PSYCHOMETRIKA—VOL. 73, NO. 3, 365–384
SEPTEMBER 2008
DOI: 10.1007/S11336-007-9046-8

BAYESIAN PROCEDURES FOR IDENTIFYING ABERRANT RESPONSE-TIME PATTERNS IN ADAPTIVE TESTING

WIM J. VAN DER LINDEN

UNIVERSITY OF TWENTE

FANMIN GUO

GRADUATE MANAGEMENT ADMISSION COUNCIL

In order to identify aberrant response-time patterns on educational and psychological tests, it is important to be able to separate the speed at which the test taker operates from the time the items require. A lognormal model for response times with this feature was used to derive a Bayesian procedure for detecting aberrant response times. Besides, a combination of the response-time model with a regular response model in an hierarchical framework was used in an alternative procedure for the detection of aberrant response times, in which collateral information on the test takers' speed is derived from their response vectors. The procedures are illustrated using a data set for the Graduate Management Admission Test® (GMAT®). In addition, a power study was conducted using simulated cheating behavior on an adaptive test.

Key words: adaptive testing, Bayesian predictive checks, cheating, collateral information, hierarchical modeling, response times.

Statistical procedures for identifying aberrances on educational and psychological tests are essential to detect flaws in the design of the test or irregular behavior by the test takers. Examples of possible aberrances include bad test items, speededness of the test, ambiguous instructions, dyslexic test takers or test takers who need special test accommodations, and answer copying during the test. The focus in this paper is particularly on the identification of aberrances in computerized adaptive testing that indicate possible cheating, such as preknowledge of some of the items and attempts to take tests only for the purpose of memorizing the items.

The traditional approach to detecting aberrant behavior is to analyze the response vectors for patterns of unexpected responses. This type of analysis is known as item-misfit or person-misfit analysis in item response theory (Meijer & Sijtsma, 1995) and belongs to the larger class of problems of outlier detection in statistics. Key to the approach is the availability of a response model that adequately represents regular behavior by the test takers.

One of the mainstream models in adaptive testing is the three-parameter logistic (3PL) model. Suppose we have to check the responses by test takers $j = 1, \ldots, N$ on the items in the test, $i = 1, \ldots, n$. The responses are denoted as binary variables U_{ij}, which take the value 1 for a correct and 0 for an incorrect response. The model postulates the following probabilities of a correct response for regular test takers

The authors have relied upon data supplied by the Graduate Management Admission Council® (GMAC®) to conduct the independent research that forms the basis for the findings and conclusions stated in this article. These findings and conclusions are the opinion of the authors only, and do not necessarily reflect the opinion of the GMAC®. The authors are indebted to Wim M.M. Tielen and Rinke H. Klein Entink for their computational support.

Requests for reprints should be sent to Wim J. van der Linden, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: w.j.vanderlinden@utwente.nl

$$\begin{aligned}\Pr\{U_{ij}=1\} &\equiv p_i(\theta_j;a_i,b_i,c_i)\\ &\equiv c_i+(1-c_i)\frac{\exp[a_i(\theta_j-b_i)]}{1+\exp[a_i(\theta_j-b_i)]},\end{aligned}\tag{1}$$

where $\theta_j \in \Re$ is the ability of test taker j and $b_i \in \Re$, $a_i \in \Re^+$, and $c_i \in [0,1]$ are the difficulty, discrimination, and guessing parameters for item i, respectively. In a typical adaptive testing program, the item parameters are estimated with enough precision to treat them as known. We therefore suppress their notation in this manuscript wherever this is convenient.

Because U_{ij} is binary, the residual response by test taker j on item i is

$$\begin{aligned}E_{ij} &\equiv U_{ij}-\mathcal{E}(U_{ij})\\ &= U_{ij}-p_i(\theta_j).\end{aligned}\tag{2}$$

It holds that E_{ij} is Bernoulli distributed with probability $p_i(\theta_j)$ of outcome $1-p_i(\theta_j)$ and probability $q_i(\theta_j)=1-p_i(\theta_j)$ of outcome $-p_i(\theta_j)$. A pattern of residuals with small probabilities might indicate a defective item or a person with aberrant behavior, where the specific shape of the pattern offers an explanation of the misfit (e.g., a pattern of several unlikely correct answers might point at cheating).

To calculate residual responses, the ability parameter θ_j in (2) has to be estimated. Although errors in the estimates of the item parameters are typically negligible in adaptive testing, this does not necessarily hold for θ_j, especially if it has to be estimated from a smaller subset of trusted items in the test. It is therefore more appropriate to allow for estimation error in θ_j and use a Bayesian version of (2).

One possible Bayesian version is to replace $\mathcal{E}(U_{ij})$ by the cross-validation predictive value of the response. Let $\mathbf{u}_j=(u_{1j},\ldots,u_{nj})$ be the responses by j on the n items in the test and suppose we evaluate the response on the ith item. The posterior distribution of the predicted response $\tilde{U}_{ij}$ given the responses on the remaining items, $\mathbf{u}_{j\setminus i}$, is given by

$$f\left(\tilde{u}_{ij}\mid\mathbf{u}_{j\setminus i}\right)=\int f\left(\tilde{u}_i\mid\theta_j\right)f(\theta_j\mid\mathbf{u}_{j\setminus i})\,d\theta_j,\tag{3}$$

where $f(\tilde{u}_k\mid\theta_j)$ follows from the response model in (1) as

$$f\left(\tilde{u}_i\mid\theta_j\right)=p_i(\theta_j)^{\tilde{u}_i}q_i(\theta_j)^{1-\tilde{u}_i}\tag{4}$$

and the posterior density of θ_j given $\mathbf{u}_{j\setminus i}$ is

$$f(\theta_j\mid\mathbf{u}_{j\setminus i})\propto f(\theta_j)\prod_{k\neq i}p_k(\theta_j)^{u_k}q_k(\theta_j)^{1-u_k},\tag{5}$$

where $f(\theta_j)$ is a prior density of θ_j.

Now, an alternative to (2) is the cross-validation predictive residual

$$\begin{aligned}E_{ij} &\equiv u_{ij}-\mathcal{E}\left(\tilde{U}_{ij}\mid\mathbf{u}_{j\setminus i}\right)\\ &= u_{ij}-f(1\mid\mathbf{u}_{j\setminus i})\end{aligned}\tag{6}$$

(e.g., Johnson & Albert 1999, Sect. 3.4; Gelman, Carlin, Stern, & Rubin, 1995, Sect. 6.2). This residual is Bernoulli distributed with probability $f(1\mid\mathbf{u}_{j\setminus i})$ of outcome $1-f(1\mid\mathbf{u}_{j\setminus i})$ and probability $f(0\mid\mathbf{u}_{j\setminus i})$ of outcome $-f(1\mid\mathbf{u}_{j\setminus i})$. For an application of this type of residual to adaptive tests, see Bradlow, Weiss, and Cho (1998) or van der Linden and van Krimpen-Stoop (2003).

These two procedures certainly do not exhaust the possibilities of person-fit analysis. In fact, the area has been well developed, and it would be beyond the scope of this paper to discuss all alternatives. For a recent review of classical procedures, the reader should consult Meijer and Sijtsma (2001); for Bayesian approaches, Glas and Meijer (2003) is recommended. The two residual procedures above have been selected because they focus on the individual responses and in this respect parallel the response-time procedures proposed in the following.

A general problem with residuals as in (2) and (6) arises, however, when the difficulties of the items and the abilities of the test takers are close. The probabilities for their two possible values then approximate 0.50, and response patterns that otherwise might point at aberrances are now much more likely with regular behavior. Thus, under this condition, a statistical test based on response residuals loses its power. The condition is automatically met in adaptive testing, where the response probabilities of the test taker on the test items quickly converge to a value close to 0.50 by design.

We therefore focus on the response times (RTs) on the items as an additional source of information on the test taker's behavior. There are at least three reasons why RTs might be a better source of information on possible aberrances than the responses themselves. First, they are continuous instead of binary. As a result, they lend themselves better to statistical treatment and offer more information on the "size" of aberrances when they happen.

Second, RTs are insensitive to the design effect mentioned above. If a test matches the test taker's ability, the selection of the items does not constrain the likelihood of his/her RTs in any systematic way, and it remains possible to discriminate between likely and unlikely patterns of RTs. As a result, an RT-based statistical check on possible aberrances can be expected to maintain its power throughout the test.

Third, RTs are the result of the speed at which the test takers work on the items as well as their time intensities (that is, the amount of work they require). No matter how fast or slow a test taker works, for a well-designed test, except for random fluctuation, the RTs have to follow the pattern of time intensities of the items in the test. If we have an RT model that allows us to separate the speed from the time intensities, we can adjust the test takers' RTs for their speed and check if the results follow the pattern of time intensities. We expect this to be a very effective check because the time intensities of the items in a pool for an adaptive test easily differ by a factor larger than five or so. Even if test takers simulated RTs to hide the fact that they are cheating, they would have to find out what a typical pattern would be for the selection of items they get. We expect this to be an impossible task, especially because it has to be executed as the time is already recorded.

To our knowledge, the only earlier publication on the use of RTs for checking on aberrances on test items in adaptive testing is van der Linden and van Krimpen-Stoop (2003). However, their RT model was a simple main-effects ANOVA model for the logarithm of the RTs used in van der Linden, Scrams, and Schnipke (1999), for which they had no adequate statistical procedures for estimating its parameters and testing its goodness of fit. In this research, we will use a model with a more realistic parameterization for which such procedures do exist. In addition, we will use a combination of the RT model with a regular response model in a hierarchical framework that allows us to use the information in the response vector as collateral information on the test taker's speed in an alternative Bayesian check on possible aberrances.

It is not our intention to advocate mechanical use of these checks. As explained earlier, test behavior may show aberrances for several reasons, and it would be wrong to jump to easy conclusions. But the following two applications of the proposed checks make sense. First, the checks could be used as a routine to flag test takers or items that may need further consideration. The next step in the analysis should then consist of collecting additional information and diagnosing the RT pattern more carefully. Second, they could be used as a more formal proof or disproof of other evidence that was collected during the test (e.g., observations by proctors).

RT Model

The RT model was proposed by van der Linden (2006). It is discussed here only briefly (for technical details, Bayesian estimators of its parameters, and procedures for analyzing its goodness of fit, see the reference). The model will be used to present the new procedures for detecting aberrant RTs presented in the next sections.

The model is a lognormal density for the distribution of the response time, T_{ij} with a parameter structure analogous to that of the logistic expression in (1):

$$f(t_i;\tau_j,\alpha_i,\beta_i)=\frac{\alpha_i}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_{ij}-(\beta_i-\tau_j)\right)\right]^2\right\}. \tag{7}$$

Parameter $\tau_j\in\Re$ represents the speed at which test taker j operates during the test, $\beta_i\in\Re$ is a parameter for the time intensity of the item, and $\alpha_i\in\Re^+$ is a discrimination parameter. The time intensity of the items is determined by the amount of work they require to produce a response to item i. Observe that a lognormal distribution of T_{ij} is the same as a normal distribution of $\ln T_{ij}$. Hence, $\ln T_{ij}$ has mean $\mu_{ij}=\beta_i-\tau_j$. The faster the test taker operates, the lower the mean. Likewise, the higher the time intensity of the item, the higher the mean. Parameter α_i modifies the relation between $\ln T_{ij}$ and $\tau_j-\beta_i$. The larger its value, the less the dispersion of $\ln T_{ij}$ around $\tau_j-\beta_i$, and hence the better the discrimination between RTs of test takers with a speed higher and lower than β_i. Parameter α_i thus also reflects the amount of "noise" that happens when a person takes an item and allows his/her thoughts to divert a little, stretches the arms and legs, and the like.

The parameters in (7) have a natural scale with a unit that is determined by the precision at which the RTs are measured (e.g., in seconds). But the two individual parameters in mean of the distribution of the log-times, $\beta_i-\tau_j$, are not yet identifiable. To establish identifiability, it is convenient to impose the constraint

$$\mu_\tau=0 \tag{8}$$

on the mean speed of the test takers.

In the remainder of this paper we will assume that the parameters α_i and β_i of the items in the pool for the adaptive test have been estimated with enough precision to treat them as known and suppress their notation wherever convenient. This assumption is usual in adaptive testing.

Posterior Predicted Response Times

Suppose the population of test takers has an approximate normal distribution for the speed parameter

$$\tau\sim N\left(\mu_\tau,\sigma_\tau^2\right). \tag{9}$$

If the constraint in (8) is used, $\mu_\tau=0$; how to estimate σ_τ as part of the calibration of the item pool will be discussed later. We use (9) as a common prior distribution of the speed parameters of the individual test takers.

Let $\mathbf{t}_j=(t_{1j},\ldots,t_{nj})$ denote the observed response times by test taker j on the items in the test. The posterior distribution of τ given $\mathbf{t}_j$ has density

$$\begin{aligned} f(\tau_j\mid\mathbf{t}_j)&\propto f(\mathbf{t}_j\mid\tau_j)f(\tau_j)\\ &=\left[\prod_{i=1}^{n}f(t_{ij}\mid\tau_j)\right]f(\tau_j), \end{aligned} \tag{10}$$

where $f(t_{ij}\mid\tau_j)$ and $f(\tau_j)$ are the lognormal and normal densities in (7) and (9), respectively.

The posterior density simplifies to a closed form with known parameters if we use $\mathbf{t}_j^* = (\ln t_{1j}, \ldots, \ln t_{nj})$ instead of $\mathbf{t}_j$. This can be shown as follows. From (7), $\ln t_{ij} \sim \mathrm{N}(\beta_i - \tau_j, \alpha_i^{-2})$. Therefore,

$$\beta_i - \ln t_{ij} \sim N\left(\tau_j, \alpha_i^{-2}\right). \tag{11}$$

Thus, (11) can be viewed as the posterior density of normally distributed observations $\beta_i - \ln t_{ij}$ with unknown mean, τ_j, and known variance, α_i^{-2}, and a normal prior for the mean. It is a standard result in statistics that the posterior density of the mean is also normal. Thus,

$$f\left(\tau_j \mid \mathbf{t}_j^*\right) = N\left(\mu_{\tau|\mathbf{t}_j^*}, \sigma^2_{\tau|\mathbf{t}_j^*}\right). \tag{12}$$

However, the posterior mean $\mu_{\tau|\mathbf{t}_j^*}$ and variance $\sigma^2_{\tau|\mathbf{t}_j^*}$ are not identical to those for the standard case of identically distributed observations (for their expressions see Gelman et al., 1995, (2.11) and (2.12)). The reason is the differences between the variances α_i^{-2} for the items. We therefore have to weigh the individual observations $\beta_i - \ln t_{ij}$ by their precision, α_i^2. Hence, the posterior mean is

$$\mu_{\tau_j|\mathbf{t}_j^*} = \frac{\sigma_\tau^{-2}\mu_\tau + \sum_{i=1}^n \alpha_i^2(\beta_i - \ln t_{ij})}{\sigma_\tau^{-2} + \sum_{i=1}^n \alpha_i^2}. \tag{13}$$

Likewise, the posterior variance is

$$\sigma^2_{\tau_j|\mathbf{t}_j^*} = \left(\sigma_\tau^{-2} + \sum_{i=1}^n \alpha_i^2\right)^{-1}. \tag{14}$$

Posterior Predictive Density of RTs

The posterior predictive distribution of the log-time on an arbitrary item i given the times on all remaining items is

$$f\left(\tilde{t}_{ij}^* \mid \mathbf{t}_{j\setminus i}^*\right) = \int \left(\tilde{t}_{ij}^* \mid \tau_j\right) f\left(\tau_j \mid \mathbf{t}_{j\setminus i}^*\right) d\tau_j, \tag{15}$$

where $\tilde{T}_{ij}^* = \tilde{t}_{ij}^*$ is the predicted log-time by test taker j on item i and $\mathbf{t}_{j\setminus i}^*$ is the vector with the actual log-times on the remaining items in the test.

The distribution can be found using the same transformation of the observations as in (11). Again, as the posterior distribution of τ_j is normal, it follows from a standard result in statistics that the distribution of $\beta_i - \tilde{T}_{ij}^*$ given $\mathbf{t}_{j\setminus i}^*$ is also normal with mean equal to the posterior mean of τ_j on the remaining items. Therefore, the mean of $\tilde{T}_{ij}^*$ given $\mathbf{t}_{j\setminus i}^*$ is

$$\mu_{\tilde{T}_{ij}^*|\mathbf{t}_{j\setminus i}^*} = \beta_i - \frac{\sigma_\tau^{-2}\mu_\tau + \sum_{k\neq i}^n \alpha_k^2(\beta_k - \ln t_{kj})}{\sigma_\tau^{-2} + \sum_{k\neq i}^n \alpha_k^2}. \tag{16}$$

Likewise, its variance is equal to the sum of the variance of the log-time on item i and the posterior variance of τ_j given the remaining items,

$$\sigma^2_{\tilde{T}_{ij}^*|\mathbf{t}_{j\setminus i}^*} = \alpha_i^{-2} + \left(\sigma_\tau^{-2} + \sum_{k\neq i}^n \alpha_k^2\right)^{-1}. \tag{17}$$

Thus, for a calibrated item pool, the predicted log-time on an arbitrary item given the times on the other items has a known distribution with a density that is easy to calculate. For the identifiability constraint $\mu_\tau = 0$ in (8), the expression for the mean even simplifies. We will use this result to propose the new checks on aberrant RT patterns below.

Using Responses as Collateral Information

If we extend our model, it becomes possible to predict the RT on an arbitrary item not only from the RTs on other items but from their joint RTs and responses. The extension we use is derived from a hierarchical framework for analyzing speed and accuracy on test items in van der Linden (2007). The framework has both an RT model and a regular IRT model as first-level models and second-level models for all their item and person parameters. Because in adaptive testing all item parameters are known and the interest is only in estimating the person parameters, we only use the population model for the person parameters. For further details on the full hierarchical framework and Bayesian estimators of the parameters, the reader should consult the reference.

Extended Model

We retain the lognormal model in (7) but, for more convenient parameter estimation later in this paper, use the three-parameter normal-ogive (3PNO) model instead of the 3PL model in (1). That is, the probability of a correct response is assumed to be

$$p_i(\theta_j) \equiv c_i + (1 - c_i)\Phi\big(a_i(\theta_j - b_i)\big), \tag{18}$$

where $\Phi(\cdot)$ denotes the normal distribution function. As is well known, the difference between (18) and (1) is mainly in scale (Lord & Novick, 1968, Sect. 17.2), and all parameters keep their interpretation.

The population model describes the joint distribution of the speed and ability parameters $\boldsymbol{\xi} = (\theta, \tau)$ in population $\mathcal{P}$ from which the test takers are assumed to be sampled. We assume that $\boldsymbol{\xi}_j$ is randomly drawn from a bivariate normal distribution

$$\boldsymbol{\xi}_j \sim f(\boldsymbol{\xi}_j; \boldsymbol{\mu}_\mathcal{P}, \boldsymbol{\Sigma}_\mathcal{P}) \tag{19}$$

with density function

$$f(\boldsymbol{\xi}_j; \boldsymbol{\mu}_\mathcal{P}, \boldsymbol{\Sigma}_\mathcal{P}) = \frac{|\boldsymbol{\Sigma}_\mathcal{P}^{-1}|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}(\boldsymbol{\xi}_j - \boldsymbol{\mu}_\mathcal{P})^T \boldsymbol{\Sigma}_\mathcal{P}^{-1}(\boldsymbol{\xi}_j - \boldsymbol{\mu}_\mathcal{P})\right], \tag{20}$$

which has mean vector

$$\boldsymbol{\mu}_\mathcal{P} = (\mu_\theta, \mu_\tau), \tag{21}$$

and covariance matrix

$$\boldsymbol{\Sigma}_\mathcal{P} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \tag{22}$$

In addition to (8), for the extended model to be identifiable, we impose

$$\mu_\theta = 0, \tag{23}$$

and

$$\sigma_\theta = 1. \tag{24}$$

In the remainder of this paper, it is not only assumed that all items have been calibrated but also that the free parameters in (22) have been estimated with enough precision to treat them as known during adaptive testing. If the Bayesian procedure in van der Linden (2007) is used, their estimation involves only an additional step in a Gibbs sampler.

Response-Based Posterior Predictive Density of RTs

In (9), the empirical population distribution of τ was chosen as the prior distribution of τ_j. However, the population model in (19, 20) now gives us the joint distribution of τ and θ. Its density function can be factorized as

$$f(\boldsymbol{\xi}_j) = f(\tau \mid \theta) f(\theta). \tag{25}$$

We will use this factorization in three different steps: First, the second factor is used as the empirical prior distribution of θ_j to derive its posterior density given the responses. Second, the posterior density of θ_j is combined with the first factor to derive the predictive posterior density of τ_j given the responses. Third, the result is combined with the RT model in (7) to derive the predictive distribution of the RT on an item given the responses and RTs on the remaining items. The first two steps mirror those in van der Linden (2008), where the reverse factorization was chosen to use RTs times as collateral information for item selection in adaptive testing.

The posterior density of θ_j given $\mathbf{u}_j$ is

$$f(\theta_j \mid \mathbf{u}_j) = \left[\prod_{i=1}^{n} f(u_{ij} \mid \theta_j) \right] f(\theta_j)\, d\theta_j, \tag{26}$$

where $f(\theta_j)$ is the second factor of (25) taken as the empirical prior of θ_j.

Combining (26) with the first factor gives the posterior predictive density of τ_j given $\mathbf{u}_j$,

$$f(\tau_j \mid \mathbf{u}_j) = \int f(\tau_j \mid \theta_j) f(\theta_j \mid \mathbf{u}_j)\, d\theta_j. \tag{27}$$

This density can be used to replace $f(\tau_j)$ as the empirical prior of τ_j in (10). Using the local independence assumption $f(t_{ij} \mid t_j, \tau_j) = f(t_{ij} \mid \tau_j)$ in the hierarchical model, the result is

$$\begin{aligned} f(\tau_j \mid \mathbf{t}_j, \mathbf{u}_j) &\propto f(\mathbf{t}_j \mid \tau_j) f(\tau_j \mid \mathbf{u}_j) \\ &= \left[\prod_{i=1}^{n} f(t_{ij} \mid \tau_j) \right] f(\tau_j \mid \mathbf{u}_j). \end{aligned} \tag{28}$$

In order to check RTs for aberrances, this posterior distribution of τ_j given $(\mathbf{t}_j, \mathbf{u}_j)$ can be used to calculate the posterior predictive density of the log-time on an arbitrary item i given the responses and RTs on the other items as

$$f\left(\tilde{t}^*_{ij} \mid \mathbf{t}^*_{j\backslash i}, \mathbf{u}_{j\backslash i}\right) = \int \left(\tilde{t}^*_{ij} \mid \tau_j\right) f\left(\tau_j \mid \mathbf{t}^*_{j\backslash i}, \mathbf{u}_{j\backslash i}\right) d\tau_j. \tag{29}$$

The posterior distribution of θ given $\mathbf{u}_j$ in (26) is not normal. Hence, unlike (15), the posterior predictive density in (29) is not normal either. Although it could easily be calculated using appropriate quadrature, a closed-form approximation of the density will be more convenient. The next section shows how to find a normal approximation.

Normal Approximation

The approximation is based on the fact that distribution of θ given $\mathbf{u}_j$ is known to converge strongly to normality under regularity conditions which are not restrictive for adaptive testing (Chang & Stout, 1993). Also, because of the adaptation, the convergence can be expected to be much faster than for standard paper-and-pencil testing. The closed-form approximation to (29) is based on this argument, which reminds one of Owen's (1969, 1975) approximate procedure for Bayesian item selection in adaptive testing. Therefore, we assume $f(\theta_j \mid \mathbf{u}_j)$ in (26) to be approximately normal with posterior mean $\mu_{\theta_j|\mathbf{u}_j}$ and variance $\sigma^2_{\theta_j|\mathbf{u}_j}$.

From (19–22), using the identifiability constraints, $f(\tau_j \mid \theta_j)$ is also normal with mean

$$\mu_{\tau_j|\theta_j} = \sigma_{\theta\tau}\theta_j \tag{30}$$

and variance

$$\sigma^2_{\tau_j|\theta_j} = \sigma^2_\tau - \sigma^2_{\theta\tau}. \tag{31}$$

Hence, the posterior predictive density $f(\tau_j \mid \mathbf{u}_j)$ in (27) is normal with mean

$$\mu_{\tau_j|\mathbf{u}_j} = \sigma_{\theta\tau}\mu_{\theta_j|\mathbf{u}_j}, \tag{32}$$

and variance

$$\sigma^2_{\tau_j|\mathbf{u}_j} = \sigma^2_\tau - \sigma^2_{\theta\tau} + \sigma^2_{\theta\tau}\sigma^2_{\theta_j|\mathbf{u}_j}. \tag{33}$$

Normality of both $f(\tau_j \mid \mathbf{u}_j)$ and the log-times implies a normal form of the posterior density $f(\tau_j \mid \mathbf{t}^*_j, \mathbf{u}_j)$ in (28) with mean

$$\mu_{\tau_j|\mathbf{t}^*_j,\mathbf{u}_j} = \frac{\frac{\sigma_{\theta\tau}}{\sigma^2_\tau - \sigma^2_{\theta\tau} + \sigma^2_{\theta\tau}\sigma^2_{\theta_j|\mathbf{u}_j}}\mu_{\theta_j|\mathbf{u}_j} + \sum_{i=1}^n \alpha_i^2(\beta_i - \ln t_{ij})}{(\sigma^2_\tau - \sigma^2_{\theta\tau} + \sigma^2_{\theta\tau}\sigma^2_{\theta_j|\mathbf{u}_j})^{-1} + \sum_{i=1}^n \alpha_i^2}, \tag{34}$$

and variance

$$\sigma^2_{\tau_j|\mathbf{t}^*_j,\mathbf{u}_j} = \left((\sigma^2_\tau - \sigma^2_{\theta\tau} + \sigma^2_{\theta\tau}\sigma^2_{\theta_j|\mathbf{u}_j})^{-1} + \sum_{i=1}^n \alpha_i^2\right)^{-1}. \tag{35}$$

Consequently, using the same argument as for (16) and (17), the posterior predictive density $f(\tilde{t}^*_{ij} \mid \mathbf{t}^*_{j\setminus i}, \mathbf{u}_{j\setminus i})$ in (29) is normal with mean

$$\mu_{\tilde{t}^*_{ij}|\mathbf{t}^*_{j\setminus i},\mathbf{u}_{j\setminus i}} = \beta_i - \frac{\frac{\sigma_{\theta\tau}}{\sigma^2_\tau - \sigma^2_{\theta\tau} + \sigma^2_{\theta\tau}\sigma^2_{\theta_j|\mathbf{u}_{j\setminus i}}}\mu_{\theta_j|\mathbf{u}_{j\setminus i}} + \sum_{k\neq i}^n \alpha_k^2(\beta_k - \ln t_{kj})}{(\sigma^2_\tau - \sigma^2_{\theta\tau} + \sigma^2_{\theta\tau}\sigma^2_{\theta_j|\mathbf{u}_{j\setminus i}})^{-1} + \sum_{k\neq i}^n \alpha_k^2}, \tag{36}$$

and variance

$$\sigma^2_{\tilde{t}^*_{ij}|\mathbf{t}^*_{j\setminus i},\mathbf{u}_{j\setminus i}} = \alpha_i^{-2} + \left((\sigma^2_\tau - \sigma^2_{\theta\tau} + \sigma^2_{\theta\tau}\sigma^2_{\theta_j|\mathbf{u}_{j\setminus i}})^{-1} + \sum_{k\neq i}^n \alpha_i^2\right)^{-1}. \tag{37}$$

This alternative density also has a mean and variance with known parameters. The posterior mean $\mu_{\theta_j|\mathbf{u}_{j\setminus i}}$ is the EAP estimate of θ from the responses $\mathbf{u}_{j\setminus i}$; its calculation is straightforward. The same holds for the posterior variance $\sigma^2_{\theta_j|\mathbf{u}_{j\setminus i}}$.

Aberrant RT Patterns

The focus in this paper is especially on aberrances that suggest cheating on adaptive tests. Two possible forms of cheating are: (i) memorization of items during the test with the intent to share them with other candidates and (ii) preknowledge of some of the items in the pool.

Both types of behavior are likely to lead to typical patterns of RTs and responses. Attempts at memorization may reveal themselves as patterns of RTs that do not reflect the time intensities of the items. In addition, because in adaptive testing it is necessary to enter a response to make the next item appear, the responses may be largely at random. Preknowledge of items may result in a combination of unlikely RTs and correct responses. The actual RTs required for test takers to memorize an item or recognize an item and check if it has not been changed in some minor way depend on the nature of the test. Examples of such RTs in empirical studies are reported in one of the later sections of this paper.

It is also possible to follow items over time and aggregate information on the RTs by the individual test takers. For example, as long as an item has not been compromised, its empirical distribution of the standardized log RTs over the test takers will be close to $N(0, 1)$. Item compromise may reveal itself as a tendency to a lower mean after some time.

The posterior predictive distributions above allow us to assign probability statement to patterns of RTs. One possible application is to identify a subset of problematic items and then calculate their predictive densities from the data on the remaining items in the test. A more explorative—but statistically less sound—approach is to calculate the density for each item in the test given the data on all other items.

Suppose we are interested in aberrances that manifest themselves by lower RTs (e.g., hypothesis of preknowledge). For each item, we can calculate the probability of a predicted RT lower than the observed RT. For item i and test taker j, the probability of a log-time lower than an observation t_{ij}^* is

$$\pi_{ij}^{l} = \int^{t_{ij}^*} f\big(\tilde{t}_{ij}^* \mid \mathbf{t}_{j\backslash i}^*\big) d\tilde{t}_{ij}^* \quad \text{or} \quad \int^{t_{ij}^*} f\big(\tilde{t}_{ij}^* \mid \mathbf{t}_{j\backslash i}^*, \mathbf{u}_{j\backslash i}\big) d\tilde{t}_{ij}^*. \tag{38}$$

Since the densities are normal with known means and variances, these probabilities are easy to calculate.

Likewise, if the check is on erratic RTs (e.g., hypothesis of memorization), we should use the probability in (38) when the observed log-time t_{ij}^* is smaller than the mean of the predictive distribution but replace it with

$$\pi_{ij}^{u} = \int_{t_{ij}^*} f\big(\tilde{t}_{ij}^* \mid \mathbf{t}_{j\backslash i}^*\big) d\tilde{t}_{ij}^* \quad \text{or} \quad \int_{t_{ij}^*} f\big(\tilde{t}_{ij}^* \mid \mathbf{t}_{j\backslash i}^*, \mathbf{u}_{j\backslash i}\big) d\tilde{t}_{ij}^*, \tag{39}$$

when it is larger.

Because of local independence, the probability of observing an entire pattern of RTs on a set of k items with a more extreme than observed RT on any of the items is equal to

$$1 - \prod_{i=1}^{k} (1 - \pi_{ij}), \tag{40}$$

with the appropriate choice of π_{ij}^{l} and π_{ij}^{u} substituted for π_{ij}.

A frequentist alternative to (40) would be to use Fisher's (1925) proposal for combining individual tests of significance. For example, for left-sided tests for the individual items, it holds

that

$$-2\sum_{i=1}^{k}\ln\pi_{ij}^{l}\sim\chi_{2k}^{2}. \tag{41}$$

A combined test of the hypothesis of no aberrances consists of checking $-2\sum\ln\pi_{ij}^{l}$ against a critical value under the chi-square distribution with $2k$ degrees of freedom.

Although such procedures may seem attractive, it should be observed that the probability in (40) goes to one with increasing numbers of items and that a test based on (41) will become sensitive to irrelevant differences under the same condition. These problems are inherent in statistical hypothesis testing. A graphical procedure may therefore be more informative. For example, we could standardize the observed log RTs using the mean and variance of the predictive distribution and plot them for each test taker or for a given item over a periodically sampled set of test takers. A band at the conventional values of $+1.96$ and -1.96 helps us to identify unexpected RTs and interpret their pattern.

The RT-based approach to the detection of aberrances in adaptive testing was used in a case study for a data set from the Quantitative section of the Graduate Management Admission Test® (GMAT®). In this study, we fitted the hierarchical model to the data, evaluated its fit, and used the graphical procedure to analyze the RT patterns for the test takers for possible cheating. The results are presented in the next section.

Analysis of real-world RT patterns do not shed any light on the statistical power of the procedures. If the type of cheating can be hypothesized as a specific change in the speed parameter of the test taker, the power of the procedure can be calculated from an explicit formula. However, it seems difficult to substantiate such hypotheses. We therefore conducted a simulation study for the same item pool from the GMAT®, in which we used our knowledge of the operational features of the test to formulate specific types of cheating directly for the RTs. This study is reported in the subsequent section.

Case Study

The GMAT® is an adaptive test from item pools calibrated under the 3PL model in (1). The data set used in this study was for the Quantitative section of the test and consisted of the responses and RT of 3,999 test takers on 431 items. The set of items did not constitute the full item pool for the section; when we estimated the time parameters for the items, only operational items with at least 100 responses were used.

Calibration and Model Validation

The model was fitted with fully Bayesian estimation of the parameters using the Gibbs sampler described by van der Linden (2007). The prior distribution for the free hyperparameters in (19–22) was the standard normal-inverse-Wishart with a low-informative choice for its parameters. The 3PNO model in (18) was chosen because it allows for the type of data augmentation discussed by Albert (1992) that is exploited in the sampler. The posterior distribution of the parameters was stable after a few hundred iterations and all analyses in this example were based on 5,000 additional draws.

As all items were previous operational items selected to fit the 3PL model and this model is numerically indistinguishable from the 3PNO model, the interest in this study was especially in the fit of the lognormal model for the RTs as well as the bivariate normal population model for the person parameters required for the procedures for identifying aberrances proposed in this paper.

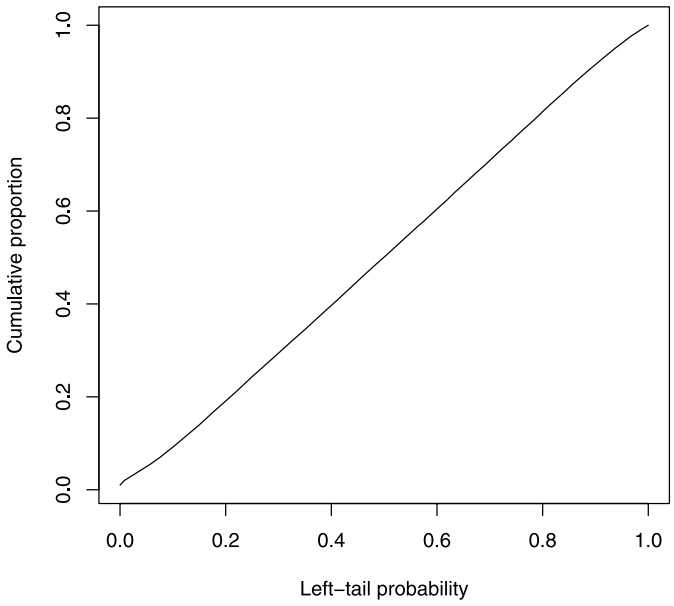

FIGURE 1.
Cumulative distribution of the left-tail posterior predictive probabilities of the observed log RTs over all person-item combinations in the data set. The closer the curve to the identity line, the better the overall fit of the RT model.

Figure 1 shows the plot of the cumulative distribution of the left-tail probabilities of the observed log-times under their posterior predictive distributions for all item-person combinations in the sample. The total number of combinations was 98,942. The probabilities were calculated directly from the draws from the posterior distribution of $(\tau_j, \alpha_i, \beta_i)$ in the Gibbs sampler. If the overall fit of the model is perfect, these cumulative plots follow the identity line (cf. the probability integral transformation theorem; e.g., Casella & Berger, 2002, p. 54). The plot reveals slight negative and positive deviations from the identity line for the left-tail and right-tail probabilities, respectively. These deviations hint at a tendency to overrepresent the longer RTs at the expense of the smaller RTs. But the deviations were small enough to warrant a satisfactory overall fit of the RT model and justify its use for the current purpose.

To give an impression of the fit of the individual items, we repeated the analyses with the left-tail probabilities aggregated over the sample of persons for each item. To restrict the amount of material, Figure 2 shows the results for every 40th item in the data set only. The plots for all other items were entirely comparable. Each of the item plots is now based on a much smaller sample of persons. The range was 101–639 persons per item. For these numbers it is impossible to get stable continuous distributions. The plots therefore show much more sample variation than in Figure 1. Generally, the plots did not show systematic deviations from the identity line, and we concluded that the items appeared to have a satisfactory fit to the model. Nevertheless, it seems safe not to overinterpret individual residuals that are in the tails of their distribution (for instance, rely blindly on probabilities of significance) but to look for *patterns* of residuals across items and test takers that point at potentially suspicious behavior.

Table 1 contains some descriptive information on the item-parameters estimates and their standard errors of estimation. We used the means of the posterior distributions of the parameter as estimates (EAP estimates) and their standard deviations as the standard errors of estimation. The most important information is on the estimates of the time intensity parameters β_i, especially on their range, which was 2.68–6.70. (Remember that these parameters are on a logarithmic scale; we have to take exponents to express them in seconds.) This finding supports our earlier claim

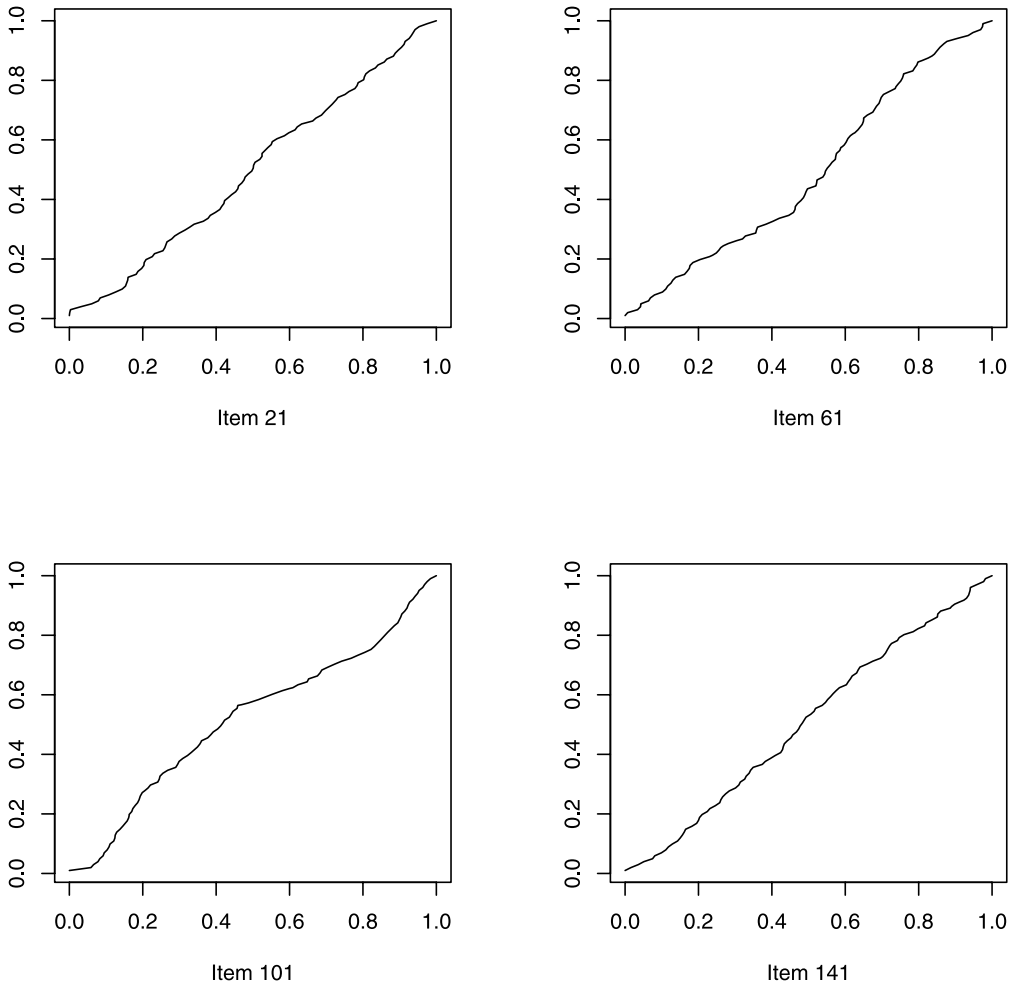

FIGURE 2.
Cumulative distribution of the left-tail posterior predictive probabilities of the observed log RTs over all persons for every 40th item in the data set. The closer the curve of an item to the identity line, the better the fit of the RT model.

TABLE 1.
Some information on the posterior means and SDs of the item parameters in the RT model.

	Alpha		Beta	
	Mean	SD	Mean	SD
Mean	1.30	0.07	4.53	0.06
Min	0.60	0.02	2.68	0.03
Max	2.43	0.14	6.70	0.15
SD	0.37	0.03	0.82	0.02

that items in a typical CAT program may differ greatly in their time intensity and it will generally be difficult for test takers to simulate realistic RTs during the test.

The estimated correlation between τ and θ in the population model was -0.25, indicating that the more able students tended to work slightly more slowly. A plot of the estimates of τ_j

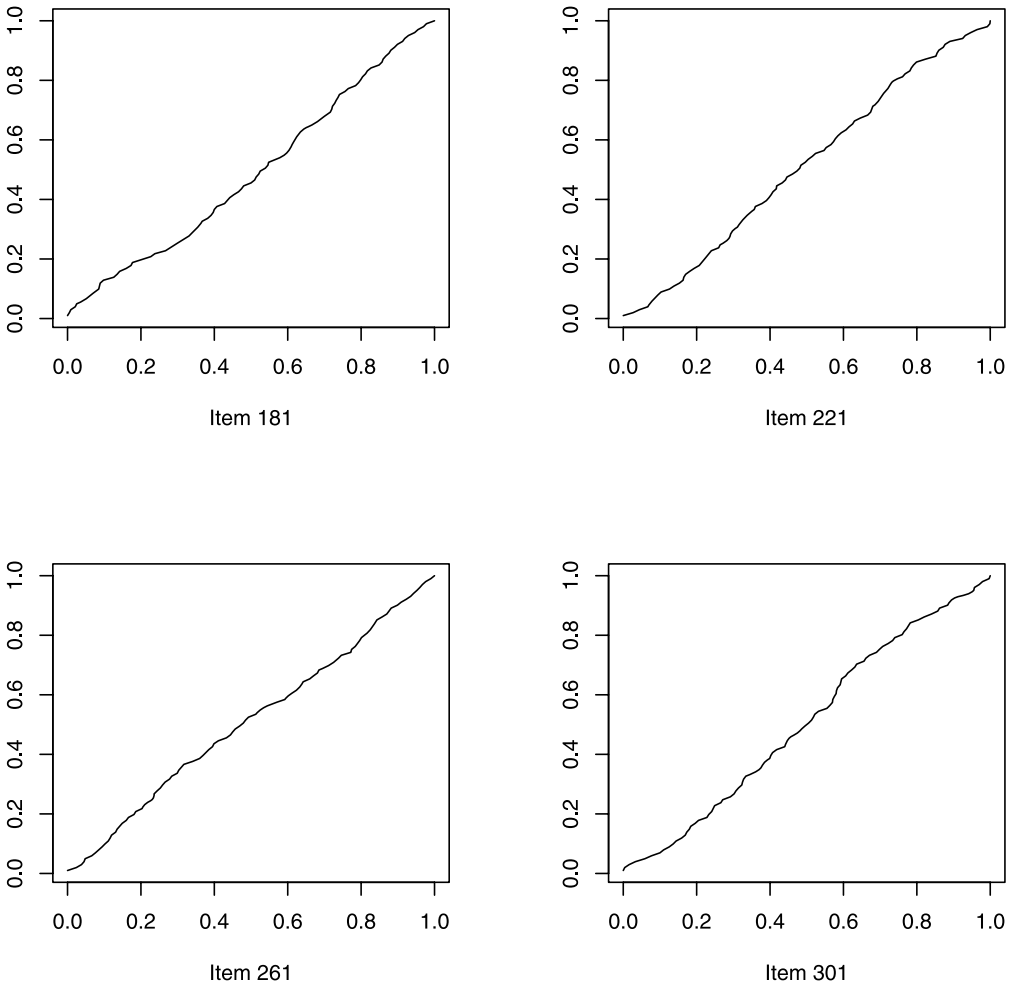

FIGURE 2.
(*Continued.*)

and θ_j showed near normality of θ but a clear tendency to bimodality for the distribution of τ. We therefore concluded that a mixture of distributions as the population model for τ would have been more realistic and that the estimate of the correlation between θ and τ based on a single bivariate distribution was biased too much to be used in an empirical study. Hence, in the empirical analyses in the next section, only the predictive posterior distribution of the RTs in (15–17) was used.

Some Results

First, we discuss a few examples of regular and aberrant RT patterns in the data set. Then, we present an example of how to combine checks on RTs with other information on the test performances to flag test takers suspected of cheating. Finally, we show how individual RTs can be combined to detect possible compromised test items.

Aberrant RT Patterns In order to identify aberrant RT patterns, each log RT was standardized using the predicted mean and standard deviation in (16) and (17) given the RTs on all other

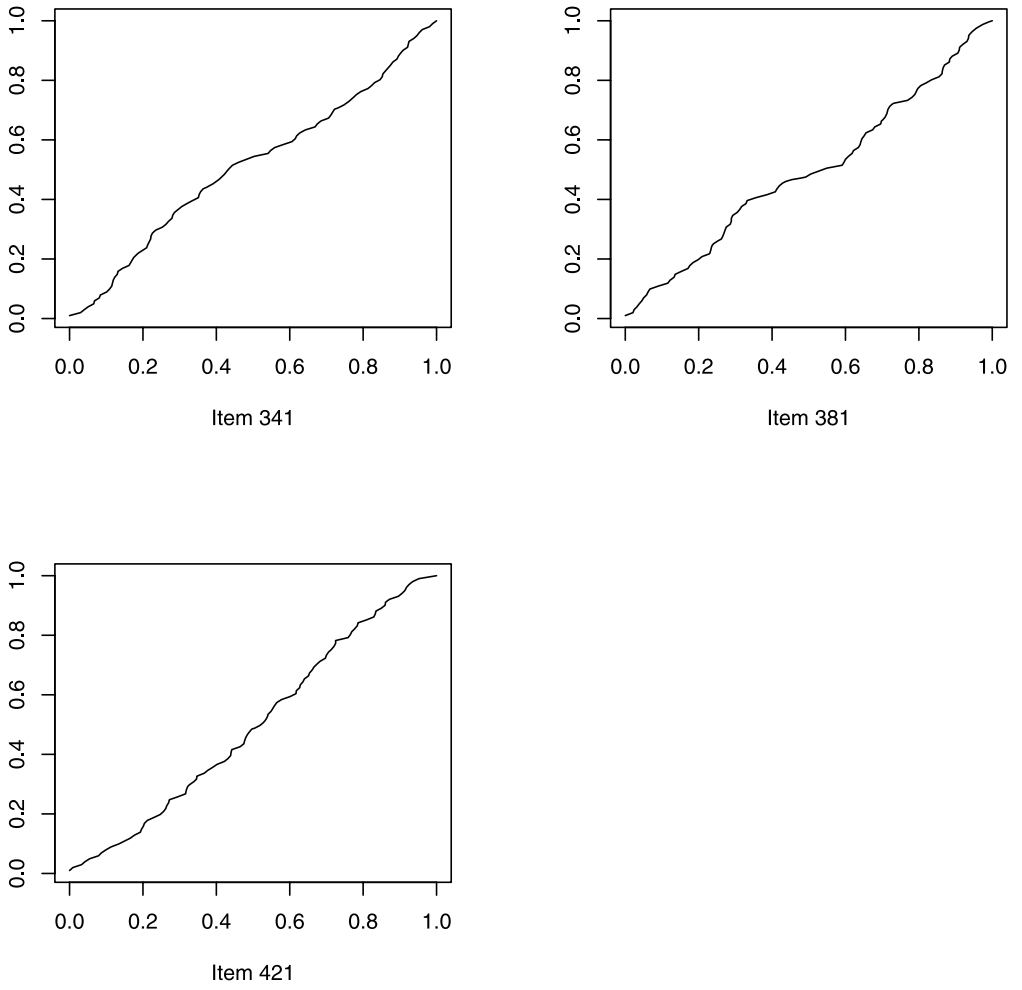

FIGURE 2.
(*Continued.*)

items by the same test taker. An RT to an item was flagged as aberrant when its standardized residual was larger than 1.96 or smaller than −1.96.

Although the length of the test was fixed, the total number of items per test taker in this study varied since, as already explained, only operational items with at least 100 responses were included in the study. The analyses included a total of 110,562 RTs across the test takers and items. Out of the total, 4,350 (3.93%) were flagged as aberrant. More specifically, 2,487 (2.25%) were flagged because more time than expected was spent on the item (positive aberrance) and 1,863 (1.69%) were flagged because less time than expected was spent (negative aberrance). These percentages are close to the nominal significance level of the test, which means that the test takers generally behaved quite regularly according to the RT model and that cheating or item compromise was certainly not a structural problem for the test.

Figure 3 shows the RT pattern of a test taker with 15 flagged RTs. The horizontal axis represents the items in the order of delivery in the test. For each item, the shaded column represents the (standardized) residual RT. It is obvious that several of the RTs were flagged because the test taker spent a large amount of time (about 52 minutes) on items 1 through 18 (especially items 2,

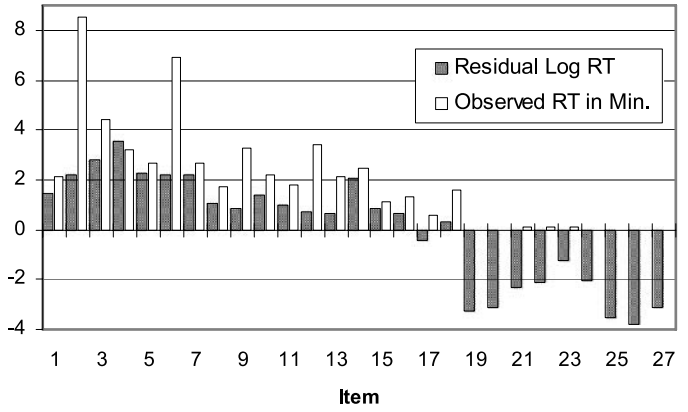

FIGURE 3.
Example of an aberrant RT pattern.

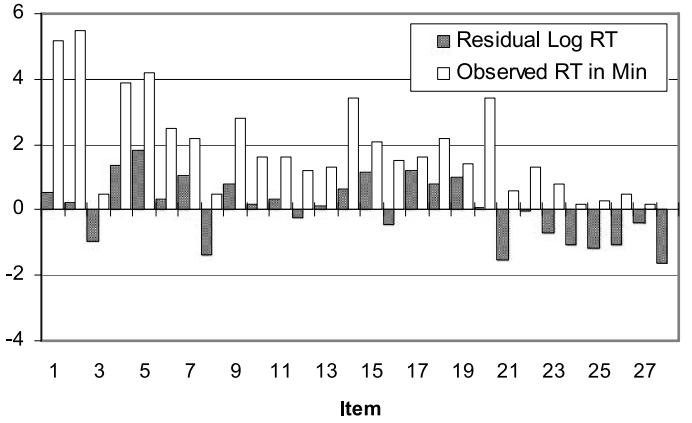

FIGURE 4.
Example of a regular RT pattern.

3, and 6) and then rushed through items 19 to 27. We concluded that these aberrances were the result of poor time management by the test taker rather than some type of cheating.

Figure 3 also shows the actual RTs in minutes for this test taker (white columns). It is important to note that the residual and observed RTs show no direct relationship to each other. For example, for item 2, the residual and observed RTs were 2.17 and 8.5 minutes, respectively, but they were 2.28 and 2.7 minutes for item 5. Thus, the former yielded a much larger observed RT than the latter but their sizes after standardization were approximately equal. The difference is explained by the fact that item 2 was more time intensive than item 5 ($\widehat{\beta}_2 = 5.05$ and $\widehat{\beta}_5 = 4.41$). Without this information, we might have flagged the RT on item 2 as extreme and on item 5 as regular whereas both were in the upper tail of their predictive distribution.

Figure 4 shows the data for a test taker with no flagged RTs. The pattern of the observed RTs suggests the same type of time management as in Figure 3 (larger RTs earlier in the test but shorter RTs toward the end). But the residual times revealed that this test taker maintained a more regular speed during the test.

The two examples illustrate how a predictive procedure for the RT model in (7) is able to discriminate between patterns of observed RTs for an adaptive test that resemble each other at a more superficial level. Because of this, it can be used to identify aberrant test behavior. However,

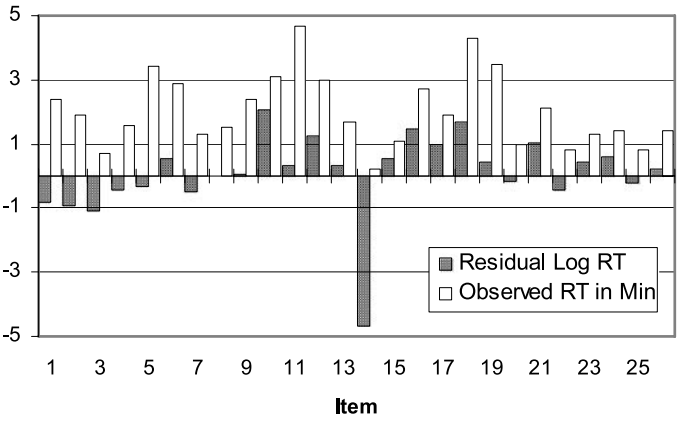

FIGURE 5.
Example of a response with possible preknowledge of the item.

the first example also shows that it would be incorrect to classify aberrancies automatically as cheating. A large-scale screening, as in this case study, should always be followed by more careful analyses of the entire patterns of RTs and responses as well as of possible more direct evidence for aberrant persons and items. The following example illustrates this more cautious approach.

Suspicious RT Patterns As noted earlier, indicators of preknowledge of an item by a test taker are:

1. an answer that is correct;
2. a large negative residual RT; and
3. a low probability of success on the item (i.e., a low estimated ability relative to the difficulty of the item).

We used these indicators to check the records of all test takers in the data set and then reviewed those that had one or more items flagged by all indicators. Figure 5 describes one of the cases we found. This test taker spent a regular amount of time on most of the items. But the residual RT for item 14 was -4.66, which was extremely negative. The estimated time intensity of this item was $\beta = 4.49$ on the logscale, which means that the median time for a test taker of average speed (i.e., $\tau = 0$) to answer the item was 88.9 seconds on the regular time scale. Also, the item was relatively difficult for this test taker since its b parameter was approximately half a standard deviation higher than his/her θ estimate. Nevertheless, miraculously, this test taker spent only 12.3 seconds on the item and answered it correctly!

Compromised Items As a first example of combining information at the level of the individual RTs to detect item compromise, we simply checked how often the RTs on the same items were flagged and if there was any pattern among the size of the residuals. Item 14 in Figure 5 had a total of four test takers with exactly the same type of record: a very short time spent on the item (approximately 12–13 seconds), a correct answer, and estimated θs much lower than the b parameter. These four events may indicate that the item was known to a few individual test takers. However, a larger number would have been necessary to conclude that there was an actual security threat.

Another useful way to diagnose an item pool for item compromise is by monitoring the RTs over time. Generally, it will take some time for an item to get compromised after its launch.

TABLE 2.
Means and SDs of standardized residual RTs in three different testing periods.

Period	Mean	SD	No. of persons
Early	0.56	0.91	134
Middle	0.61	0.95	132
Late	0.40	1.10	82

Before this happens, the residuals will be regular but thereafter they can be expected to show a trend toward a decrease. In order to illustrate the procedure, we divided a testing window into three periods of approximately equal length. For each period, the means and standard deviations of the residual RTs on all items were calculated.

A typical result is given in Table 2. The mean residual RTs of each period were very close. (One-way ANOVA did not show any significant difference between the periods.) In fact, the analyses did not yield any item that could have been compromised. Although the result was negative, Table 2 suggests how to use residual RTs to monitor items for security breaches when the test is operational.

An alternative to the ANOVA approach would be to use CUSUM-based statistics on RTs on individual items collected over time. The same type of statistics were used by van Krimpen-Stoop and Meijer (2001) to analyze the response vectors of individual test takers for person-misfit in adaptive testing. Because CUSUM statistics originated in the analysis of time series in industrial statistics for quality control purposes, it will be quite natural to apply them for checking whether series of RTs on individual items in adaptive testing remain within quality bounds.

Conclusion The GMAT® did not seem to have suffered much from cheating. The overall number of RTs flagged as aberrant was close to the level of significance. The patterns of aberrant RTs that were detected seemed to point more toward other types of behavior, such as problems with time management during the test, than toward cheating. To make sure that the results were not due to a lack of power of the method, a study was conducted using simulated data (next section). The item pool in this study was the same pool for the GMAT®, and the specific types of cheating that were simulated were suggested by our knowledge of the operational features of the test.

Power Study

When the check is on aberrances that can be represented as a change of speed by the test taker on a subset of the items, it is straightforward to calculate the power of the test. Suppose that τ_0 is the speed by the test takers on the items for which regular RTs are produced and we want to test the null hypothesis $\tau = \tau_0$ against one of the two alternatives $\tau < \tau_0$ or $\tau > \tau_0$ for another subset of the items. If the hypothesis is tested against the right-sided alternative, the critical value of the α-level test is the log-time $\ln t_\alpha^*$ for which

$$1 - F\left(\ln t_\alpha^*; \tau_0, \alpha_i, \beta_i\right) = \alpha, \tag{42}$$

where $F(\cdot)$ is the cumulative distribution function of the normal distribution with the parameterization in (7). The power function of the test is

$$1 - F\left(\ln t_\alpha^*; \tau, \alpha_i, \beta_i\right), \quad \tau > \tau_0. \tag{43}$$

TABLE 3.
Proportion of test takers flagged for preknowledge as a function of the time spent on the known item (δ).

# Regular items (m)	$\alpha = 0.05$			$\alpha = 0.01$		
	$\delta = 10$	$\delta = 20$	$\delta = 30$	$\delta = 10$	$\delta = 20$	$\delta = 30$
2	0.26	0.03	0.00	0.02	0.00	0.00
4	0.85	0.50	0.26	0.64	0.25	0.08
6	0.77	0.36	0.15	0.54	0.14	0.03
8	0.84	0.45	0.22	0.60	0.22	0.05
10	0.87	0.48	0.26	0.68	0.24	0.08
20	0.87	0.45	0.33	0.70	0.34	0.16
30	0.83	0.36	0.31	0.60	0.31	0.17

To correct for estimation error in τ_0, the test would have to be based on the predictive distribution with the mean in (16) and variance in (17) instead of the model distribution. However, the null distribution would still be normal.

This power analysis does not have much practical value because the alternative hypothesis is not yet specific. In fact, it seems even impossible to equate specific types of aberrant behavior with certain values of τ. To get a better impression of the power of the procedure in applied settings, we therefore conducted a simulation study in which we simulated "realistic" cases of cheating on adaptive tests by directly manipulating the RTs on subsets of the items instead of the speed parameter, and evaluated how frequently the procedure would detect the aberrancy. Both types of cheating addressed in this paper were simulated: preknowledge of some of the items and attempts to memorize a portion of the test. The tests were simulated from the same item pool from the Quantitative section of the Graduate Management Admission Test® (GMAT®) as in the case study in the preceding section, and we used our knowledge of the nature of the test to choose the test takers' simulated strategies for cheating.

Item Preknowledge The test takers were simulated to behave regularly except for one item, which they were assumed to already know. The RTs on the regular items were drawn from (7). For the item assumed to be known, the RTs were equal to $\delta = 10$, 20, or 30 seconds. These three alternatives were considered to be in a realistic range for a test taker to recognize an item, check if it is has not been changed in some minor way, and enter the response.

The size of the set of regular items from which the residual RTs were estimated varied between $m = 2$ and 30. The conditions with the smaller values of m were included to see where the procedure breaks down when the RTs on many items in the test are suspicious.

For each combination of δ and m, the number of replications was equal to 800. The proportion of times the residual RTs on the problematic item were flagged are given in Table 3. The results show that the method did not function well for $m = 2$ (although it still flagged a remarkable proportion of the test takers for $\delta = 10$), but can be recommended for larger numbers of regular items. Generally, the hit rates decreased with δ: for $\delta = 10$, they were quite high; for $\delta = 20$ we considered them as acceptable; but for $\delta = 30$ seconds they were low. It should be noted, however, that these rates were only for a single item. If a test taker knows more than one item, or when we aggregate the residual RTs to detect compromised items, the power of the procedure would go up immediately.

Memorizing Items The scenario we chose was that of test takers answering the items seriously for some time to reach an estimated level ability close to their true ability. Thereafter, they were assumed to use their time only to memorize a fixed number of items. The remaining time was divided equally between these items. Also, during memorization the test takers were

TABLE 4.
Distribution of number of items flagged per test taker for memorization as a function of the number of items memorized (κ).

# Flagged items	$\alpha = 0.05$			$\alpha = 0.01$		
	$\kappa = 5$	$\kappa = 10$	$\kappa = 15$	$\kappa = 5$	$\kappa = 10$	$\kappa = 15$
0	0.07	0.34	0.51	0.22	0.59	0.71
1	0.26	0.19	0.20	0.33	0.20	0.20
2	0.35	0.13	0.13	0.30	0.10	0.08
3	0.20	0.12	0.05	0.12	0.06	0.01
4	0.10	0.11	0.04	0.03	0.03	0.00
5	0.04	0.07	0.03	0.01	0.01	0.00
6		0.03	0.03		0.00	0.00
7		0.11	0.01		0.00	0.00
8		0.00	0.00		0.00	0.00
9		0.00	0.00		0.00	0.00
10		0.00	0.00		0.00	0.00
11			0.00			0.00
12			0.00			0.00
13						0.00
14						0.00
15						0.00

not assumed to actually solve the items. Their responses were therefore produced randomly with a probability equal to 0.20 (five-choice items).

The Quantitative section of the GMAT® has a time limit of two hours. The two periods were therefore taken to be 30 and 90 minutes. The number of memorized items was equal to $\kappa = 5$, 10, or 15. Again, for each value of κ, the number of replications was equal to 800.

Table 4 shows the distributions of the numbers of items flagged per test taker. Clearly, for the majority of the test takers, one or more items were detected. Less favorable results were obtained only for test takers trying to memorize $\kappa = 15$ items. But this number may have been chosen too large; at least it is considerably larger than Miller's (1956) magic number, seven plus or minus two, which is generally taken to be an upper bound on the number of items in human information processing.

Concluding Remark

Checking the response behavior of test takers for possible aberrances is one of the prime methods of quality control in the testing industry. As argued in this paper, when the test is adaptive, traditional person-fit checks based on the response model lose their power. But a useful alternative can be found in checks based on the RTs under a model with item and person parameters for their distribution. Although the simulation study showed quite satisfactory power for the proposed checks, we repeat our earlier warnings against mechanical use of them. There are other explanations for aberrant RTs besides cheating, and blind conclusions from statistically significant RT residuals could easily be wrong. One of the explanations identified in the case study was bad time management by test takers during the test. The only safeguards against wrong explanations are careful qualitative analyses of the entire pattern of RT residuals for flagged test takers or items and corroborating evidence in the form of reported irregularities during the testing session.

The focus in this paper was on checks of RTs to detect cheating but the number of potential applications is much larger. The same type of checks can be used to supplement traditional item

analysis. For example, a hypothesis of differential item functioning (DIF) is confirmed more convincingly if the distributions of residual RTs for the focal and reference groups differ systematically. Traditional response-based analyses easily confound DIF with multidimensionality and distributional differences between focal and reference groups on nuisance abilities. But systematic differences between residual RTs hint at the fact that the item actually invokes different processes in the two groups. Other examples are checks on possible differential speededness of the test (e.g., between selections of items given to more and less able test takers), flaws in the design of the test (e.g. ambiguous instructions), or fatigue toward the end of it. For all such applications, we expect the information in the RTs to be at least as powerful as that in the responses.

References

Albert, J.H. (1992). Bayesian estimation of normal-ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*, 261–269.

Bradlow, E.T., Weiss, R.E., & Cho, M. (1998). Bayesian detection of outliers in computerized adaptive tests. *Journal of the American Statistical Association*, *93*, 910–919.

Casella, G., & Berger, R.L. (2002). *Statistical inference* (2nd ed.). Pacific Grove: Duxbury.

Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52.

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

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

Glas, C.A.W., & Meijer, R.R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*, 217–233.

Johnson, V.E., & Albert, J.H. (1999). *Ordinal data modeling*. New York: Springer.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Meijer, R.R., & Sijtsma, K. (1995). Detection of aberrant item response patterns: A review of recent developments. *Applied Measurement in Education*, *8*, 261–272.

Meijer, R.R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.

Miller, G.A. (1956). The magic number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Owen, R.J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ, Educational Testing Service.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351–356.

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

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

van der Linden, W.J. (2008). Using response times for item selection in adaptive tests. *Journal of Educational and Behavioral Statistics, 33*. In press.

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

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

van Krimpen-Stoop, E.M.L.A., & Meijer, R.R. (2001). CUSUM-based person fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*, 199–218.

Manuscript received 18 DEC 2006
Final version received 21 SEP 2007
Published Online Date: 9 JAN 2008