Predictive Control of Speededness in Adaptive Testing

Wim J. van der Linden, University of Twente

An adaptive testing method is presented that controls the speededness of a test using predictions of the test takers’ response times on the candidate items in the pool. Two different types of predictions are investigated: posterior predictions given the actual response times on the items already administered and posterior predictions that use the responses on these items as an additional source of information. In a simulation study with an adaptive test modeled after a test from the Armed Services Vocational Aptitude Battery, the effectiveness of the methods in removing differential speededness from the test was evaluated. Index terms: adaptive testing, Bayesian prediction, collateral information, differential speededness, response times

Test items differ in the amount of time they require to produce a response to them. These differences are due, for instance, to the amount of information decoding and processing they involve, the complexity of their problem-solving process, and the numbers of concepts or relations that have to be retrieved from memory. It is not unusual to find differences between the time intensities of the items in a pool for an adaptive test equal to a factor of five or larger (van der Linden & Guo, 2008).

It is important to distinguish the time intensity of test items from the speed at which the test takers respond to them. Within certain limits, test takers have a choice of speed at which to work. When making such a choice, they have to deal with what is known in reaction-time research in psychology as a speed–accuracy trade-off (e.g., Luce, 1986, sect. 6.5). The trade-off, which is a within-person effect, implies that if a test taker chooses to operate at a higher speed, the accuracy of problem solving will go down.

Accuracy in problem solving is measured by the probability of producing a correct solution. In test theory, the probability of success on a test item is typically modeled as an increasing function of an ability parameter. Thus, one can also refer to the trade-off as a speed–ability trade-off. The trade-off implies that the test taker’s ability level during the test depends on his or her choice of speed. The specific ability level that is the result of this choice will be denoted as the effective ability of the test taker.

Most large-scale adaptive testing programs have a fixed length. This feature allows the program to impose the same content distributions on the test for each test taker and to schedule the administrations using uniform time slots. However, each test taker gets a different selection of items. Because the items differ in time intensity, some of the test takers may therefore have to work under a high time pressure whereas others have ample time to complete the test. The same phenomenon of differential speededness may occur in multistage testing or in other multiple-form testing.
formats such as “linear-on-the-fly testing,” where the computer selects a different linear test for each test taker. But the effects of differential speededness are expected to be more serious in adaptive testing. Typically, more difficult items tend to be more time intensive. In their study, van der Linden, Scrams, and Schnipke (1999) found a correlation between the ability and speed estimates under a regular item response theory (IRT) model and response-time model equal to .65. Adaptation thus has the side effect of more able test takers getting more time-intensive items. For a fixed time limit, more able test takers therefore have a risk of running out of time and are forced to make a choice between increasing their speed and accepting a lower effective ability or maintaining it and guessing on the items that are not reached. In either case, the effect is a systematically lower test score than if less time-intensive items had been administered.

Empirical studies on the effects of differential speededness have been reported in Bejar (1985), Bridgeman and Cline (2004), Evans and Reilly (1972), Kingston and Dorans (1984), and van der Linden, Breithaupt, Chuah, and Zhang (2007). But differences in speededness have an impact beyond test scoring. For instance, as shown in studies by Kingston and Dorans (1984), Oshima (1994), Wollack, Cohen, and Wells (2003), and Yamamoto and Everson (1997), it also biases the results of item parameter estimation and test equating.

More important than the diagnosis of differential speededness are methods to prevent it. One such method was presented in van der Linden et al. (1999), who used the actual response times (RTs) on the items during an adaptive test to predict the RT distributions on the remaining items in the pool and then used this information to constrain the selection of the items to meet the time limit for the test takers. However, the RT model they used for the predictions was a simple analysis of variance model with parameters for main effects for the persons and items. Also, they had no method to estimate the model from incomplete data but substituted sample statistics for parameters while ignoring the fact that the data were collected during operational adaptive testing.

The current research was based on a recent RT model with a more realistic parameterization that has been shown to have good fit to RTs in operational testing (van der Linden, 2006). For this model, a Bayesian method for estimating the model parameters using Markov chain Monte Carlo sampling from their posterior distribution was available. Also, in the current research a new method was used for the prediction of RTs on test items that exploits the responses on the test items as an additional source of information on the speed at which the test taker works. The effectiveness of the earlier and new prediction methods was evaluated in an empirical study using an item pool from the Armed Services Vocational Aptitude Battery (ASVAB).

**Models**

As is customary in adaptive testing, the items in the pool were assumed to have been calibrated with enough precision to treat them as known during the test. The item pool in the empirical example below was calibrated using the three-parameter logistic (3PL) model; that is, the probability of a correct response, $U_j = 1$, was modeled as

$$\Pr(U_j = 1) = p_j(\theta_j) \equiv c_i + (1 - c_i)\Psi(a_i(\theta_j - b_i)),$$

where $\Psi(\cdot)$ is the logistic distribution function; $\theta_j \in [\infty, \infty]$ is the ability of test taker $j$; and $a_i \in (0, \infty], b_i \in [-\infty, \infty]$, and $c_i \in (0, 1)$ are the discrimination, difficulty, and guessing parameters for item $i$, respectively (Birnbaum, 1968).

The RT model is for the time taken to produce a response on a test item, which can be correct or incorrect with probabilities given by the response model in equation (1). (The model thus differs
from others in reaction-time research in psychology that measure the time until a correct response.) Let \( T_{ij} \) denote the RT of test taker \( j \) on item \( i \). The model posits a log-normal distribution for \( T_{ij} \) with a parameter structure that, except for guessing parameter \( c_i \), is entirely analogous to that of the 3PL model in equation (1):

\[
f(t; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left[ \alpha_i (\ln t_{ij} - (\beta_i - \tau_j))^2 \right] \right\}, \tag{2}
\]

where \( \tau_j \in [-\infty, \infty] \) is the speed at which test taker \( j \) operates during the test and \( \alpha_i \in (0, \infty] \) and \( \beta_i \in (-\infty, \infty] \) are the discrimination and time-intensity parameters of item \( i \). Because RTs have a natural zero, the model does not need any analogue of the guessing parameter in equation (1).

The assumption of a log-normal distribution of \( T_{ij} \) is identical to that of a normal distribution of \( \ln T_{ij} \); that is, \( \ln T_{ij} \sim N(\beta_i - \tau_j, \alpha_i^{-2}) \). Because of their natural zero, RT distributions on the natural time scale are typically skewed to the right. The logarithmic transformation removes the skewness from these distributions. The derivations of the prediction equations below capitalize on the property of a normal distribution of \( \ln T_{ij} \).

The property also helps to interpret the item parameters. The mean of \( \ln T_{ij} \) is equal to \( \beta_i - \tau_j \). Thus, the faster the test taker works, the lower this mean. Likewise, the higher the ability to identify an item, the higher the mean. The standard deviation of \( \ln T_{ij} \) is equal to \( \alpha_i^{-1} \). Thus, the larger \( \alpha_i \), the less \( \ln T_{ij} \) is dispersed around its mean \( \tau_j - \beta_i \), and the better the discrimination between the RT distributions of test takers working at a speed just above and below \( \beta_i \).

Although the distribution in equation (2) has an upper tail that tapers off to zero, it is not bounded. Researchers should be aware of this when using the model in a simulation study. For example, the study described later in this article replicated extremely large numbers of draws from the RT distributions for simulated test takers. Because of the nonzero probability of an outlier, a few of them might have been drawn. Therefore, the distribution was truncated at an arbitrary large value.

A Bayesian method for parameter estimation with Gibbs sampling from the posterior distribution is given in van der Linden (2006). The sampler is easy to implement because its conditional full posterior distributions of the model parameters are standard options in libraries of statistical procedures. It is therefore relatively straightforward to estimate the item parameters of the RT model as part of the regular calibration of the pool. The only thing required is the recording of the RTs when the items are pretested.

In, remainder of this article, the parameters \( \alpha_i \) and \( \beta_i \) of the items in the pool for the adaptive test are assumed to have been estimated with enough precision to treat them as known. In addition, the speed parameters in the population of test takers are assumed to be approximated by a normal distribution,

\[
\tau \sim N(\mu_{\tau}, \sigma_{\tau}^2). \tag{3}
\]

Parameter \( \sigma_{\tau} \) can be estimated easily as part of the regular calibration of the item pool. How to do so is discussed below. Because the zero of \( \tau_j \) and \( \beta_i \) in equation (2) are not yet determined, they are fixed by setting \( \mu_{\tau} = 0 \). It is thus not necessary to estimate \( \mu_{\tau} \).

**First Method of RT Prediction**

Let \( i = 1, \ldots, I \) denote the items in the pool and \( k = 1, \ldots, n \) the items in the test. Thus, \( i_k \) is the index of the item in the pool administered as the \( k \)th item in the test. Suppose \( k - 1 \) items have already been administered. Let \( u_{k-1} = (u_1, \ldots, u_{k-1}) \) and \( t_{k-1} = (t_1, \ldots, t_{k-1}) \) denote the observed responses and RTs on these items, where, for convenience, the index for the test taker
has been omitted. The prediction equations derived in the following sections are for the selection of the $k$th item in the adaptive test.

**Posterior Distribution of Speed Parameter**

When the test begins, the only unknown parameters are the test taker’s ability $\theta$ and speed $\tau$. Procedures for updating the estimates of $\theta$ during the test belong to the standard tools of adaptive testing. Therefore, the focus is on updating the estimates of $\tau$, and a Bayesian approach is used to update the posterior distribution of this parameter.

From equation (3), the initial prior distribution of $\tau$ is taken to be

$$\tau \sim N(0, \sigma^2_{\tau}),$$

(4)

where $\sigma^2_{\tau}$ is known. The posterior distribution of $\tau$ given $t_{k-1}$ has density

$$f(\tau|t_{k-1}) \propto f(t_{k-1}|\tau)f(\tau) = \prod_{i=1}^{k-1} f(t_i|\tau) f(\tau),$$

(5)

with $f(t_i|\tau)$ and $f(\tau)$ the log-normal and normal densities in equations (2) and (4), respectively. Observe that equation (5) is based on the assumption of conditional independence between the RTs given the test taker’s speed. The assumption is analogous to that of “local independence” in IRT and can be motivated similarly: If speed is the only factor of the test taker that has an impact on RT and the test taker works at a stationary speed (which is what the model assumes), there is no source of variation left that could make the RTs on the items correlate.

When using $t^*_n = (\ln t_1, \ldots, \ln t_{k-1})$ instead of $t_{k-1}$, the likelihood becomes that for a normal distribution. Because a normal prior is conjugate for a normal likelihood with unknown means and known variances, the posterior distribution is also normal (e.g., Gelman, Carlin, Stern, & Rubin, 1995, sect. 2.6). Thus,

$$f(\tau|t^*_n) = N(\mu_{\tau|t^*_n}, \sigma^2_{\tau|t^*_n}).$$

(6)

But the posterior mean $\mu_{\tau|t^*_n}$ and variance $\sigma^2_{\tau|t^*_n}$ of $\tau$ are not identical to those for the standard case of observations from identical normal distributions (Gelman et al., 1995, eqs. 2.11-2.12). Each of the observations in $t^*_n$ is from a normal distribution with a partially unknown mean $\beta_i - \tau$ and a known variance, $\alpha_i^{-2}$. If the observations are redefined as $\beta_i - \tau^*$, they are normally distributed with a common unknown mean $\tau$ but different variances $\alpha_i^{-2}$. Therefore, the observations $\beta_i - \tau^*$ have to be weighted by their precision $\alpha_i^2$ and the expressions adjusted for the posterior mean and variance for the standard normal-normal model as

$$\begin{align*}
\mu^*_{\tau|t^*_n} &= \frac{\sum_{i=1}^{k-1} \alpha_i^2 (\beta_i - \tau^*)}{\sigma^{-2} + \sum_{i=1}^{k-1} \alpha_i^2}, \\
\sigma^2_{\tau|t^*_n} &= \left(\sigma^{-2} + \sum_{i=1}^{k-1} (\alpha_i^2)^{-1}\right)^{-1}.
\end{align*}$$

(7)

(8)
Posterior Predictive Density of Log Time

Let $T^*_{ik} = i_k$ be the predicted log time on candidate item $i_k$. The posterior predictive density of $T^*_{ik}$ is

$$f(T^*_{ik} | t^*_{k-1}) = \int (t^*_{ik} | \tau) f(t^*_{k-1}) dt_j.$$  \hspace{1cm} (9)

This density is the standard tool of prediction in Bayesian analysis and is used, for instance, to predict new observations in a model validation study or impute randomly missing data in surveys (e.g., Gelman et al., 1995, sect. 1.3).

Using the same argument as before, because both densities in the integrand are normal, it follows that the posterior predictive distribution of $T^*_{ik}$ given $t^*_{k-1}$ is normal with (a) mean equal to the posterior mean of $\beta_i - \tau$ given $t^*_{k-1}$ and (b) variance equal to the sum of the variance of $t^*_{ik}$ and the posterior variance of $\tau$ given $t^*_{k-1}$. Thus,

$$f(T^*_{ik} | t^*_{k-1}) = N\left(\beta_i - \tau, \sigma^2_t + \sum_{i=1}^{k-1} \alpha_i^2\right).$$  \hspace{1cm} (10)

with

$$\mu_{T^*_{ik} | t^*_{k-1}} = \beta_i - \frac{\sum_{i=1}^{k-1} \alpha_i^2 (\beta_i - t^*_i)}{\sigma^2_t + \sum_{i=1}^{k-1} \alpha_i^2} \hspace{1cm} (11)$$

and

$$\sigma^2_{T^*_{ik} | t^*_{k-1}} = \alpha_k^2 + \left(\sigma^2_t + \sum_{i=1}^{k-1} \alpha_i^2\right)^{-1}. \hspace{1cm} (12)$$

The two expressions in equations (11) and (12) contain only known parameters and log times recorded on the earlier items. It is therefore easy to update the predictive densities for the items in the pool each time a new item is administered. In fact, the update is most simple because the second terms of equations (11) and (12) are constants across all items in the pool. How to use these updates to control the speededness of an adaptive test during its administration is explained later in this article.

Second Method of RT Prediction

The idea is to predict the RTs on the remaining items in the pool not only from the RTs that have already been recorded but also from the responses. This use of the responses as an additional source of information on the speed at which the test taker operates becomes possible by using a portion of the hierarchical framework for speed and accuracy on test items presented in van der Linden (2007). The part of this framework needed for the current application is an extension of the second-level model in equation (3) to the joint distribution of the two person parameters $\theta$ and $\tau$. This distribution connects parameters from two different first-level models, namely, the response model in equation (1) and the RT model in equation (2).
Let $\xi = (0, \tau)$ denote the vector with the two person parameters. Assume that the distribution of $\xi$ in the population of test takers is approximately bivariate normal. That is,

$$f(\xi) = N(\mu, \Sigma),$$

with mean vector

$$\mu = (\mu_0, \mu_\tau)$$

and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_0^2 & \sigma_{0\tau} \\ \sigma_{0\tau} & \sigma_\tau^2 \end{pmatrix}.$$  

In addition to setting $\mu_\tau = 0$ to fix the speed scale, $\mu_0 = 0$ and $\sigma_0 = 1$ are set to fix the zero and unit of the ability scale. Thus, the only new parameters that have to be estimated when using this larger population model are $\sigma_\tau^2$ and $\sigma_{0\tau}$. The estimates can be calculated from the RTs and responses in the calibration sample during regular item calibration using the extended version of the Gibbs sampler in van der Linden (2007).

The normal density of the population distribution in equation (13) can be factorized as

$$f(\xi) = f(t|y)f(y).$$

Because $f(\xi)$ is normal, the same holds for the conditional density of $\tau$ given $0$:

$$f(\tau|0) = N(\mu_{\tau|0}, \sigma_{\tau|0}^2).$$

Its mean and variance are equal to

$$\mu_{\tau|0} = \sigma_{0\tau}0$$

and

$$\sigma_{\tau|0}^2 = \sigma_\tau^2 - \sigma_{0\tau}^2,$$

respectively.

The marginal density of $0$ in equation (16) is used as the prior distribution to calculate the density $f(0|u_{k-1})$ of the posterior distribution of $0$ given $u_{k-1}$. The density is combined with the conditional density of $\tau$ given $0$ in equation (17) to calculate the posterior predictive density of $\tau$ given $u_{k-1}$, that is, $f(\tau|u_{k-1})$. This density contains the information on the test taker’s speed parameter $\tau$ in the responses on the first $k - 1$ items. Therefore, it makes sense to replace the prior distribution $f(0)$ in equation (5) by the updates of the more informative empirical density $f(0|u_{k-1})$ during the test. Similar choices were made in van der Linden (2008) for the reverse problem of updating the estimate of $0$ in adaptive testing and in van der Linden and Guo (2008) to detect aberrances in RTs due to cheating by the test takers. The current choice leads to the following results.

**Alternative Posterior Distribution of Speed Parameter**

The posterior density of $0$ given $u_{k-1}$ is

$$f(0|u_{k-1}) = \prod_{i=1}^{k-1} f(u_i|0) f(0) d0.$$
Because the likelihood in equation (20) is not normal, neither is the posterior density \( f(\theta|u_{k-1}) \). But the density has been shown to converge strongly to normality (Chang & Stout, 1993). As the test is adaptive, the convergence is faster than for a standard fixed test. Following Owen (1975), the posterior distribution of \( \theta \) given \( u_{k-1} \) was assumed to be approximately normal:

\[
f(\theta|u_{k-1}) \approx N\left(\mu_{0|u_{k-1}}, \sigma^2_{0|u_{k-1}}\right).
\]

(21)

Observe that \( \mu_{0|u_{k-1}} \) is the posterior mean of \( \theta \), routinely used as its expected a posteriori (EAP) estimate in adaptive testing, and \( \sigma^2_{0|u_{k-1}} \) is a standard measure of its accuracy. These quantities are then automatically available.

The posterior predictive density of \( \tau \) given \( u_{k-1} \) is

\[
f(\tau|u_{k-1}) = \int f(\tau|\theta)f(\theta|u_{k-1})d\theta.
\]

(22)

Because both densities in the integrand are approximately normal, this posterior is also approximately normal. Hence, from equations (18) and (19),

\[
f(\tau|u_{k-1}) \approx N(\mu_{\tau|u_{k-1}}, \sigma^2_{\tau|u_{k-1}}),
\]

(23)

with

\[
\mu_{\tau|u_{k-1}} = \sigma_{0\tau}\mu_{0|u_{k-1}}
\]

(24)

and

\[
\sigma^2_{\tau|u_{k-1}} = \sigma^2_{0\tau} - \sigma^2_{0\tau} + \sigma^2_{0\tau}\sigma^2_{0|u_{k-1}}.
\]

(25)

The goal was to replace \( f(\tau) \) by \( f(\tau|u_{k-1}) \) as the empirical prior in the posterior distribution of \( \tau \) in equation (5). Thus,

\[
f(\tau|t^*_k, u_{k-1}) \propto f(t^*_k|\tau)f(\tau|u_{k-1}) = \prod_{i=1}^{k-1} f(t^*_i|\tau) f(\tau|u_{k-1}).
\]

(26)

Normality of \( f(t^*_i|\tau) \) and approximate normality of \( f(\tau|u_{k-1}) \) imply

\[
f(\tau|t^*_k, u_{k-1}) \approx N(\mu_{\tau|t^*_k, u_{k-1}}, \sigma^2_{\tau|t^*_k, u_{k-1}}).
\]

(27)

The posterior mean and variance are found by adjusting the expressions for the standard normal-normal model (Gelman et al., 1995, eqs. 2.11-2.12) as

\[
\mu_{\tau|t^*_k, u_{k-1}} = \frac{\sigma_{0\tau}}{\sigma^2_{0\tau} - \sigma^2_{0\tau} + \sigma^2_{0\tau}\sigma^2_{0|u_{k-1}}} \mu_{0|u_{k-1}} + \frac{1}{\sigma^2_{0\tau} - \sigma^2_{0\tau} + \sigma^2_{0\tau}\sigma^2_{0|u_{k-1}}} \sum_{i=1}^{k-1} \beta_i (\beta_i - \ln t_i),
\]

(28)

and

\[
\sigma^2_{\tau|t^*_k, u_{k-1}} = \left(\sigma^2_{0\tau} - \sigma^2_{0\tau} + \sigma^2_{0\tau}\sigma^2_{0|u_{k-1}}\right)^{-1} + \sum_{i=1}^{k-1} \alpha_i^2.
\]

(29)
Alternative Posterior Predictive Density of Log Time

The proposed alternative posterior predictive density of $\tilde{t}_i^*$ is

$$f(\tilde{t}_i^* | t_{k-1}^*, u_{k-1}) = \int (\tilde{t}_i | \tau) f(\tau | t_{k-1}^*, u_{k-1}) d\tau.$$  \hspace{1cm} (30)

Using a similar argument as for the derivation of equations (9) to (12), the density is approximately normal:

$$f(\tilde{t}_i^* | t_{k-1}^*, u_{k-1}) \approx N(\mu_{i|t_{k-1}, u_{k-1}}, \sigma_{i|t_{k-1}, u_{k-1}}^2),$$  \hspace{1cm} (31)

with

$$\mu_{i|t_{k-1}, u_{k-1}} = \beta_i - \frac{\sigma_{0r}}{\sigma_{0r}^2 + \sigma_{0c}^2} \mu_{0|u_{k-1}} + \sum_{i=1}^{k-1} (\beta_i - \ln t_i)$$  \hspace{1cm} (32)

and

$$\sigma_{i|t_{k-1}, u_{k-1}}^2 = \alpha_i^2 + \left(\sigma_{0r}^2 - \sigma_{0c}^2 \sigma_{0u_{k-1}}^2 \right)^{-1} + \sum_{i=1}^{k-1} (\alpha_i^2).$$  \hspace{1cm} (33)

These expressions contain the same (known) parameters and log times recorded for the earlier items as in equations (11) and (12). The additional information in the responses is summarized by the known variance $\sigma_r^2$ and covariance $\sigma_{cr}$ as well as the posterior mean, $\mu_{i|u_{k-1}}$, and variance, $\sigma_{i|u_{k-1}}^2$, of $\theta$ given the responses $u_{k-1}$. As already noted, the calculation of the two posterior quantities is a standard routine in Bayesian adaptive testing.

Controlling for Differential Speededness

The case of adaptive testing with a fixed test length and time limit is discussed next. Let $S_{k-1}$ denote the set of $k - 1$ items that have already been administered and $R_k$ the set of remaining items in the pool. The $k$th item is selected from $R_k$. Because the actual RTs recorded for the items in $S_{k-1}$ and the predicted RTs on all items in $R_k$ are available, the information is used to constrain the selection of the items to meet the time limit for the test. The idea was implemented in van der Linden et al. (1999) using the shadow-test approach (STA) to adaptive testing (van der Linden, 2005, chap. 9). The STA selects the items from a shadow test, which contains the projection of the remaining portion of the test that is maximally informative and meets all content constraints. In the current application, the shadow tests were assembled subject to the constraint that the sum of the actual RTs on the items already administered and the predicted RTs on the remaining items met the time limit for the test. As a result, the adaptive test automatically satisfied the time limit.

Let $t_{\text{lim}}$ be the time limit for the test and $x_i$ the decision variable for the selection of item $i = 1, \ldots, I$; that is, $x_i = 1$ if item $i$ is selected and $x_i = 0$ if it is not. The total time recorded for the first $k - 1$ items is

$$\sum_{i \in S_{k-1}} t_i x_i.$$  \hspace{1cm} (34)
To predict the remaining portion of the test time, the following expression is used:

$$\sum_{i \in R_k} \exp(\tilde{t}^*_i)x_i,$$

where $\tilde{t}^*_i$ is a well-chosen realization of the predicted log time $\tilde{T}^*_i$ in equation (10) or (31). The exponential function transforms the RTs from the logarithmic to their regular scale. An obvious choice for $\tilde{t}^*_i$ is a quantile in the posterior predictive distribution. Early in the test, the median of the distributions could simply be chosen, but toward the end of the test a more conservative choice, such as their .95 quantile, is recommended.

For an adaptive test with maximization of the information on $\theta$ as the objective, the model for the assembly of the shadow test is

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

subject to

$$\sum_{i \in R_k} \exp(\tilde{t}^{\pi_k})x_i \leq t_{\text{lim}} - \sum_{i \in S_{k-1}} t_i x_i,$$

$$\sum_{i \in S_k} x_i = k - 1,$$

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

and

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

The objective function in equation (36) maximizes the information in the test at the current estimate $\hat{\theta}_{k-1}$. The constraint in equation (37) requires the predicted time on the remaining portion of the test not to be larger than the time available, where $\tilde{t}^{\pi_k}$ denotes the quantile of the predicted distribution used for the selection of the $k$th item. The constraint in equation (38) fixes to 1 the values of the decision variables of the items that were already administered, whereas the length of the adaptive test is fixed in equation (39). The last set of constraints defines the range of the decision variables. The model, which can easily be extended with other constraints to deal with the entire set of specification for the adaptive test, can be solved using a standard solver for integer programming such as CPLEX (ILOG, Inc., 2003). One of the attractive features of linear integer programming is that there are no interactions between the constraints (that is, one can be formulated independently of the others). Also, efficient implementations of the solver are available that enable the selection of shadow tests with hundreds of constraints within a split second (van der Linden, 2005, sect. 9.1.5).

The item selected for administration is the most informative one among the $n - k + 1$ free items in the shadow test. Because the shadow test is required to meet the constraint in equation (37) for the selection of each item, the entire adaptive test meets the time limit, $t_{\text{lim}}$. 
Empirical Study

Simulation studies were conducted to evaluate the effectiveness of the approach in eliminating differences in speededness between test takers under realistic conditions. The test was a 15-item adaptive test from an item pool from the ASVAB Arithmetic Reasoning test.

Three different factors were manipulated:

1. The time limit for the test, which controls the time pressure under which the test takers have to work, was varied by $t_{lim} = 39, 34, \text{ and } 29 \text{ min.}$ The first time was the actual limit for the ASVAB test; the other two conditions were added to see what would happen if the degree of speededness of the test were increased. The same levels were chosen for the earlier procedure in van der Linden et al. (1999).

2. The two methods of predicting the time on the remaining items in the pool were varied, that is, prediction with and without the use of the responses as an additional source of information.

3. For the second method, the correlation between speed and ability was varied. The values $\rho_{yt} = .0(.2).8$ were used. This correlation does not play any role in the first method.

In addition, a simulation was run in which the time limit was removed and the test was not controlled for speededness. The results from this simulation show the potential differential speededness inherent in the adaptive test and can be used as a baseline for evaluating the effectiveness of the control of the degree of speededness in the other conditions.

The data set for the ASVAB test was from its operational use. Because the test length was quite short, the set was too sparse to get stable estimates of the correlations between the item parameters in the IRT and RT models. Therefore, an item pool was generated with parameter values that matched the ranges of the estimates of the IRT item parameters used in operational testing as well as the estimates for the RT model from an earlier study for the same ASVAB test, in which the model showed excellent fit to the data (for a report, see van der Linden, 2006). In addition, to allow comparison of the results of the current study with those in van der Linden et al. (1999), the same correlation between the item difficulty and time intensity parameters was assumed as that found in this earlier study ($\rho_{dt} = .65$). (It is important to distinguish this correlation between two different item parameters from the correlation between the test takers’ speed and ability parameters, $\rho_{yt}$, used below.) More specifically, a pool of 350 items was generated with parameters $a_i$, $b_i$, and $c_i$ in the response model as well as parameters $x_i$ in the RT model drawn from uniform distributions over the same ranges as for the ASVAB data set. In addition, the parameters $\beta_i$ were drawn from a normal distribution with mean and standard deviation conditional on $b_i$ to realize the earlier correlation $\rho_{\beta\beta}$.

In each condition, adaptive test administrations for different combinations of values for the speed and ability parameters were simulated. For the predictive method in equation (10), the values of these parameters were fixed at

1. $\tau = -.68, -.34, 0, .34, .68$;
2. $\theta = -2.0, -1.5, \ldots, 2.0$.

Both sets of values covered a range of $\pm 2$ standard deviations for the ASVAB population. (The population distribution to which the items were scaled had a mean of $\mu_0 = 0.0$ and a standard deviation $\sigma_0 = 1$; in addition, $\sigma_1$ was estimated to be equal to .34.) For the predictive method in equation (31), the values of $\tau$ were also fixed but the $\theta$s were sampled conditionally on $\tau$ to realize the desired level of correlation. The procedure was as follows:

1. $\tau = -.68, -.34, 0, .34, .68, -.34$;
2. $\theta$ was drawn from a normal distribution with mean $\mu_{\theta | \tau}$ and standard deviation $\sqrt{1 - \rho_{\theta \tau}^2}$. To obtain the same range as for the first method, all values outside the interval $[-2, 2]$ were ignored. As already indicated, the procedure was repeated for $\rho_{yt} = .0(.2).8$. 


The number of test administrations was 1,800 for each value of \( t \). For the conditions with \( y \) fixed, 300 administrations were simulated for each value of \( y \). For the conditions with \( y \) randomly sampled conditional on \( t \), the values were grouped in intervals of size .25 about \(-2\), \(-1.5\), \ldots, 2 to produce results comparable to those for the conditions with \( y \) fixed.

The items in the adaptive test were selected from the shadow tests in equations (36) to (40) and calculated using the integer solver in CPLEX. For the selection of item \( k = 1 \) in the adaptive test, the point predictions \( \tilde{\theta}_i \) in the constraint in equation (37) were calculated as the median of the posterior predictive distributions of the RTs. Thereafter, it moved in equal steps to the .95 quantiles for the selection of items \( k = 14 \) and 15. The abilities were estimated using the method of EAP estimation with a uniform prior over \([-4, 4]\). For each simulated test administration, both the total time and the error in the final estimate of \( y \) were recorded. The former helped to evaluate the impact of the time constraint on the degree of speededness of the test, the latter to evaluate a loss in the accuracy of the ability estimates that possibly had to be paid as a price for the constraint on the selection of the items.

**Results**

The effectiveness of the first method can be evaluated using the results illustrated in Figures 1-3. Each of the plots in Figure 1 shows the average total time as a function of the speed \( \tau \) at which the test takers worked for the different ability groups. To reduce the number of curves, only those for the
ability groups $\theta = 2, 1, 0, -1, \text{ and } -2$ are shown. (The curves for the other groups reveal the same pattern.) The curves are ordered from high (top) to low ability (bottom). This order reflects the fact that in an adaptive test the more able students get the more difficult items, which tend to be more time intensive. Obviously, in each of the plots, the average total time increased with decreasing $t$. The top-left plot for the baseline case with no time limit reveals only a slight tendency to differential speededness of the test: The test takers with the highest ability and lowest speed tended to run out of time. However, their average time was only 30 sec higher than the time limit for the test. In van der Linden et al. (1999), a more serious level of differential speededness was observed. It is now clear that this impression must have been the result of a much less flexible RT model and poor parameter estimation. Both contributed to a poor fit of the estimated model.

The plot for the actual time limit ($t_{\text{lim}} = 2,340$ sec) shows that the constraint was effective in removing the risk of running out of time during the test for the groups of test takers that were potentially threatened by its degree of speededness. However, the true power of the constraint is shown in the plots for which the limit was lowered artificially by 5 min ($t_{\text{lim}} = 2,040$ sec) and 10 min ($t_{\text{lim}} = 1,740$ sec). A comparison of these two plots with the first plot helps with inferring what groups of test takers could have been threatened by these more stringent limits. Nevertheless, none of them ran out of time.

The bias in the final estimates of $\theta$ in Figure 2 varied about zero and was largest for the extreme ability groups. The sign of the bias reflected an inward bias typical of adaptive testing with EAP estimation of $\theta$ (positive bias for negative $\theta$s and negative bias for positive $\theta$s). Their absolute size

Note. Prediction is without the use of responses as additional information; the bolder the curve, the higher the ability.
was as expected for a regular adaptive test of this length. Again, the horizontal curves do not suggest any bias related to the speed at which the test takers operated, and there were no systematic differences between the conditions with and without the time constraint.

The plots with the mean square errors in the final estimates of $\hat{\theta}$ in Figure 3 show the possible impact of the time constraint on the accuracy of the ability estimation. The lines for the different ability groups run horizontally in each plot; this shows that the mean square error was independent of the speed at which the test takers operated. Also, the mean square errors for the three conditions with time constraints were entirely comparable to the baseline case with no constraint. The results indicate that the price paid for control of speededness through a constraint on the item selection was nil. Of course, the size of the mean square errors is associated with the true value of $\theta$. Generally, mean square errors close to .1 are in agreement with what the author has seen in earlier simulations of adaptive tests of comparable lengths.

Figures 4-6 illustrate the version of the constraint based on prediction of the RTs on the remaining items in the pool from the earlier RTs and responses jointly. For lack of space, only the results for a correlation in the middle of the range ($\rho_{t\tau} = .4$) are presented; the results for the other correlations were entirely comparable.

A comparison between the plots with the total time in Figures 1 and 4 shows hardly any differences between the two methods. The reason, of course, is that the previous method was already most effective and did not leave any space for improvement. The only difference was a slightly

Figure 3
Mean Square Error ($MSE$) of $\hat{\theta}$ for Different Ability Groups as a Function of Speed Parameter $\tau$ for Conditions With No Time Limit and With the Time Constrained to 2,340, 2,040, and 1,740 Sec

![Figure 3](image-url)

Note. Prediction is without the use of responses as additional information; the bolder the curve, the higher the ability. $MSE = \text{mean square error.}$
smaller variation in total time between the ability groups for the second method. The decrease was probably a consequence of the fact that using the actual responses accounted for the differences in ability between the test takers. From the point of view of test standardization, this should be considered a favorable effect.

Consequently, the plots in Figures 5 and 6 show the same pattern for the bias and mean square error functions as for the first method. The only difference is somewhat smaller errors for the second method, but the effect may not be systematic. Because the values of $\theta$ for the second method were grouped in small intervals, the curves display some random variation.

Concluding Remarks

The simulated adaptive test showed less differential speededness than was expected from the results in an earlier study. There was a large variation in the time needed to complete the test between the different simulated groups of test takers, but none of them ran completely out of time because the actual time limit was too tight. The current impression of the degree of speededness of the test is expected to be closer to reality than that from the earlier study because of a more flexible RT model and better parameter estimation.

Nevertheless, manipulating the time limit of the test demonstrated the effectiveness of the method of control of speededness for tighter limits. The time usage across the groups of test takers became much more uniform, but still none of the test takers ran out of time. For this form of

Note. Prediction is with the use of responses as additional information ($r_{it} = .4$); the bolder the curve, the higher the ability.
The predictive control of test speededness in this study was implemented under the assumption that measuring $\Theta$ was the primary goal of the test and speed could be treated as a nuisance factor. The assumption was obvious from the fact that the time constraint in equation (37) adapted the selection of the items to the actual speed of the test takers and did not prescribe any minimally acceptable level. The assumption is reasonable for the current generation of adaptive tests, which do not offer their candidates much support in the way of time management. Typically, the only source of information they have is a display of the time elapsed or remaining on the computer screen. As the test takers receive one item at a time and have to guess how much time their future items will require, appropriate management of their time must be a challenge. Under such conditions, it would be unfair to punish test takers seriously for unanticipated changes in the time intensity of the items.

On the other hand, when test takers are provided with appropriate tools for time management and speed is part of the definition of the construct measured by the test, a simple adjustment of the method can be used. Researchers should then choose a minimally acceptable level of speed $\tau_0$ for the test and predict the RT distributions on the candidate items in the pool at this fixed $\tau_0$ instead of the estimated actual speed of the test takers. For these predictions, the constraint in equation (37) guarantees that a test taker working at the minimally acceptable level of speed will meet the time limit for the test.

**Figure 5**
Bias of $\hat{\Theta}$ for Different Ability Groups as a Function of Speed Parameter $\tau$ for Conditions With No Time Limit and With the Time Constrained to 2,340, 2,040, and 1,740 Sec

*Note.* Prediction is with the use of responses as additional information ($\rho_{hi} = .4$); the bolder the curve, the higher the ability.

standardization, the price paid in greater errors in the final $\Theta$ estimates appeared to be negligible, even for the conditions with more stringent time limits. In this study, the two prediction methods did not differ in their effectiveness. Because the first method already performed completely as desired, there was not much left for the second method to improve on.
Mean Square Error (MSE) of $\hat{\theta}$ for Different Ability Groups as a Function of Speed Parameter $\tau$ for Conditions With No Time Limit and With the Time Constrained to 2,340, 2,040, and 1,740 Sec

**Note.** Prediction is with the use of responses as additional information ($r_{bi} = .4$); the bolder the curve, the higher the ability. $MSE = \text{mean square error}$.

### References


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


**Acknowledgments**

This study received funding from the Law School Admissions Council (LSAC). The opinions and conclusions contained in this article are those of the author and do not necessarily reflect the policy and position of the LSAC. The author is indebted to Wim M. M. Tielen for his computational support and to the Defense Manpower Data Center for making available the data set.

**Author’s Address**

Address correspondence to W. J. van der Linden, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands; e-mail: w.j.vanderlinden@utwente.nl.