect on the probabilities of success on the items. Realistic versions of such models are not available yet.

Empirical studies of the relation between response time and ability have become possible through the introduction of computerized testing but are still hard to find. A favorable exception is a recent study of the response times in a field test of a computerized version of the National Board of Medical Examiners (NBME) Step 2 Licensure Exam (Swanson, Featherman, Case, Luecht, & Nungester, 1997, April). In this study, items in linear subtests were timed, and no correlation between response time and ability was found for the various subtests. However, a replication of the study for the NBME Step 1 Licensure Exam showed moderate correlation toward the end of the subtests apparently due to the use of more stringent time limits (Swanson, personal communication, December 18, 1997). These results seem to suggest that the only effective way to deal with scoring problems due to unreached items is to design the test such that they do not occur. However, given the variability in response time among test takers, for a conventional linear test this approach would imply either the need of a shorter test with loss of accuracy in ability estimation for the faster test takers or a more generous time limit for the test and thus an increase in costs.

In computer adaptive testing (CAT), an attractive solution to this test design dilemma is possible. If a model for the response-time distributions of the test takers on the items in the pool is available, the actual response times on the items administered to the current test taker can be recorded and used to update the estimates of these distributions for the remaining items in the pool. These distributions can then be used to constrain the selection of the next items in the test to give the test the same degree of speededness for all test takers. Of course, this procedure is only feasible if a model for the response-time distributions with a satisfactory fit to actual response-time data is available.

It is the purpose of this paper to present an algorithm for adaptive testing that builds on this idea. In addition to the usual update of the estimate of the ability parameter in an IRT model, a lognormal model for the response-time distributions is used to update the response-time estimates for the items in the pool. Response-time constraints are derived from these estimates and imposed on the item selection using a 0–1 linear programming (LP) algorithm for constrained adaptive testing. The following sections of this paper introduce the model and the algorithm. The algorithm is then studied empirically using an item pool and estimates of response-time parameters for a large-scale, operational adaptive test. The main purpose of the study was to ascertain the effects of the response-time constraints on the statistical properties of the ability estimator. The final section of the paper discusses some remaining aspects of the implementation of the algorithm in the practice of educational and psychological testing.

Model for Response Times

The response time of test taker *j* on item *i* is denoted by a variable T_{ij}. The variable is assumed to be random because replications of tasks by the same subject are generally known to show variation in the time needed to complete them (Luce, 1986, sect. 1.2; Townsend & Ashby, 1983, chap. 3). The following decomposition for the (natural) logarithm of T_{ij} is assumed as a model for its distribution:

$$\ln T_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij} \ , \tag{1}$$

with

$$\varepsilon_{ij} \sim N(0,\sigma^2), \tag{2}$$

where μ is the grand mean or general response time level for the item pool and population of test takers, τ_j is an effect parameter for the slowness of test taker *j*, δ_i for the amount of time demanded by item *i*, and ε_{ij} is a normally distributed residual or interaction term. Together, Equations 1–2 imply a lognormal distribution for the observed response times of a fixed test taker taking a fixed item. The model was proposed in Scrams and Schnipke (in preparation). The effect terms τ_j and δ_i are defined to have expectations equal to zero across test takers and items, respectively. The marginal distribution of log response time across test takers for a fixed item also depends on the distribution of τ. This distribution is examined as one test of the goodness of fit of the model later in this paper.

Observe that the distributions in Equations 1–2 vary in location across test takers and items but have a common variance. The last assumption is stringent but allows us to estimate the parameters in the model in a straightforward way. In addition, since the model will be used to constrain item selection using only a percentile in the upper tail of the distributions toward the end of the test, a slight misfit of the model would not seem to lead to serious item selection errors . However, whenever using the model, it should be standard practice to check its assumptions.

For future reference, note that

$$\mu \equiv E_{ij}\left(\ln T_{ij}\right), \tag{3}$$

$$\delta_i \equiv E_j\left(\ln T_{ij}\right) - \mu, \tag{4}$$

$$\tau_j \equiv E_i\left(\ln T_{ij}\right) - \mu, \tag{5}$$

$$\sigma^2 \equiv E_{ij}\left[\ln T_{ij} - \delta_i - \tau_j\right]^2. \tag{6}$$

Throughout this paper subscripts at expectation signs denote indices over which expectations are taken. A different use of the lognormal distribution as a model for response times is made in Thissen's (1983) model for timed testing. In his model, the lognormal distribution is parameterized to be dependent on the latent ability measured by the items and becomes part of the likelihood function used for estimating the test taker's ability from the joint distribution of the item scores and the response times. Thissen also found adequate fit for a series of tests, except for one which showed an overrepresentation of fast responses due to the relative easiness of its items. Other distributions used to study response times on test items are the Weibull (Roskam, 1997) and the gamma distribution (Verhelst, Verstalen, & Jansen, 1997). In a previous study, the lognormal distribution showed a good fit to the response time distributions on an item pool for an operational, large-scale CAT, outperforming the Weibull and gamma distributions (Schnipke & Scrams, 1997). These results will be further discussed below when an empirical example is presented.

IRT Model

It is assumed that the item pool has been calibrated using an IRT model. In the empirical example later in this paper, the item pool was calibrated using the 3-parameter logistic (3-PL) model. The model describes the probability of a correct response on item *i* as:

$$p_i(\theta) \equiv \text{Prob}\left\{\text{U}_\text{i} = 1 \mid \theta\right\} \equiv c_i + \left(1 - c_i\right)\left\{1 + \exp\left[-a_i\left(\theta - b_i\right)\right]\right\}^{-1}, \tag{7}$$

where θ is the unknown ability of the test taker and $a_i \in [0,\infty]$, $b_i \in [-\infty,\infty]$, and $c_i \in [0,1]$ are the discrimination, difficulty, and guessing parameter for item *i*, respectively (Lord, 1980, chap. 2).

A key quantity in IRT is Fisher's information on the unknown ability parameter. For a test of *n* items, the measure is defined as:

$$I_{U_1,\ldots,U_n} \equiv -E\left[\frac{\partial}{\partial\theta^2} \ln L(\theta \mid U_1, \ldots, U_n)\right] \tag{8}$$

In Equation 8, $L(\theta \mid U_1, \ldots, U_N)$ is the likelihood statistic associated with the (random) response vector $U_1, \ldots, U_n$. For the 3-PL model in Equation 7 it holds that

$$I_{U_1,\ldots,U_n}(\theta) = \sum_{i=1}^{n} \frac{\left(p_i'(\theta)\right)^2}{p_i(\theta)\left(1-p_i(\theta)\right)}, \tag{9}$$

with

$$p_i'(\theta) \equiv \frac{\partial}{\partial\theta} p_i(\theta) \tag{10}$$

(Lord, 1980, chap. 5). The term in Equation 9 for item i will be denoted as $I_i(\theta)$ and is the item information used in the maximum-information item selection criterion in the CAT algorithm below.

Response-Time Constraints in CAT

It is assumed that the item pool consists of items indexed by $i = 1, \ldots, I$. In addition, the CAT is assumed to consist of items indexed by $k = 1, \ldots, n$. Thus, index i_k represents the event of the ith item in the pool being administered as the kth item in the CAT. The index values of the first $k - 1$ items in the CAT are denoted by the set $S_{k-1} \equiv \{i_1, \ldots, i_{k-1}\}$. The remaining items in the pool are denoted by the set $R_k \equiv \{1, \ldots, I\} \setminus S_{k-1}$. The kth item in the test is chosen from the set R_k.

The basic idea is to update an estimate of the test taker's slowness parameter τ_j in Equations 1–2 during the test given accurate estimates of μ, item parameters δ_i, $i = 1, \ldots, I$, and the residual variance, σ^2. The improved estimates of τ_j are used to update projections of the time needed to complete each of the remaining items in the pool. The next item is then selected subject to a constraint based on these projections as well as the time available to complete the remaining portion of the test. A Bayesian framework is used to update the response time projections whereas the response time constraints are incorporated in the item selection procedure using a 0–1 linear programming (LP) model for constrained CAT that maximizes the information on θ in the test and also allows for additional constraints that can be used to guarantee its content validity.

Updating Response-Time Estimates

It is assumed that δ_i, $i = 1, \ldots, I$, and σ^2 have been estimated precisely enough to be considered as known. Estimates can easily be obtained from the response times in the calibration sample using the Equations in 4 and 6. However, if test taker j is tested, τ_j is an unknown parameter; it is assumed to have a normal prior distribution:

$$\tau_j \sim N\left(\mu_{0j}, \sigma_{0j}^2\right). \tag{11}$$

The model in Equations 1–2 yields a normal likelihood with unknown mean and known variance that has the normal distribution as its family of conjugate priors (Gelman, Carlin, Stern & Rubin, 1995, sect. 2.6). Hence, noting $\hat{\tau}_j \equiv \ln(t_{ij}) - \mu - \delta_i$, the posterior distribution of τ_j after the response times on items $i_1, \ldots, i_{k-1}$ have been recorded, is normal with mean and variance:

$$E\left(\tau_j \mid t_{i_1} j, \ldots, t_{i_{k-1}j}\right) = \left[\sigma^2\mu_{0j} + \sigma_{0j}^2 \sum_{p=1}^{k-1}\left[\ln\left(t_{i_pj} - \mu\,\delta_{i_p}\right)\right] / \left[\sigma^2 + (k-1)\sigma_{0j}^2\right]\right] \tag{12}$$

$$Var\left(\tau_j \mid \tau_{i_1} j, \ldots, t_{i_{k-1}j}\right) = \sigma_{0j}^2\sigma^2 / \left((k-1)\sigma_{0j}^2 + \sigma^2\right) \tag{13}$$

Also, the predictive density for the logarithm of the response time of test taker j on item i after items i_1, ..., i_{k-1} is normal with mean equal to the posterior mean and variance equal to the sum of the prior and posterior variances:

$$E\left(\ln T_{ij} \mid t_{i_1 j},\ldots,t_{i_{k-1} j}\right) = E\left(\tau_j \mid t_{i_p j},\ldots,t_{i_{k-1} j}\right)+\mu+\delta_i \tag{14}$$

$$Var\left(\ln T_{ij} \mid t_{i_1 j},\ldots,t_{i_{k-1} j}\right) = \sigma^2_{0j} + Var\left(\tau_j \mid t_{i_p j},\ldots,t_{i_{k-1} j}\right). \tag{15}$$

As the test takers are assumed to be exchangeable, an obvious choice for the parameters in this prior is to equate them to the mean and variance of the population of test takers:

$$\mu_{0j} \equiv E(\tau) = E_j\, E_{i|j}\left(\ln T_{ij}\right) = 0, \tag{16}$$

$$\sigma^2_{0j} \equiv \sigma^2_\tau . \tag{17}$$

for all j. Using Equations 16–17, the mean and variance in Equations 14–15 specialize to:

$$E\left(\ln T_{ij} \mid t_{i_1 j},\ldots,t_{i_{k-1} j}\right) = \mu+\delta_i+\frac{\sum_{p=1}^{k-1}\left(\ln\left(t_{i_p j}\right)-\mu-\delta_{i_p}\right)}{\sigma^2/\sigma^2_\tau+k-1} \tag{18}$$

$$Var\left(\ln T_{ij} \mid t_{i_1 j},\ldots,t_{i_{k-1} j}\right) = \frac{\sigma^2+k\sigma^2_\tau}{1+(k-1)\sigma^2_\tau/\sigma^2}. \tag{19}$$

Because δ_i, $i = 1, \ldots, I$, and σ^2 are assumed to be estimated using Equations 4 and 6, respectively, the expressions in Equations 18 and 19 have only known constants and are easy to calculate.

Let t^α_{ij} be the αth certain percentile in this posterior predictive density for $\ln(T_{ij})$ transformed back to the original time scale. The choice of item i_k will now be constrained using this percentile for all remaining items in the pool, $i \in R_k$. As will become clear later, it makes sense to choose a value for α near the middle of the density in the beginning of the test, for example, the expected value of the posterior predicted density in Equation 18, and move to percentiles in the upper tail toward the end of the test.

Constrained CAT Algorithm

The kth item is selected according to an algorithm for constrained CAT presented in van der Linden and Reese (1998). To select the initial item, the algorithm first selects a full test that meets all constraints to be imposed on the selection of items in the CAT and has maximum information at the initial ability estimate. The item actually administered is the one from the assembled test with maximum information at this ability estimate. At each next step, the test is reassembled to have maximum information at the updated ability estimate fixing the items already administered. Again, the item to be administered is selected from the new portion of the test to have maximum information at the updated ability estimate. The procedure is repeated until the last item is selected.

The fact that a full test is assembled at each step rather than a single item keeps the actual item selection feasible with respect to the set of constraints. Because both the test assembly and the selection of the individual item have the objective of maximum information, the ability estimator can be expected to be maximally informative too. All test assembly is done while the test taker takes the test and is based on a 0–1 LP model that represents all test specifications.

To discuss the model, decision variables x_i, $i = 1, \ldots, I$, are introduced that take the value 1 if item i is selected in the test and the value 0 otherwise. The total amount of time available for the CAT is denoted as t_{tot}. In addition, it is assumed that the composition of the test is constrained with respect to a variety of categorical attributes, such as content, cognitive level, and item format. These attributes partition the item pool into a collection of sets V_g, $g = 1, \ldots, G$, each of which is defined by one or more attribute values. Also, the composition of the test can be constrained with respect to several quantitative attributes a_{hi}, $h = 1, \ldots, H$, such as word counts, exposure rates, and IRT parameter values. Finally, let $\hat{\theta}_{k-1}$ be the estimate of θ after $k - i$ items have been administered.

The decision variables are used to formulate the following linear model for selecting item *k* for the current test taker:

$$\text{maximize} \sum_{i=1}^{I} I_i\left(\hat{\theta}_{k-1}\right) x_i \tag{20}$$

subject to

$$\sum_{i \in S_{k-1}} t_{ij} x_i + \sum_{i \in R_k} t_{ij}^{\alpha} x_i \leq t_{tot} \, , \tag{21}$$

$$\sum_{i \in S_{k-1}} x_i = k - 1, \tag{22}$$

$$\sum_{i=1}^{I} x_i = n \tag{23}$$

$$\sum_{i \in V_g} x_i \geq n_g^{(1)}, \; g = 1, \ldots, G, \tag{24}$$

$$\sum_{i \in V_g} x_i < n_g^{(2)}, \; g = 1, \ldots, G, \tag{25}$$

$$\sum_{i - 1}^{I} a_{hi} x_i \geq n_h^{(l)}, \; h = 1, \ldots, H, \tag{26}$$

$$\sum_{i=1}^{I} a_{hi} x_i < n_h^{(2)}, \; h = 1, \ldots, H, \tag{27}$$

$$x_i \in \{0,1\}, \; i = 1, \ldots, I. \tag{28}$$

The objective function in Equation 20 optimizes the information in the test at $\hat{\theta}_{k-1}$. The total length of the test is set at *n* items in Equation 23, whereas Equation 22 fixes the values of the decision variables of all items that have already been administered to the test taker at 1. The key constraint in this paper is the one on response times in Equation 21 which requires the remaining $n - k + 1$ items to be selected such that the sum of the αth percentiles of their predicted response time distributions plus the actual response time on the first *k–i* items not be larger than the total amount of time available. In Equations 24–25, the numbers of items with the various attribute categories are required to be between lower and upper bounds $n_g^{(1)}$ and $n_g^{(2)}$, respectively. Finally, the constraints in Equations 26–27 guarantee that the sums of values for the various quantitative attributes are between the bounds $n_h^{(1)}$ and $n_h^{(2)}$.

At step *k*, the model thus selects *n–k* + 1 new items from the set R_{k-1}. The item actually administered is the one selected from this set that is most informative at $\hat{\theta}_{k-1}$. The cycle is then repeated to select item *k* + 1.

As already noted, it makes sense to choose t_{ij}^{α} close to the means of the posterior predictive response time densities in the beginning of the test but to move toward their upper tails later on. This suggestion is motivated by the fact that the sum of the mean predicted response times is a good predictor of the actual time for a large set of items but a more conservative predictor is needed if the set becomes smaller.

Other types of constraints can be added to the model to deal with possibly remaining test specifications. Several examples of possible additions are given in van der Linden (1998) and van der Linden and Reese (1998). The set of specifications should be large enough to guarantee a CAT with satisfactory content validity.

Empirical Example

A previous pool from an operational, large-scale CAT was used to study the behavior of the algorithm. The pool consisted of 186 items used for the adaptive version of an arithmetic reasoning test. The length of the test was 15 items. The items in the pool were calibrated using the 3-PL model in Equation 7.

Response-time data were recorded for 38,357 test takers who took the test in 1997.

The parameters in the response-time model in Equations 1–2 were estimated substituting sample statistics into Equations 3–6. The following results were obtained: $\hat{\mu} = 4.093, \hat{\sigma} = .515$, whereas the estimated item and person effects, $\hat{\delta}_i$ and $\hat{\tau}_j$ were distributed about zero with a standard deviation equal to .497 and .344, respectively. The correlation between $\hat{\theta}$ and $\hat{\tau}$ was equal to .035 indicating that ability and speed were independent variables for this test.

Note that these standard deviations overestimate the standard deviations of the distributions of the true effects. However, because the sample of test takers is large, the bias in $\hat{\sigma}_\delta$ is negligible. Moreover, this estimate is only reported here as a descriptive statistic; it does not play any role in the adaptive procedure in this example. The bias in $\hat{\sigma}_\tau$ is expected to be larger but this quantity serves as the variance of the prior for τ_j in the adaptive procedure. The result is thus a less informative prior, and hence a more conservative adaptive test. Also, note that these parameters were estimated ignoring the missing entries in the data matrix. This procedure is not assumed to yield biased estimates since ability and speed were estimated to be independent variables.

The main goal of the study was to estimate the effect of the response-time constraints in Equation 21 on the statistical properties of the ability estimator. In particular, the bias and root mean squared error (RMSE) functions of the estimator were studied for various values of τ.

The ability values of the simulated test takers were chosen to be equal to $\theta = -1.5, -0.5, 0.5$, and 1.5. In addition, the response times were simulated at the values of $\tau = -.60, -.30, .00, .30$, and .60. Recall that the values of τ in the empirical example of test takers were estimated to be distributed about .00 with a standard deviation equal to .515. The simulated τ values thus cover the range of values in the sample of test takers. The number of replications for each combination of the θ and τ values was equal to 120. The ability estimator used was the expected value of the posterior ability parameter (EAP estimator) with a uniform prior distribution. The first item was selected to be optimal at $\theta = 0$; this selection is known to introduce a more favorable RMSE at this ability value and more unfavorable RMSEs toward the ends of the ability scale. The LP model was solved using the First Acceptable Integer Solution Algorithm from the ConTEST test assembly software package; a detailed description of the algorithm is available in the manual (Timminga, van der Linden, & Schweizer, 1996, sect. 6.6). On a PC with Pentium/133MHz processor the time needed to update the ability estimate, solve the LP model, and select the most informative item was always less than 1 second.

The values of the parameter t_{ij}^{α} in Equation 21 were set equal to the 50th percentile in the posterior predictive distributions for the selection of the item $k = 1$ and moved to the 95th percentile for the selection of item $k = 13$ in equal steps. The last value was maintained for items $k = 14$ and 15. The total amount of time available, t_{tot}, was set equal to 39 minutes. The same time limit was used in the actual CAT.

Fit of Response-Time Distribution

The response-time distribution in Equations 1–2 was tested for its assumption of a lognormal shape against the assumptions of a normal, gamma, and Weibull distribution for the response times. Detailed results for the current data set are given in Schnipke and Scrams (1997). Since only single observations were recorded for each item and test taker combination, the distributions of the log response times on the items were inspected pooling the data across the test takers. The distribution of the estimated test taker parameters was approximately normal. Hence, the marginal distribution of the log response times for each item is also normal. This feature was checked in depth for 30 of the 186 items. These items were selected not to involve any figures and to be answered by at least 1,000 test takers. The samples were randomly split into halves used for parameter estimation and checking the distributional assumptions. The parameters of the four candidate models for the response-time distributions were estimated using the method ML estimation.

Double probability plots were produced for each model separately for each item, with the observed cumulative probability function along the abscissa and the estimated function along the ordinate. A typical example is given in Figure 1. The lognormal distribution provided the best fit (indicated by most points falling along the diagonal), followed by the gamma, Weibull, and normal. The same essential result was found for all other items.

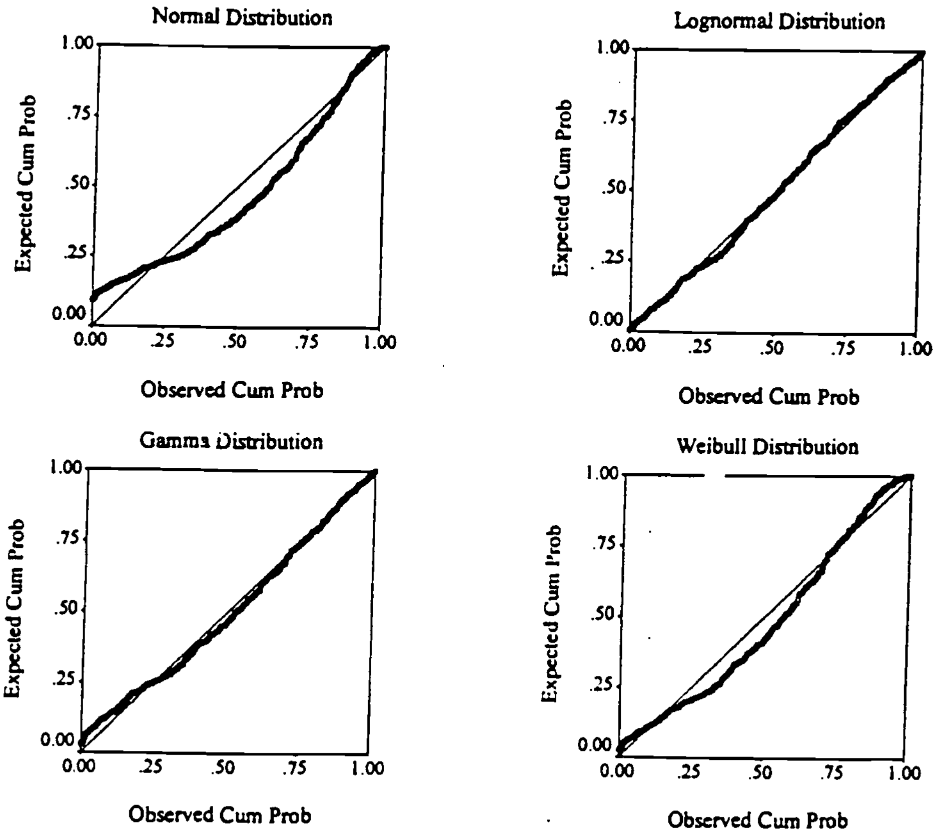

FIGURE 1. *Double probability plot of the four response-time distributions for a typical item*

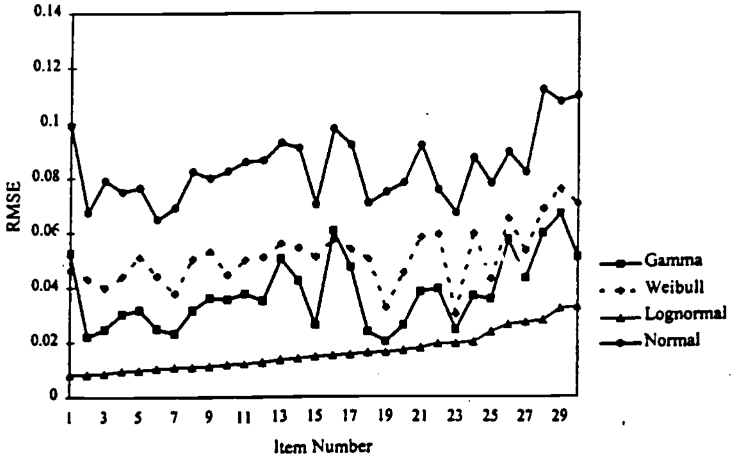

FIGURE 2. *RMSE of the items for the four response-time distributions*

Fits were also examined using the RMSE calculated between the observed and estimated distribution functions at each fifth percentile. The results for all 30 items are provided in Figure 2. For readability, items were ranked according to the quality of fit provided by the lognormal distribution. Again, the lognormal distribution provided the best overall fit, and the other three distributions were ranked as before.

The assumption of the common standard deviation across items was tested plotting the sample estimates of the standard deviations of the log response times for each of the 186 items against sample sizes in Figure 3. Different items were administered different numbers of times. For smaller samples much of the observed variability may be attributed to sampling variation. However, for larger samples, the estimated standard deviations should stabilize about an identifiable mean. Figure 3 shows this stabilization to hold indeed about the value of 0.63 estimated for the common standard deviation.

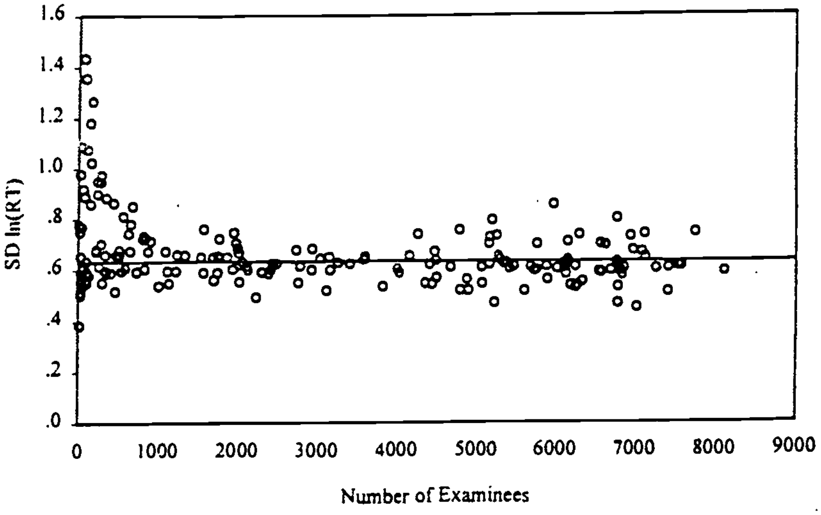

FIGURE 3. *Standard deviation of the log response times for the items as a function of sample size*

RMSE and Bias Function of Ability Estimator

The RMSE and bias functions of the ability estimator in the CAT algorithm were estimated as

$$\sum\left[(\hat{\theta}-\theta)^2\,|\theta\right]/120 \text{ and } \sum\left[\hat{\theta}-\theta\,|\theta\right]/120,$$

respectively. The results for $\tau = -.60, -.30, .00, .30$, and .60. are shown in Figure 4 (n = 15). Both for the bias and RMSE function, no systematic differences between the curves were obtained. Also, the results are of the same order as those for unconstrained adaptive testing with a variety of item-selection criteria in van der Linden (1998). The only conspicuous feature is the tendency of a slight positive bias, and hence a slightly larger RMSE, at $\theta = -1.50$. However, since the tendency is the same for all five τ values, the result is believed to be an artifact due to the composition of the item pool.

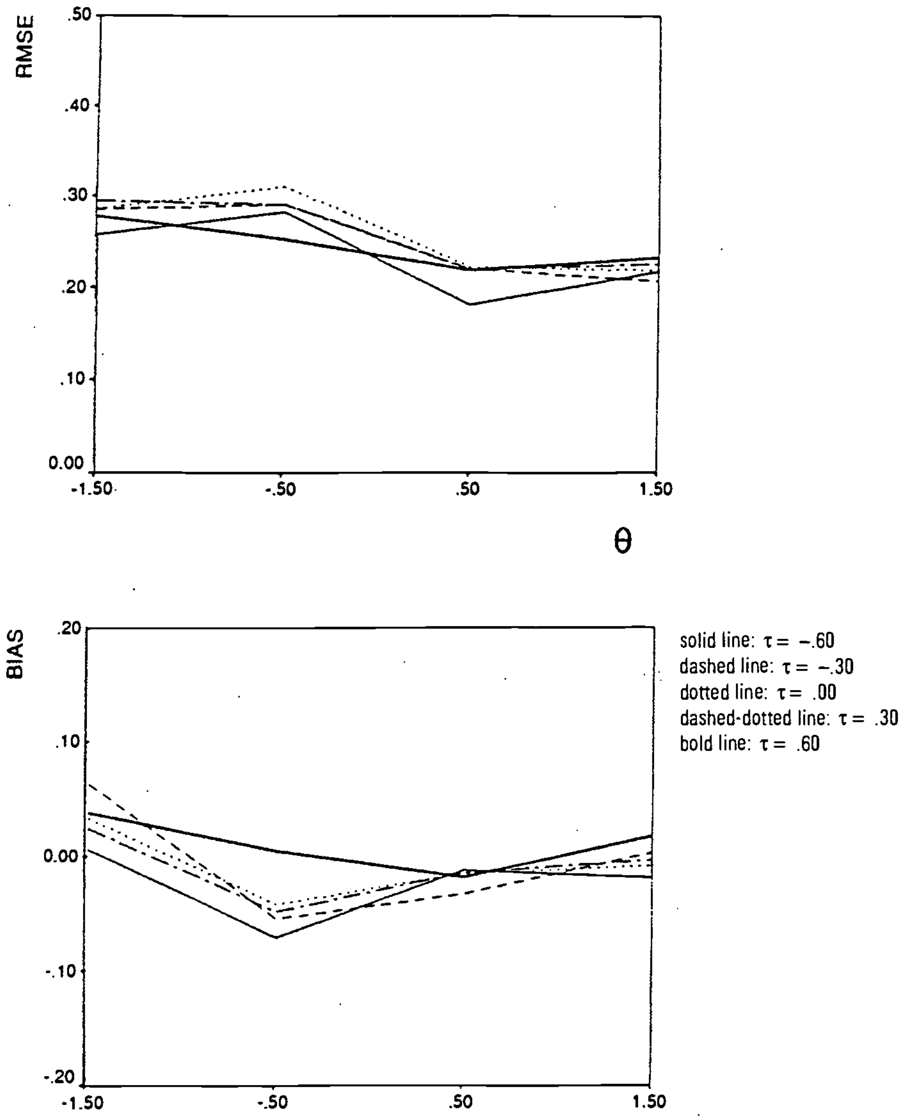

FIGURE 4. *Estimated RMSE and bias functions of the ability estimator after 15 items*

Conclusion

Differential speededness among test takers is a problem in educational measurement. In standardized testing, solutions to the problem include relaxing the time limit or reducing the number of items administered. However, these solutions either result in greater administration costs or reduced measurement precision for all test takers. The solution proposed in this paper is to make the test adaptive and constrain item selection according to the time needed by the test takers. Response-time constraints on item selection are built into an adaptive procedure that maximizes the statistical information in the test about the test taker's ability. The model for the response-time distributions used is the lognormal. The data in the empirical example showed a satisfactory fit to the model. Also, for a variety of θ and τ values, the bias and the RMSE of the ability estimator in this example did not show any anomalies due to the presence of the response-time constraint in the item-selection procedure.

References

Gelman, A., Carlin, J. B., Stern, H., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

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

Luce, D. R. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187-208). New York: Springer-Verlag.

Schnipke, D. L., Scrams, D. J. (1997). *Representing response-time information in item banks* (LSAC Computerized Testing Report No.97-09). Newtown, PA: Law School Admission Council.

Scrams, D. J., & Schnipke, D. L. (in preparation). *A lognormal response time model*. Princeton, NJ: Educational Testing Service.

Swanson, D. B., Featherman, C., Case, S. M., Luecht, R. M., & Nungester, R.J. (1997, April). *Relation of response latency to test design, examinee proficiency, and item difficulty in computer-based test administration*. Paper presented at the annual eeting of the National Council on Measurement in Education, Chicago.

Thissen, D. (1983). Timed testing: An approach using item response testing. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

Timminga, E., van der Linden, M. J., & Schweizer, D. A. (1996) *ConTEST, test assembly system* (sect. 6.6) [Computer software and manual]. St. Paul, MN: Assessement Systems Corporation.

Townsend, J. T., & Ashby, G. F. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.

van der Linden, W. J. (1998a). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201–216.

van der Linden, W. J. (Ed.). (1998b). Optimal assembly of educational and psychological tests, with a bibliography [Special issue]. *Applied Psychological Measurement, 22*(3).

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

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169-186). New York: Springer-Verlag.