Speededness and Adaptive Testing

Wim J. van der Linden
Xinhui Xiong
CTB/McGraw-Hill

Two simple constraints on the item parameters in a response–time model are proposed to control the speededness of an adaptive test. As the constraints are additive, they can easily be included in the constraint set for a shadow-test approach (STA) to adaptive testing. Alternatively, a simple heuristic is presented to control speededness in plain adaptive testing without any constraints. Both types of control are easy to implement and do not require any other real-time parameter estimation during the test than the regular update of the test taker’s ability estimate. Evaluation of the two approaches using simulated adaptive testing showed that the STA was especially effective. It guaranteed testing times that differed less than 10 seconds from a reference test across a variety of conditions.

Keywords: adaptive testing; Luecht heuristic; lognormal response–time model; mixed integer programming; shadow-test approach; test speededness

Introduction

Adaptive tests are typically organized as fixed-length tests with test takers signing up for fixed time slots. However, the items in the pool may show considerable variation in the time they require to process all their information and find a solution. As different test takers get different selections of items from the pool, the time they need to reply to their selection may also vary considerably, with serious differential speededness of the test as a result. Because the time intensity of the items typically correlates with their difficulty, and the difficulty of the selected items correlates highly with the ability of the test takers, usually the more able test takers experience much greater time pressure during adaptive testing than their less able counterparts. Empirical evidence demonstrating this counterintuitive result has been presented, for instance, by Bridgeman and Cline (2003) and van der Linden (2009a).

A method for the control of speededness in adaptive testing was presented by van der Linden (2009a). The method was based on the same response–time (RT) model as in the current study (which will be reviewed below). It consisted of the following components: (a) use of the RTs to permanently update an estimate of
the test taker’s speed during the adaptive test; (b) use of the speed estimate along with the estimates of the time intensities for the remaining items in the pool to predict the time necessary for each of them; and (c) selection of the items in the adaptive test subject to a constraint on their predicted times that guarantees completion of the test within the remaining time. Bayesian methodology was used for updating both the estimate of the speed parameter and the predicted times for the test taker on the remaining items in the pool in the method.

Two major differences exist between this earlier study and the method presented in this article. First, the control in the earlier study was one sided in that it only imposed the time limit as an upper bound on the completion time for the test but did not prevent test takers from finishing too early. In fact, because of the low time intensity of the majority of items selected by the adaptive algorithm, the evaluation of the method showed many test takers who actually did finish much earlier (van der Linden, 2009a, figure 1). The new method is designed to offer two-sided control; it prevents candidates from running out of time but at the same time selects combinations of items that imply full use of the time available for the test. We believe this additional feature is important because it allows for more effective scheduling of test takers across the fixed time slots typically used in adaptive tests (no idle testing stations or unrest created by test takers leaving early).

Second, the new method does not require any real-time estimation of speed parameters for the test takers nor any updating of the predicted distributions of the time the test takers would spend on each of the remaining items in the pool. It thus entirely avoids the computational burden involved in the use of a Bayesian predictive methodology during adaptive testing and produces results not liable to any errors in the estimates of these parameters and distributions. Instead, the new method imposes two simple constraints on the RT parameters of the items that are selected. Consequently, the method is effective directly from the beginning of the test; we do not have to wait for stabilization of any parameter estimates more toward the end of it.

Another key element of the new method is its direct focus on the probability of a test taker running out of time on the adaptive test. Following van der Linden (2011a), the probability is defined as follows: Let $T_i$ denote the random response time by a test taker on the items $i = 1, \ldots, n$ in the test. The total time for the test taker is $T_{tot} = \sum_{i=1}^{n} T_i$. Observe that $T_{tot}$ is random as well; its cumulative distribution function is denoted as $F_{T_{tot}}(t)$. Hence, the probability (or risk) of the test taker running out of time is

$$\pi = \Pr \left\{ \sum_{i=1}^{n} T_i > t_{lim} | \tau, \alpha, \beta \right\}$$

$$= 1 - F_{T_{tot}}(t_{lim} | \tau, \alpha, \beta),$$

where $\tau$ is the test taker’s speed, $\beta = (\beta_1, \ldots, \beta_n)$ is the vector of parameters for the time intensities of the items in the test, and $\alpha = (\alpha_1, \ldots, \alpha_n)$ is the vector of
discrimination parameters for the items in the response–time model in Equation 2 given below. Notice that the lower the risk, the less stringent the time limit will be. Also, as follows from the RT model below, the risk will be lower for test takers who operate faster, but it will be higher for selections of items that are more time intensive. Each of these relationships makes intuitive sense.

For a given test taker, the probability of running out of time in Equation 1 can be controlled by constraining the vectors $\alpha$ and $\beta$ for the selection of items in the adaptive test. The same principle was applied in van der Linden (2011b) to assemble fixed test forms to be equally speeded as a reference form, fixed forms with the same degree of speededness but some of the other test specifications changed, as well as the problem of adjusting the degree of speededness of a testing program with a fixed time slot to a newly selected level. As demonstrated below, application of the principle to adaptive testing becomes extremely straightforward for a direct approximation to the distribution function in Equation 1, which leads to two constraints each to be imposed on a simple sum of item

![Graph of total-time distributions](image-url)
parameters. The only thing we need to do during adaptive testing is updating the sums for the items selected in the test.

**Total-Time Distribution**

The response time for a test taker with speed \( \tau \in (-\infty, \infty) \) on item \( i \) is modeled using the lognormal density

\[
f_{T_i}(t; \tau, \alpha_i, \beta_i) = \frac{\alpha_i}{t_i \sqrt{2\pi}} \exp\left\{-\frac{1}{2} \left[\alpha_i (\ln t_i - (\beta_i - \tau))^2\right]\right\},
\]

with parameters \( \beta_i \in (-\infty, \infty) \) for the time intensity or amount of labor required by item \( i \) and \( \alpha_i \in (0, \infty) \) for its discriminating power. The model follows directly from a fundamental equation for response-time modeling derived from the definition of speed on a test as the rate of change in the amount of labor performed on the items with respect to time (van der Linden, 2009b). Bayesian estimation of the parameters with Gibbs sampling from their posterior distributions is described in van der Linden (2006). Alternatively, the item parameters can be estimated using confirmatory factor analysis (Finger & Chuah, 2009). For various methods of checking the fit of the model (Bayesian predictive checks; the deviance information criterion [DIC]; Bayes factors; Bayesian residuals), see Klein Entink, Fox, and van der Linden (2009), and van der Linden (2006, 2007). Throughout this article, we assume that the fit of the model in Equation 2 has been checked carefully with satisfactory results and its parameters have been previously estimated with enough precision to treat them as known.

The density of the distribution of the total time for a test taker with speed \( \tau \) on two items follows from Equation 2 as

\[
f_{T_{\text{tot}}}(t; \tau, \alpha, \beta) = \int f_{T_1}(z; \tau, \alpha_1, \beta_1) f_{T_2}(t - z; \tau, \alpha_2, \beta_2) dz,
\]

an expression known as the convolution integral of the separate densities for the two items. Repeated application of the integral, each time adding an extra item to the test, gives the distribution of \( T_{\text{tot}} \) as the \( n \)-fold convolution of the densities for the item distributions. The convolution does not have any known closed form, and its integrals are numerically intractable, even for the case of a test consisting of only a few test items. However, as shown in van der Linden (2011a), it can be approximated excellently by a standard lognormal density using Fenton’s (1960) method of matching the first two cumulants. We only summarize the approximation here; for its derivation as well as examples illustrating its extremely accurate fit, see van der Linden (2011a).

We reparameterize the RT model in Equation 2 and define new item parameters

\[
q_i \equiv \exp(\beta_i + \alpha_i^{-2}/2),
\]

and
As a result, although the distribution of $T_{tot}$ has an unknown shape, its mean and variance can be shown to be equal to

$$\text{Mean}(T_{tot}) = \exp(-\tau) \sum_{i=1}^{n} q_i$$

(6)

and

$$\text{Var}(T_{tot}) = \exp(-2\tau) \sum_{i=1}^{n} r_i.$$  

(7)

(van der Linden, 2011a, equations 20–21).

The standard family of lognormal distributions family has density

$$f(x; \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left( \frac{\ln{x - \mu}}{\sigma} \right)^2 \right\}, \mu, \sigma > 0.$$  

(8)

with parameters $\mu$ and $\sigma^2$. The member we need to approximate the density of $T_{tot}$ has parameters $\mu$ and $\sigma^2$ that can be shown to follow from the new parameters in Equations 4 and 5 as

$$\mu = -\tau + \ln\left( \sum_{i=1}^{n} q_i \right) - \ln\left( \frac{\sum_{i=1}^{n} r_i}{[\sum_{i=1}^{n} q_i]^2} + 1 \right)/2,$$

(9)

and

$$\sigma^2 = \ln\left( \frac{\sum_{i=1}^{n} r_i}{[\sum_{i=1}^{n} q_i]^2} + 1 \right).$$

(10)

Observe that $\mu$ and $\sigma^2$ are not the mean and variance of the (otherwise unknown) distribution of $T_{tot}$ in Equations 6 and 7. Rather, as revealed by the parameter structure of Equation 8, they are the equivalent of the mean and variance of the normal distribution of an arbitrary variate $\ln X$. However, substitution of Equations 9 and 10 into 8 gives us its member with the same mean and variance as the distribution of $T_{tot}$ in Equations 6 and 7.

Controlling Speededness in Adaptive Testing

A surprisingly useful feature of Equations 6 and 7 is their factorization into two parts depending exclusively on the test taker and the item parameters, respectively. The factorization guarantees that, whatever the speed $\tau$ of the test taker, the mean and variance of his or her total-time distribution on any two test forms with matching sums of these item parameters are identical. In fact,
Equations 8 through 10 show that the same principle holds for the entire shape of the approximating distribution; for any $t$, identical sums of $q_i$ and $r_i$ parameters guarantee identical distributions of $\ln T_{tot}$, and thus of $T_{tot}$.

Another useful feature of each expression in Equations 7 and 8 is that they are based only on sums of the $q_i$ and $r_i$ parameters. The additivity is particularly convenient in adaptive testing. It allows us to immediately evaluate the impact of any extra item on the properties of the total-time distribution; the only thing we need to do is add the $q_i$ and $r_i$ parameters for the candidate item to the two sums.

In order to control the speededness of an adaptive test, our basic idea is to identify a reference form with a proven, excellent level of test speededness, and build the adaptive test to produce the same level. For instance, the reference form could be a fixed form used before a program became adaptive or the set of items in a previous adaptive test. It is essential that the actual total times on it have been checked for test takers working at a realistic range of speed and that they have been found to be all right, given the current time limit. In addition, an evaluation of the subjective time pressure experienced by the test takers during the test may play a role. Observe that we only have to check these data for a realistic range of $t$ values. As for any two matching sums of $q_i$ and $r_i$ parameters, the conditional distributions of $T_{tot}$ given $t$ are automatically identical for each value of $t$, differences between the distributions of $t$ in the populations of test takers for the reference and new forms do not matter.

In fact, as the two sums of the $q_i$ and $r_i$ parameters are the only things we need from the reference test, we could just as well select a priori values for them that are plausible for a test of the intended length, study the degree of speededness in Equation 1 associated with them (for instance, using the risk curves in van der Linden, 2011a, figure 2), modify their values if necessary, and use the final result to set the target values in Equations 11 and 12 below.

As identical total-time distributions on the adaptive and reference test for each test taker implies identical risks of running out of time, we control the speededness during the adaptive test administrations in the strongest possible sense. For instance, for any given population of test takers, the method of control guarantees that the same number of test takers will experience exactly the same levels of time pressure on both tests. Observe that we are able to guarantee this without actually having to specify any explicit limit on the risk in Equation 1. We thus entirely avoid the problem of having to specify a minimally acceptable level of speed required to calculate such a limit.

Let $j = 1, \ldots, n$ denote the items in the reference form. We use the item parameters in Equations 4 and 5 for the reference form to define target values

$$T_q = \sum_{j=1}^{n} q_j,$$  \hspace{1cm} (11)

and
Each adaptive test is built to meet these two values. In order to achieve this goal, two different implementations of the idea are suggested. The first is a shadow-test method in which constraints based on the target values are inserted in the test-assembly model for each shadow test. A key advantage of this approach is that the constraints can be easily combined with whatever other constraints are required to meet the full set of content, statistical, and practical specifications for the test. The second method is a simple heuristic. It can be used for a plain adaptive test without any other constraints than the target values in Equations 11 and 12.

$$T_r \equiv \sum_{j=1}^{n} r_j,$$

(12)

FIGURE 2. a. Average difference between the time limits for the reference and adaptive tests for the conditions with no control, shadow-test approach (STA), and two versions of the heuristic in seconds ($\pi = .05$).

Downloaded from http://jebs.aera.net by Wim van der Linden on July 8, 2013
Implementations

Shadow-Test Approach (STA)

In STA, the selection of each item is preceded by the assembly of a shadow test from the item pool. Shadow tests are defined as fixed-form tests of the same length as the adaptive tests that (a) are maximal informative at the test takers current estimate of $\theta$, (b) meet all constraints to be imposed on the test, and (c) contain all items already administered to the test taker (van der Linden, 2005, chap. 9). Basically, each next shadow test involves a reassembly of the remaining portion of the adaptive test, such that it still meets all constraints but now is maximally informative at the new $\theta$ estimate. The next item to be
administered is the most informative item in the shadow test not yet seen by the test taker. All other free items are returned to the item pool. Because each next shadow tests meets all constraints and is maximally informative, the same automatically holds for the adaptive test.

In principle, shadow tests can be assembled using any method of constrained test assembly that is fast enough for use in real time. In the example later in this article, we used mixed integer programming (MIP) in combination with a fast commercial solver (IBM ILOG OPL version 6.3; International Business Machines Corporation, 2009) to identify the optimal shadow tests. For an application in R based on a free solver, see Diao and van der Linden (2011). The first step in this approach is the definition of 0-1 decision variables $x_i, i = 1, \ldots, I$, for

FIGURE 2. c. Average difference between the time limits for the reference and adaptive tests for the conditions with no control, STA, and two versions of the heuristic in seconds ($\pi = .15$).

No Control

STA

Heuristic (Standardized Weights)

Heuristic (Weight 50% Greater)
the items in the pool, which take the value \( x_i = 1 \) if item \( i \) is selected for the new shadow test and \( x_i = 0 \) if it is not. These variables are used to model the objective function for the test as well as all necessary constraints implied by the test specifications. For a more comprehensive review of the possibilities of MIP modeling for automated test assembly, refer to van der Linden (2005).

Suppose the shadow test prior to the selection of the \( k \)th item in the adaptive test needs to be reassembled. Let \( S_{k-1} \) denote the indices of the \( k-1 \) items already administered and \( I_i(\hat{\theta}_{k-1}) \) the value of the information function for item \( i \) at the current estimate \( \hat{\theta}_{k-1} \). The core of the model for the reassembly of the shadow test is

\[
\text{maximize } \sum_{i=1}^{I} I_i(\hat{\theta}_{k-1})x_i \quad \text{(maximum information at } \hat{\theta}_{k-1}) \tag{13}
\]

subject to

\[
\sum_{i=1}^{I} q_i x_i \leq T_q + \delta_q, \quad \text{(upper limit on sum of } q_i), \tag{14}
\]

\[
\sum_{i=1}^{I} q_i x_i \geq T_q - \delta_q, \quad \text{(lower limit on sum of } q_i), \tag{15}
\]

\[
\sum_{i=1}^{I} r_i x_i \leq T_r + \delta_r, \quad \text{(upper limit on sum of } r_i), \tag{16}
\]

\[
\sum_{i=1}^{I} r_i x_i \geq T_r - \delta_r, \quad \text{(lower limit on sum of } r_i), \tag{17}
\]

\[
\sum_{i \in S_{k-1}} x_i = k - 1, \quad \text{(items already administered)}, \tag{18}
\]

\[
x_i \in \{0, 1\}, i = 1, \ldots, I. \quad \text{(range of variables).} \tag{19}
\]

where \( \hat{\theta}_{k-1} \) is the estimate of \( \theta \) after the previous \( k-1 \) items and \( \delta_q \) and \( \delta_r \) are tolerance parameters used to set small intervals about the target values \( T_q \) and \( T_r \), respectively.

In a real-world application, the model may have to be extended with several other constraints, for instance, on the content distribution of the items in the test, to guarantee a desired answer-key distribution, prevent the administration of “enemy items,” constrain the range of readability indices, and so on. The solution to the model found by the solver is the string of 0s and 1s for the variables \( x_i, i = 1, \ldots, I \), that identifies the items for the shadow test that meet all constraints and provide maximum information at \( \hat{\theta}_{k-1} \).
Heuristic

The heuristic we propose is analogous to a suggestion for automated test assembly by Luecht (1998). However, instead of the original application to a target for the test information function as the criterion for item selection, the current application is with respect to a combination of the targets $T_q$ and $T_r$ and the goal of maximum formation at $\theta$.

The method consists of the calculation of the differences between the targets in Equations 11 and 12 and the sums of the parameters for the $k - 1$ items already administered to the test taker, division of the two differences in equal portions for the $n - k + 1$ remaining item slots in the adaptive test, and use of the result to select the $k$th item. More formally, we define

\[ T_{qk} = \left( T_q - \sum_{i \in S_{k-1}} q_i \right) / (n - k + 1), \]  
\[ T_{rk} = \left( T_r - \sum_{i \in S_{k-1}} r_i \right) / (n - k + 1), \]  

as the target values for the selection of the $k$th item. The $k$th item is selected to have both maximum information at $\theta_{k-1}$ and $q_i$ and $r_i$ parameters as close as possible to $T_{qk}$ and $T_{rk}$, respectively. The three objectives are combined in the following criterion for item selection:

\[ i_k = \arg \max_j \left\{ I_i(\theta_{k-1}) - w_q |q_i - T_{qk}| - w_r |r_i - T_{rk}|; j \in R_k \right\}, \]  

where $R_k = \{ 1, \ldots, I \} \setminus S_{k-1}$ is the set of indices of the remaining items in the pool after $k - 1$ items have been administered. Observe that the algorithm is “forward looking” in that it uses the current gap between the actual sums of the $q_i$ and $r_i$ parameters and the targets in Equations 11 and 12 to set a criterion for the next item. In doing so, it prevents from a scenario in which all best items are immediately selected, leaving us with the lesser items to finalize the test. In addition, if we are occasionally forced to select an item with a contribution too low or high relative to Equations 20 and 21, the recalculation of these values at the next step automatically compensates for the fact.

The weights $w_q$ and $w_r$ in Equation 22 can be used to prioritize the individual objectives. This should be done after a correction for the scale differences between them. In the simulation studies below, we used the average values of $I_i(\theta_{k-1})$ and the $q_i$ and $r_i$ parameters for the remaining set of items $R_k$ in the pool to correct for scale differences and opted for equal priorities for all three objectives. As the numbers of items cancel, the choice amounts to

\[ w_{qk} = \frac{\sum_{i \in R_k} I_i(\theta_{k-1})}{\sum_{i \in R_k} q_i}, \]  

428
and

\[
W_{jk} = \frac{\sum_{i \in R_k} I_i(\hat{\theta}_{k-1})}{\sum_{i \in R_k} r_i}.
\]

(24)

**Evaluation Study**

Simulation studies were conducted to evaluate the effectiveness of the STA and heuristic in eliminating differences in speededness between test takers under a variety of conditions. More particularly, we looked at the extent to which the two methods approximated the desired degree of speededness for the adaptive test for test takers with different abilities operating at different speeds. In addition, we assessed the possible price to be paid for the control of the speededness in the form of loss of precision of ability estimation due to the extra constraints or objectives enforced on the item selection.

**Data**

Adaptive test administrations were simulated from a pool consisting of a set of 185 items sampled from a retired pool from a large-scale testing program. The \(a_i\) for \(\beta_i\) parameters for items in the original pool were estimated from a data set with the RTs for more than 10,000 test takers, using a Bayesian method with Gibbs sampling from the posterior distributions of the parameters (for a description of the method, see van der Linden, 2006). We only selected items that showed satisfactory fit to the model. The ranges of the estimates were [1.32, 2.46] for the \(a_i\) and [3.01, 5.08] for the \(\beta_i\) parameters. The mean estimate of the speed parameters \(\tau\) for the test takers in the data set was centered on \(\mu_{\tau} = 0\). The estimates had an empirical standard deviation of approximately \(\sigma_{\tau} = .35\). The estimated parameter values were used as the true values in the simulation study. The same holds for the item parameters in the three-parameter logistic (3PL) response model; the true values used for these parameters were the estimates that had been used operationally in the testing program. Table 1 gives some descriptive statistics of the item parameter distribution in the item pool. For later reference, observe the positive correlation between the \(a_i\) and \(b_i\) parameters and the \(b_i\) and \(\beta_i\) parameters. The former implies an item pool that is generally less informative at the lower end of the scale. The latter is a potential source of differential speededness; without any control, it would have the item-selection algorithm select the more time-intensive items for the more able test takers.

Adaptive tests with a fixed length of 20 items for test takers with a true ability parameter equal to \(\theta = -2, -1, 0, 1,\) or 2 were simulated. For each \(\theta\), we replicated 500 simulations. Item selection was according to the
maximum-information criterion. For ability estimation, we used expected a posteriori (EAP) estimation with a uniform prior over \([-4, 4]\). The first item was chosen to be optimal at \(\theta_0 = 0\).

Four different conditions were simulated:

1. No control of speededness;
2. Control of speededness using the STA in Equations 13–19;
3. Control of speededness using the item-selection heuristic in Equation 22 with the standardized weights in Equations 23 and 24.
4. Control of speededness as in Condition 3, but now with weights that were 50\% greater than the standardized weights in Equations 23 and 24.

The fourth condition was included to explore the effects of relatively greater weights for the last two objectives in Equation 22. In principle, increasing these weights means better control of speededness, but possibly at the price of a less informative adaptive test. The STA was implemented using tolerances \(\delta_q\) and \(\delta_r\) in Equations 14 to 17 set equal to 0.5\% of the target values.

The reference test consisted of 20 items randomly sampled from the pool. The target values for the reference test in Equations 11 and 12 were \(T_q = 1,601\) and \(T_r = 49,592\). Figure 1 shows the total-time distributions in Equations 8–10 for this single reference test for test takers working at speed \(\tau = -.70, .35, 0, .35,\) and \(.70\). The speed levels were chosen to cover the approximate range of \(\mu_t \pm 2\sigma_t\) for the test takers in the original data set. Clearly, variation of the speed parameter is consequential. The two most extreme distributions, which were four standard deviations apart (\(\tau = -.70\) and \(\tau = .70\)), had locations at approximately 800 and 3,200 s.

For each of the five \(\tau\) levels, we calculated the time limit \(t_{lim}\) required for the reference test to realize the risk levels of \(\pi = .05, .10,\) and \(.15\) using

\[
t_{lim} = F_{\tau_{lim}}^{-1}(1 - \pi | \alpha, \beta),
\]

<table>
<thead>
<tr>
<th>Correlation</th>
<th>(a_i)</th>
<th>(b_i)</th>
<th>(c_i)</th>
<th>(\alpha_i)</th>
<th>(\beta_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_i)</td>
<td>0.88</td>
<td>1.21</td>
<td>1.39</td>
<td>1.50</td>
<td>1.74</td>
</tr>
<tr>
<td>(b_i)</td>
<td>-2.81</td>
<td>-1.07</td>
<td>-0.10</td>
<td>-0.20</td>
<td>0.68</td>
</tr>
<tr>
<td>(c_i)</td>
<td>0.02</td>
<td>0.13</td>
<td>0.17</td>
<td>0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>(\alpha_i)</td>
<td>1.32</td>
<td>1.73</td>
<td>1.89</td>
<td>1.89</td>
<td>2.03</td>
</tr>
<tr>
<td>(\beta_i)</td>
<td>3.01</td>
<td>3.79</td>
<td>4.10</td>
<td>4.08</td>
<td>4.33</td>
</tr>
</tbody>
</table>
which is the value of the quantile function for the total-time distribution (inverse of the distribution function in Equation 1) at $1/C_0$. The range of risk levels was chosen to represent a testing program with an intentionally low degree of speededness. The results in Table 2 show time limits that varied widely across the chosen combinations of speed and risk. The limit required for the slowest speed in combination with the lowest risk level was approximately 4 times higher than the one for the highest speed in combination with the highest risk level ($t_{\text{lim}} = 66'51''$ vs. $15'09''$).

As the methods of control of differential speededness did not assume estimation of any speed parameter during testing, it was not necessary to manipulate this parameter in our study. The only thing we had to do to evaluate the two methods was to record the $\alpha_i$ and $\beta_i$ parameters in the RT model for the items administered to the test takers. The actual level of speededness was evaluated afterward, comparing the time limits for the reference test in Table 2 with the time limits that would have been required for the actual items in the adaptive tests to create the same risk at the same levels of speed. This choice of evaluation criterion allowed us to evaluate speededness in a directly interpretable metric, namely as a difference between actual and required testing time.

The estimation errors $\hat{\theta}_k - \theta$ recorded during the simulation of the adaptive tests were used to calculate the bias and MSE functions for the final ability estimates as

$$E[(\hat{\theta}_k - \theta)|\theta],$$

and

$$E[(\hat{\theta}_k - \theta)^2|\theta].$$

### Results

Figure 2 shows the average difference between the actual and required time limits in seconds as a function of the $\tau$ values for a risk of $\pi = .05$ (panel a), $\pi = .10$
(panel b), and \( \pi = .15 \) (panel c) for the four conditions. The large differences for the condition of no control show the seriousness of the problem of differential speededness in adaptive testing. The problem was especially dramatic for test takers with high ability who worked slowly; on average, they would have needed at least 15 more minutes to finish in time. On the other hand, low-ability test takers at the same level of speed had an average of approximately 25 min left when they finished the test. This differential effect is entirely due to the combination of the already observed positive correlation between the difficulties and time intensities of the items (\( r = .63 \)) and the adaptive algorithm selecting items with difficulties matching the test takers’ abilities. The same combination explains the general order of the curves for the \( \theta \) levels in each of the plots in Figure 2.

The shadow-test method did a superior job controlling the speededness of the adaptive test. No matter the level of risk, speed, or ability, the presence of the constraints in Equations 14 to 17 guaranteed a maximum absolute difference between the time limit for the reference test and the limits required for the adaptive test smaller than 10 seconds. The two versions of the heuristic method had more difficulty controlling the degree of differential speededness, but all plots show results that tend to be much closer to those for the shadow-test method than for the condition without control. Generally, as expected, the results for the higher weights were somewhat better, especially for the low ability group.

In principle, small differences between average levels of speededness do not imply anything at the level of the individual test takers. The standard deviations of the time limits required by the adaptive tests in Figure 3, however, confirm the differences in control between the four conditions. The variation between the limits for the condition without any control was huge, whereas the shadow-test method showed negligible variation (all standard deviations smaller than 10 seconds). Again, the results for the heuristic were closer to those for the shadow-test method than for the condition without any control. For all ability groups, the two versions of the heuristic yielded approximately equal standard deviations, with the exception of some improvement for the low-ability group for the version with the higher weights. Thus, the better results for the average differences in time limit by the latter were accompanied by a somewhat smaller variability across the test takers. The same general pattern of results for the four methods is shown by the 1st and 99th percentiles of their distributions of the time limit for the reference test minus the limits for the adaptive tests in Table 3. Note that the case for \( \tau = .70 \) in this table is the one with the largest standard deviations in Figure 3. The results for \( \pi = .05, .10, \) and .15 were entirely comparable.

Generally, the control of speededness resulted in negligible loss of statistical quality of the ability estimates. The plots in Figure 4 show the estimated bias and MSE functions for the two control methods that are essentially the same as for adaptive testing without any control. The only differences are a
positive bias for the shadow-test method and a somewhat larger MSE for the two versions of the heuristic method at $\theta = -2$. As the precision of estimation at all other $\theta$ values was virtually identical, we expect these incidental exceptions to be due to less information in the items at the lower end of the scale (role of the guessing parameter and positive correlation between the $a_i$ and $b_i$ parameters, $r = .60$).

**Concluding Remarks**

The empirical study showed how serious the effects of differences in item selection on the testing time can be if we do not control for differential speededness in adaptive testing. It also showed that the effects can be effectively
removed by imposing two simple constraints on the RT parameters of the items during item selection.

Another novelty of the method is the absence of any estimation of the test taker’s actual speed during the test. Besides, it does not require any projection of the time needed by the test taker for the remaining portion of the test. And the idea of matching the time characteristics of the adaptive tests with a reference test of proven quality is also practically convenient: It prevents the setting of a minimally acceptable level of speed for the test takers required for direct control of the risk in Equation 1. In spite of control of differential speededness up to a few seconds by the STA method, there appears to be hardly any price in the form of deteriorated ability estimation.

It is possible to compare the results for the current STA method with those for the earlier method in van der Linden (2009a). Both studies had
FIGURE 3. c. Standard deviation of the required time limits for the adaptive test for the conditions with no control, STA, and two versions of the heuristic (π = .15).

TABLE 3
First and 99th Percentiles of Distributions of the Time Limits for the Reference Test Minus the Limits for the Adaptive Tests in Seconds (τ = −.70; π = .10)

<table>
<thead>
<tr>
<th>Percentile</th>
<th>1st</th>
<th>99th</th>
<th>1st</th>
<th>99th</th>
<th>1st</th>
<th>99th</th>
<th>1st</th>
<th>99th</th>
<th>1st</th>
<th>99th</th>
<th>1st</th>
<th>99th</th>
</tr>
</thead>
<tbody>
<tr>
<td>No control</td>
<td>1,609</td>
<td>884</td>
<td>1,353</td>
<td>442</td>
<td>893</td>
<td>−130</td>
<td>231</td>
<td>−483</td>
<td>−332</td>
<td>−1,179</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STA</td>
<td>17</td>
<td>1</td>
<td>17</td>
<td>−15</td>
<td>17</td>
<td>−15</td>
<td>15</td>
<td>−16</td>
<td>12</td>
<td>−16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heuristica</td>
<td>1,057</td>
<td>−3</td>
<td>381</td>
<td>−24</td>
<td>202</td>
<td>−232</td>
<td>−46</td>
<td>−374</td>
<td>3</td>
<td>−412</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heuristicb</td>
<td>721</td>
<td>−18</td>
<td>276</td>
<td>−70</td>
<td>75</td>
<td>−206</td>
<td>−76</td>
<td>−375</td>
<td>3</td>
<td>−306</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aStandardized weights.
bWeights 50% greater.
an item pool for the adaptive test sampled from the same inventory of items and an identical setup of the simulated adaptive tests. The only differences existed in a much smaller item pool for the current study (185 vs. 350 items) in combination with a longer test (20 vs. 15 items). The former favors the earlier study; the latter our current study. The relevant comparison is between the average differences between the actual and required time limits in our current study (Figures 2 and 3) and the average time spent on the adaptive test in the earlier study (van der Linden, 2009a, figure 1; this figure also shows the time limits that were simulated). While the current study demonstrated control up to differences smaller than 10 s for each of the ability groups working at any of the levels of speed, the earlier study yielded much greater variation. In fact, only the differences for the test takers with ability $\theta = 2$ working at the slowest simulated speed in the earlier study came close to our current results; for all other conditions, much larger differences were obtained (running up to some 1,700–1,800 seconds for the group with $\theta = -2$ working at the highest simulated speed). The explanation is the one-sided nature of the earlier method. As already indicated, it was designed only to guarantee completion of the test within the available time, which it did remarkably well. In spite of the much more homogenous, two-sided control by the new method, a comparison between the bias and MSE plots in Figure 4 with those in the earlier study (van der Linden, 2009, figures 5–6) show results with exactly the same pattern: hardly any bias or MSE for the simulated ability levels $\theta = -1, 0, 1$ and similar bias and MSE at $\theta = -2$ and 2.

Finally, observe again that the two constraints are only on two different sums of item parameters. It is thus not necessary to match these parameters with the ones for the reference test on an item-by-item basis. The leeway

FIGURE 4. Estimated bias and MSE as a function of $\theta$ for the conditions with no control, shadow-test approach (STA), and two versions of the heuristic.
generated by this relaxation explains why there was no significant loss of statistical precision in the ability estimates for the two methods of speededness control.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References


Authors

WIM J. VAN DER LINDEN is Chief Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; wim_vanderlinden@ctb.com. His primary research interests include test theory, applied statistics, and research methods.

XINHUI XIONG is Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; xinhui_xiong@ctb.com. Her primary research interests include computerized adaptive testing, equating and linking, multilevel IRT and structure equation modeling.

Manuscript received November 28, 2011
First revision received April 2, 2012
Second revision received July 5, 2012
Accepted July 31, 2012