LSAC RESEARCH REPORT SERIES

- **Testing for Aberrant Behavior in Response Time Modeling**

Sukaesi Marianti
Jean-Paul Fox
Marianna Avetisyan
Bernard P. Veldkamp
University of Twente, Enschede, the Netherlands

- **Law School Admission Council**
Research Report 14-02
March 2014

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation that provides unique, state-of-the-art products and services to ease the admission process for law schools and their applicants worldwide. Currently, 218 law schools in the United States, Canada, and Australia are members of the Council and benefit from LSAC's services. All law schools approved by the American Bar Association are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also members. Accredited law schools outside of the United States and Canada are eligible for membership at the discretion of the LSAC Board of Trustees; Melbourne Law School, the University of Melbourne is the first LSAC-member law school outside of North America. Many nonmember schools also take advantage of LSAC's services. For all users, LSAC strives to provide the highest quality of products, services, and customer service.

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

This study is published and distributed by LSAC.

Table of Contents

Executive Summary

Many standardized tests are now administered via computer rather than paper-and-pencil format. In a computer-based testing environment, it is possible to record not only the test taker's response to each question (item), but also the amount of time spent by the test taker in considering and answering each item. Response times (RTs) provide information not only about the test taker's ability and response behavior but also about item and test characteristics. The current study focuses on the use of RTs to detect aberrant test-taker responses. An example of such aberrance is a correct answer with a short response time on a difficult question. Such aberrance may be displayed when a test taker or test takers have preknowledge of the items. Another example is rapid guessing, wherein the test taker displays unusually short response times for a series of items. When rapid guessing occurs at the end of a timed test, it often indicates that the test taker has run out of time before completing the test.

In the current study, a model for detecting various types of aberrant RT patterns is proposed and evaluated. In simulation studies, the model was successful in identifying aberrant response patterns. Further investigations are required to analyze flagged patterns more thoroughly, possibly by applying additional information.

Introduction

Many standardized tests rely on computer-based testing (CBT) because of its operational advantages. CBT reduces the costs involved in the logistics of transporting the paper forms to various test locations, and it provides many opportunities to increase test security. CBT also benefits the candidates. It enables testing organizations to record scores more easily and to provide feedback and test results immediately. In computerized adaptive testing (CAT), a special type of CBT, the difficulty level of the items is adapted to the response pattern of the candidate; this advantage also holds for multistage testing. Multimedia tools can even be included, and automated scoring of open-answer questions and essays can be supported. CBT can be used for online classes and practice tests.

An advantage of CBT is that it offers the possibility of collecting response time (RT) information on items. RTs provide information not only about test takers' ability and response behavior but also about item and test characteristics. With the collection of RTs, the assessment process can be further improved in terms of precision, fairness, and minimizing costs.

The information that RTs reveal can be used for routine operations in testing, such as item calibration, test design, detection of cheating, and adaptive item selection. In general, once RTs are available, they could be used both for test design and diagnostic purposes.

In the 1990s, psychometric analysis of RTs to improve the quality of assessment measurements was suggested by Masters and Keeves (1999), Weiss and Schleisman (1999), Schnipke and Scrams (1997, 1999a, 1999b, 2002), Schnipke and Pashley (1997, March), Hornke (1997, 2000), and Bergstrom, Gershon, and Lunz (1994, April), among others. Test takers' speed became an important component influencing response accuracy, and suggestions were made to develop test models including test takers' response time. Further research in this area was

done by Wainer, Dorans, Flaugher, Green, and Mislevy (2000), Wainer and Eignor (2000), Schnipke and Scrams (1997, 1999b), (Hornke, 2005), Jansen (2007), and Jansen and Glas (2001).

In general, two types of test models can be recognized: (a) separate RT models that only describe the distribution of the RTs given characteristics of the test taker and test items, and (b) test models that describe the distribution of RTs as well as responses. With respect to the second one, Thissen (1983) defined the timed testing modeling framework, where item response theory (IRT) models are extended to account for speed and accuracy within one model. However, these types of models have been criticized because problems with confounding were likely to occur.

Van der Linden (2006, 2007) advocated the first type of modeling and proposed a latent variable modeling approach for both processes. He defined a model for the RTs and a separate model for the response accuracy, where latent variables (person level and item level) explain the variation in observations and define conditional independence within and between the two processes. The RT process is characterized by RT observations, speed of working, and labor intensity, which are in a comparable way defined in the RT process by observations of success, ability, and item difficulty. This framework has many advantages and recognizes two distinct processes: It adheres to the multilevel data structure, and it allows one to identify within, between, and cross level relationships.

The item characteristics of the RT distribution can be recognized by a time-intensity parameter and a time-discrimination parameter. The time-intensity parameter reflects the average time needed for completing the item, and the time-discrimination parameter characterizes the sensitivity of the item for different speed levels of the test takers. As analogues to item parameters of the IRT model, the RT parameters can be applied for diagnostic purposes and for test assembly. The sum of the time intensities is a measure of the total test time, whereas the RT discriminations can be used to control for variable speed, to identify regions where items measure accurately, and to define the contributions of each item to the total speed measurement.

This modeling framework provides many features, and a log-normal RT distribution can be applied to model response behavior in educational research (van der Linden, 2006, 2007). Unfortunately, not all respondents behave according to the model. Besides random fluctuation, aberrant response behavior also occurs due to, for example, item preknowledge, cheating, or test speededness. Focusing on RTs might have several advantages in revealing various types of aberrant behavior. RTs are continuous and therefore more informative and easier to evaluate statistically. One other advantage, especially for CAT, is that RTs are insensitive to the design effect in adaptive testing, since the selection of test items does not influence the distribution of RTs in any systematic way. RT models are defined to separate speed from time intensities; this makes it possible to compare the pattern of time intensities with the pattern of RTs.

Different types of aberrant behavior have been introduced and studied. Van der Linden and Guo (2008) introduce two types of aberrant response behavior: (a) attempts at memorization, which might reveal themselves by random RTs; and (b) item preknowledge, which might result in an unusual combination of a correct response and RTs. RT patterns are considered to be suspicious when an answer is correct and the RT is relatively small while the probability of success on the item is low. Schnipke and Scramms (1997) studied rapid guessing, where part of the items show unusually small RTs. Bolt, Cohen, and Wollack (2002) focused on test

speededness toward the end of a test. For some respondents who run out of time, one might observe unexpected small RTs during the last part of the test.

For all of these types, it holds that response behavior either conforms to an RT model representing normal behavior or it does not (i.e., it is aberrant behavior). We propose using a log-normal RT model to deal with various types of aberrant behavior. Based on this log-normal RT model, a general approach to detect aberrant response behavior can be considered in which checks can be used to flag respondents or items that need further consideration. Van der Linden and Guo (2008) already indicated that test takers may show aberrant behavior for several reasons, and it would be wrong to jump to conclusions. Checks could be used routinely in order to flag test takers or items that may need further consideration or to support observations by proctors or other evidence.

In this report, a log-normal RT model will be introduced first; we developed an R-package to estimate this model. In a simulation study, we compare our new R-package with WinBUGS and an existing R-package for the case of log-normal RT models to check the performance of the new software. Then we test the log-normal RT approach by simulating various types of aberrant response behavior and studying the power to detect the aberrancies. We evaluate the results and present several directions for future research.

RT Modeling

Van der Linden (2006) proposed a log-normal distribution for RTs on test items. In this model, the logarithm of the RTs is assumed to be normally distributed. The model is briefly discussed since it is used to derive new procedures for detecting aberrant RTs. The proposed tests for detecting aberrant response behavior are based on log-normally distributed RTs. The log-normal density for the distribution of RTs is specified by the mean and the variance. The mean term represents the expected time the test taker needs to answer the item, and the variance term represents the variance of measurement errors.

In log-normal RT models, each test taker is assumed to have a constant working speed during the test. Let

$p = 1, ..., N$	be an index for the test takers
$i = 1, ..., I$	be an index for the items
ζ_p	denote the working speed of test taker p
λ_i	denote the time intensity of item i
T_{ip}	denote the RT of test taker p to item i

Subsequently, the logarithm of T_{ip} has mean $\mu_{pi} = \lambda_i - \zeta_p$ (see also, van der Linden, 2006). The lower the time intensity of an item, the lower the mean. In the same way, the faster a test taker operates, the lower the mean. This model can be extended by introducing a time-discrimination parameter to allow variability in the effect of increasing the working speed to reduce the mean. Let

ϕ_i denote the time discrimination of item i.

With this extension, the mean is parameterized as $\mu_{pi} = \phi_i\left(\lambda_i - \zeta_p\right)$, such that the reduction in RT by operating faster is not constant over items. The higher the time discrimination of an item, the higher the reduction in the mean when operating faster. For example, when a test taker operates a constant C faster, the mean is represented by $\mu_{pi} = \phi_i\left(\lambda_i - \left(\zeta_p + C\right)\right) = \phi_i\left(\lambda_i - \zeta_p\right) - \phi_i C$, such that the item-specific reduction is defined by $\phi_i C$.

Observed RTs will deviate from the mean term (i.e., expected times), and the errors are considered to be measurement errors. The response behavior of test takers can deviate slightly during the test, leading to different error variances over items. Test takers might stretch their legs or might be distracted for a moment, and so on. These measurement errors are assumed to be independently distributed given the operating speed of the test taker, the time intensities, and time discriminations. Let

σ_i^2 denote the error variance of item i.

In the log-normal RT model, σ_i^2 could vary over items. The errors are expected to be less homogenous, when, for example, items are not clearly written, when items are positioned at the end of a time-intensive test, or when test conditions vary during an examination and influence the performance of the test takers (e.g., noise nuisance).

With this mean and variance, the log-normal model for the distribution of T_{ip} can be represented by

$$p\left(t_{ip}\left|\zeta_p, \lambda_i, \phi_i, \sigma_i^2\right.\right) = \frac{1}{\sqrt{2\pi\sigma_i^2 t_{ip}}} \exp\left[-\frac{1}{2\sigma_i^2}\left(\ln t_{ip} - \phi_i\left(\lambda_i - \zeta_p\right)\right)^2\right]. \quad (1)$$

We will refer to the time-intensity and time-discrimination parameters as the item's *time characteristics* in order to stress their connection with the definition of *item characteristics* (i.e., item difficulty and item discrimination) in IRT.

This parameterization and its interpretation deviate slightly from the model of van der Linden (2006), since a time-discrimination parameter is introduced. In the model of van der Linden, the representation of working speed can be directly related to a physical meaning of speed, since differences in RTs are due to differences in either working speed or time intensities.

With the introduction of a time-discrimination parameter, differences in working speed do not lead to a homogeneous change in RTs over items. A differential effect of speed on RTs is allowed, which is represented by the time-discrimination

parameters. The idea is that working speed is modeled by a latent variable representing the ability to work with a certain level of speed. Furthermore, it is assumed that this construct comprehends different dimensions of working speed. Depending on the item, this construct can relate, for example, to a physical capability, a cognitive capability, or a combination of both. For example, consider two items with the same time intensity, where one item concerns writing a small amount of text and the other doing analytical thinking. Differences between the RTs of two test takers can be explained by the fact that one works faster. However, differences in RTs between test takers are not necessarily homogenous over items. One item appeals to the capability of writing faster and the other to thinking or reasoning faster, and it is unlikely that both dimensions influence RTs in a common way.

Identification

The observed times have a natural scale, which is defined by a unit of measurement (e.g., seconds). However, the metric of the scale is undefined due to our parameterization. First, the mean of the scale is undefined due to the speed and time intensity parameters in the mean, $\lambda_i - \zeta_p$. To identify the mean of the scale, the mean speed of the test takers is set to zero. Second, the variance of the scale is also undefined due to the time-discrimination parameter and the population variance of the speed parameter. The variance of the scale is identified by setting the product of discriminations equal to one. It is also possible to fix the population variance of speed (e.g., to set it equal to one).

A Bayesian Log-Normal RT Model

Prior distributions can be specified for the parameters of the distribution of RTs in Equation (1). The population of test takers is assumed to be normally distributed such that

$$\zeta_p \sim N\left(\mu_\zeta, \sigma_\zeta^2\right) \tag{2}$$

where $\mu_\zeta = 0$ to identify the mean of the scale. An inverse gamma hyper prior is specified for the variance parameter. The prior distribution for the time intensity and discrimination parameters give support to partial pooling of information across items. When the RT information for a specific time intensity leads to an unstable estimate, RT information from other items is used to obtain a more stable estimate. This partial pooling of information within a test is based on the principle that the items in the test have an average time intensity and an average time discrimination. Each individual item can have characteristics that deviate from the average depending on the information in the RTs.

Partial pooling of information is also defined for item-specific parameters. The time intensity and discrimination parameter in Equation (1) relate to the same item, and are allowed to correlate. A bivariate normal distribution is used to describe the relationship between the parameters,

$$\begin{pmatrix} \phi_i \\ \lambda_i \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_\lambda \\ \mu_\phi \end{bmatrix}, \begin{bmatrix} \sigma_\phi^2 & \rho \\ \rho & \sigma_\lambda^2 \end{bmatrix} \right). \tag{3}$$

The mean time intensity of the test is denoted by μ_λ and represents the average time it takes to complete the test. The mean time discrimination is denoted by μ_ϕ and represents the effect of reducing the mean test time when increasing the working speed. The common covariance parameter ρ across items represents for each item the linear relation between both parameters. For example, items that are more time intensive might discriminate better between individual performances. The hyper priors will be normal distributions for the mean parameters and an inverse Wishart distribution for the covariance matrix. Although the modeling approach supports partial pooling of information, the hyper priors are specified in such a way that partial pooling of information is diminished and the within-item RT information is the most important source of information to estimate the time-intensity and time-discrimination parameters.

The measurement error variance parameters σ_i^2 are assumed to be independently inverse gamma distributed. The errors of a test taker are assumed to be independently distributed given the speed of working and the item's time characteristics.

The specification of the log-normal model leads to the following random effects model to model the logarithm of RTs:

$$\left.\begin{aligned} \log T_{ip} &= \phi_i\left(\lambda_i - \zeta_p\right) + \varepsilon_{ip} \\ \phi_i &= \mu_\phi + r_{1i} \\ \lambda_i &= \mu_\lambda + r_{2i} \\ \zeta_p &= \mu_\zeta + e_p \end{aligned}\right\} \begin{aligned} &\text{Modeling time observations} \\ &\text{Item specification} \\ &\text{Test-taker specification,} \end{aligned} \tag{4}$$

where three levels can be recognized. At Level 1, time observations are modeled using a normal distribution for the logarithm of RTs and three random effects to address the influence of the test taker's speed of working and of the item's time characteristics. The test item's properties are modeled as multivariate normally distributed random effects and are modeled at the level of items. Finally, the test taker's working speed is modeled at the level of persons.

The Estimation Procedure for Log-Normal RT Models

The model parameters and the test statistics are computed using a Bayesian estimation procedure. With the Markov chain Monte Carlo (MCMC) method referred to as Gibbs sampling, samples are obtained from the posterior distributions of the model parameters. Gibbs sampling is an iterative estimation method where, in each

iteration, a sample is obtained from the full conditional distributions of the model parameters. Methods for sampling directly from the posterior distributions have been described by Gelman, Carlin, Stern, and Rubin (2004) and Gelfand and Smith (1990). To apply Gibbs sampling, the full conditional distributions of the model parameters need to be specified. For the log-normal model, the technical details of the estimation method are given by Klein Entink, Fox, and van der Linden (2009), van der Linden (2007) , and Fox, Klein Entink, and van der Linden (2007).

Test for Aberrant RT Patterns

One of the most popular fit statistics in person-fit analysis is the l_z statistic (Drasgow, Levine, & Williams, 1985), which is the standardized likelihood-based person-fit statistic l_o of Levine and Rubin (1979). This person-fit statistic has received much attention in educational measurement. Studies have shown that it almost always outperforms other person-fit statistics, and it is commonly accepted as one of the most powerful person-fit statistics to detect aberrant response patterns. With this in mind, we propose a person-fit statistic for aberrant response behavior for RT patterns.

The log-likelihood of the RTs is used to evaluate the fit of a response pattern consisting of RTs. We will use $t_{ip}^{*} = \ln\left(t_{ip}\right)$ to denote the logarithm of the RT of test taker p on item i. Our likelihood-based person-fit statistic for RTs requires knowledge of the density of the response pattern. This follows directly from the normal model for the logarithm of RTs; that is,

$$l_o\left(\zeta_p, \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2; \mathbf{t}_p^{*}\right) = -2\log p\left(\mathbf{t}_p^{*} \middle| \zeta_p, \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2\right) = \sum_{i=1}^{I} l_{oi}. \tag{5}$$

The l_0 statistic can be evaluated over all items in the test, but it is also possible to consider a subpart of the test. A large value of the statistic indicates a misfit, since it represents a departure of the RT observations from expected RTs under the model. The posterior distribution of the statistic can be used to examine whether a pattern of observed RTs is extreme under the model.

Given the model specification in Equation (1), the probability density function of a response pattern is represented by the product of individual RTs. The probability density of response pattern $\mathbf{t}_p^{*} = \left(t_{1p}^{*}, ..., t_{Ip}^{*}\right)$ is given by

$$\begin{aligned} -2\log p\left(\mathbf{t}_p^{*} \middle| \zeta_p, \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2\right) &= -2\sum_{i=1}^{I} \log p\left(t_{ip}^{*} \middle| \zeta_p, \lambda_i, \phi_i, \sigma_i^2\right) \\ &= \sum_{i=1}^{I}\left(\left(\frac{t_{ip}^{*} - \mu_{ip}}{\sigma_i}\right)^2 + \log\left(2\pi\sigma_i^2\right)\right) \\ &= \sum_{i=1}^{I}\left(Z_{ip}^2 + \log\left(2\pi\sigma_i^2\right)\right), \end{aligned} \tag{6}$$

where Z_{ip} is standard normally distributed, since it represents the standardized error of the normally distributed logarithm of RT.

The test statistic l_0 depends on various model parameters. It is possible to compute statistic values given values for the model parameters or given posterior distributions of the model parameters. In the last case, the posterior mean statistic value is estimated by integrating over the posterior distributions of the model parameters.

In the person-fit literature, the standardized person-fit statistic, which is usually denoted as l_z, receives much attention because it has an asymptotic standard normal distribution. Drasgow et al. (1985) showed that for tests longer than 80 items, the l_z statistic is approximately normally distributed. Other studies (e.g., Meijer & Nering, 1997; Molenaar & Hoijtink, 1990) showed that for shorter tests the distribution of the test statistic was negatively skewed, violating the assumption of symmetry of the normal distribution. Snijders (2001) proposed an adjustment to standardize the l_z statistic, thereby accounting for the fact that parameter estimates are used to compute the statistic value.

The standardized version of the l_0^t for RTs, denoted as l_z^t, requires an expression for the expected value and the variance of the statistic in Equation (5). In Appendix A, it is shown that the conditional expectation is given by

$$E\left[l_o\left(\zeta_p,\boldsymbol{\lambda},\phi,\boldsymbol{\sigma}^2\right)\middle|\mathbf{t}_p^*,\zeta_p,\boldsymbol{\lambda},\phi,\boldsymbol{\sigma}^2\right]=\sum_i\left(1+\ln\left(2\pi\sigma_i^2\right)\right) \tag{7}$$

and the variance is given by

$$Var\left[l_o\left(\zeta_p,\boldsymbol{\lambda},\phi,\boldsymbol{\sigma}^2\right)\middle|\mathbf{t}_p^*,\zeta_p,\boldsymbol{\lambda},\phi,\boldsymbol{\sigma}^2\right]=2I, \tag{8}$$

where I is the total number of test items. Subsequently, the standardized version, l_z^t, is derived by standardizing the statistic in Equation (5) using the terms in Equations (7) and (8). It follows that

$$l_z^t\left(\zeta_p,\boldsymbol{\lambda},\phi,\boldsymbol{\sigma}^2;\mathbf{t}_p^*\right)=\frac{\left(\sum_{i=1}^{I}Z_{ip}^2+\log\left(2\pi\sigma_i^2\right)\right)-\left(\sum_{i=1}^{I}1+\log\left(2\pi\sigma_i^2\right)\right)}{\sqrt{2I}}=\frac{\sum_{i=1}^{I}Z_{ip}^2-I}{\sqrt{2I}}. \tag{9}$$

To ease the notation, the statistic's dependency on the model parameters is ignored, leading to $l_z^t\left(\zeta_p,\boldsymbol{\lambda},\phi,\boldsymbol{\sigma}^2;\mathbf{t}_p^*\right)=l_z^t\left(\mathbf{t}_p^*\right)$. In the computation of l_z^t, model parameters are assumed to be known, or the posterior expectation is taken over the unknown model parameters.

The Null Distribution

In order to come to a person-fit statistic, the null distribution of l_z^t has to be derived. First we introduce some notation. The logarithm of RTs is represented by a random variable T_{pi}^*, which is normally distributed, where the observed values are denoted by t_{pi}^*. An RT pattern of test taker p is represented by $\mathbf{T}_p^*$. Given this

notation, the null distribution of $l_z^t\left(\mathbf{T}_p^*\right)$ can be derived in three different ways, resulting in three different person-fit statistics for $\mathbf{T}_p^*$ under the log-normal model.

First, the null distribution of the $l_z^t\left(\mathbf{T}_p^*\right)$ follows from the fact that the errors Z_{ip} (see Equation (9)) are standard normally distributed. The sum of squared errors, which are standard normally distributed, is known to be chi-squared distributed with I degrees of freedom. Box, Hunter, and Hunter (1978, p. 118) showed that a chi-squared distributed variable T with I degrees of freedom, the distribution of $\left(T-I\right)/\sqrt{2I}$ is approximately standard normal. Therefore, the null distribution of the $l_z^t\left(\mathbf{T}_p^*\right)$ can be considered to be approximately standard normal.

Second, an exact null distribution can be obtained by considering a nonstandardized version of the $l_z^t\left(\mathbf{T}_p^*\right)$, which is the sum of squared standardized errors:

$$l^t\left(\mathbf{T}_p^*\right)=\sum_{i=1}^{I}Z_{ip}^2\,. \tag{10}$$

This sum of squared errors, which are standard normally distributed, is known to be chi-squared distributed with I degrees of freedom.

Third, the Wilson–Hilferty transformation can be used to standardize the person-fit statistic $l^t\left(\mathbf{T}_p^*\right)$ in such a way that it is approximately standard normal distributed. This leads to

$$l_s^t\left(\mathbf{T}_p^*\right)=\frac{\left(\sum_{i=1}^{I}Z_{ip}^2/I\right)^{1/3}-\left(1-2/(9I)\right)}{\sqrt{2/(9I)}}\,. \tag{11}$$

Summarized, three person-fit statistics for RTs are considered that differ in the way the null distribution is derived (Table 1).

TABLE 1
Person-fit statistics for RT data under the lognormal model

Statistic	Type Null Distribution	Exact or Approximation	Probability of Significance
l_z^t	Normal	Approximation	$P\left(l_z^t\left(\mathbf{T}_p^*\right)>C\right)\approx\Phi\left(l_z^t\left(\mathbf{T}_p^*\right)>C\right)$
l^t	Chi-squared	Exact	$P\left(l^t\left(\mathbf{T}_p^*\right)>C\right)=P\left(\chi_I^2>C\right)$
l_s^t	Normal	Approximation	$P\left(l_s^t\left(\mathbf{T}_p^*\right)>C\right)\approx\Phi\left(l_s^t\left(\mathbf{T}_p^*\right)>C\right)$

Bayesian Testing of Aberrant RT Patterns

To assess the extremeness of the pattern of RTs, the posterior probability can be computed such that the estimated statistic value, say $l^t\left(\mathbf{t}_p^*\right)$, is greater than a certain threshold C. This threshold C defines the boundary of a critical region, which is the set of values for which the null hypothesis is rejected if the observed statistic value is located in the critical region. The critical value C can be determined from the null distribution; that is,

$$P\left(l^t\left(\mathbf{T}_p^*\right) > C\right) = P\left(\chi_I^2 > C\right) = \alpha, \tag{12}$$

since the null distribution is a chi-squared distribution with I degrees of freedom, where α is the level of significance. When the observed statistic value, $l^t\left(\mathbf{t}_p^*\right)$, is larger than C, the RT pattern will be flagged.

Given the sampled parameter values in each MCMC iteration, it is also possible to compute a function of the model parameters (e.g., a probability statement). To illustrate this, consider the tail-area event as specified in Table 1. Given sampled values from the posterior distribution of the model parameters, the posterior probability can be computed as

$$\begin{aligned} P\left(l^t\left(\mathbf{T}_p^*\right) > C\right) &\approx \sum_{m=1}^{M} P\left(l^t\left(\mathbf{T}_p^*\right) > C\right) p\left(\zeta_p^{(m)}, \boldsymbol{\lambda}^{(m)} \middle| \mathbf{t}_p^*\right) \\ &= \sum_{m=1}^{M} \Phi\left(l^t\left(\mathbf{T}_p^*\right) > C\right) p\left(\zeta_p^{(m)}, \boldsymbol{\lambda}^{(m)} \middle| \mathbf{t}_p^*\right) \end{aligned}, \tag{13}$$

where m denotes the MCMC iteration number. The terms to standardize the test statistic depend on the model parameters. In each iteration, the test statistic is computed using the sampled model parameters, and the average posterior probability approximates the marginal posterior probability of obtaining a test statistic larger than a criterion value C. The uncertainty in the parameters is taken into account in the computation of the posterior probability.

Note that in Equation (13), draws are used from the posterior distribution to compute the marginal posterior probability. When using posterior draws, the posterior distribution of the model parameters might be distorted by RT data that do not fit the model. An alternative would be to use draws from the prior distribution. Then, most often a much larger number of draws will be required to obtain an accurate estimate of the marginal posterior probability. Moreover, a misspecification of the priors might lead to a biased posterior probability estimate.

Besides testing whether a pattern of RTs is in a critical area defined by a threshold C, it is also possible to quantify the extremeness of the observed RT pattern by computing the right-tail area probability under the model. This right-tail probability represents the posterior probability of observing a more extreme statistic value under the model. The estimated statistic value is constructed from the sum of squared errors, and an extreme statistic value indicates that the RT pattern is not likely to be produced under the log-normal model. When the posterior probability is close to zero, it can be concluded that the pattern is unlikely under the posited log-normal model and the pattern is considered to be aberrant given the observed data.

Note that the decision to flag an RT pattern as extreme depends on the size of the statistic value but also on the posterior uncertainty. When the distribution of the test statistic is rather flat, it is less likely to conclude with high posterior probability that an RT pattern is extreme in comparison to a highly peaked distribution. Given accurate information, a more definitive decision can be made about the extremeness of the RT pattern.

Dealing With Nuisance Parameters

The test statistic depends on the model parameters, which follows directly from the definition of Z_{pi}. To compute the marginal posterior probability of observing a more extreme value than the observed one, an integration needs to be performed over all model parameters:

$$P\left(l^t\left(\mathbf{T}_p^*\right) > C\right) = \int_{\boldsymbol{\lambda}} \int_{\zeta_p} P\left(l^t\left(\mathbf{T}_p^*\right) > C \middle| \zeta_p, \boldsymbol{\lambda}\right) p\left(\zeta_p, \boldsymbol{\lambda}\right) d\zeta_p d\boldsymbol{\lambda} . \quad (14)$$

The marginal posterior probability is obtained by integrating over the model parameters. MCMC can be used to obtain draws from the posterior distribution of the model parameters. For each draw, the probability that the computed statistic value is above a threshold value C can be computed. The average posterior probability over MCMC iterations is an estimate of the marginal posterior probability as specified in Equation (12).

In Equation (14), the distribution of the statistic is assumed to be known, and the assessment of the test statistic is known as a *prior predictive test* (Box, 1980). Given prior distributions for the model parameters, it is assessed how extreme the observed statistic value is. Prior predictive testing is usually preferred, since the double use of the data in posterior predictive assessment is known to bias the distribution of estimated tail-area probabilities. When the data are used to estimate the model parameters and to assess the distribution of the test statistic, the tail-area probabilities are often not uniformly distributed. This makes it more difficult to interpret the estimated probabilities. In the prior predictive assessment approach, as stated in (12) and (14), the double use of the data is avoided and the tail-area probability estimates can be correctly interpreted.

To assess whether an RT pattern is extreme, a classification is made based on the value of the test statistic. The exact or an accurate approximation of the null distribution of the statistic is known but depends on unknown model parameters. When the statistic is computed by plugging in parameter estimates, the corresponding tail-area probability might be biased. Therefore, the probability that an RT pattern will be flagged as extreme is evaluated in each MCMC iteration. An accurate decision can be made in each MCMC iteration given values for the model parameters. Let random variable F_p take on a value of one when the RT pattern of test taker p is flagged, or a value of zero otherwise. Thus,

$$F_p = \begin{cases} 1 \text{ if } & P\left(l^t\left(\mathbf{T}_p^*\right) > l^t\left(\mathbf{t}_p^*\right)\right) < \alpha \\ 0 \text{ if } & P\left(l^t\left(\mathbf{T}_p^*\right) > l^t\left(\mathbf{t}_p^*\right)\right) \geq \alpha. \end{cases} \quad (15)$$

Interest is focused on the marginal posterior probability that the RT pattern of test taker p will be flagged, which is computed by

$$\begin{aligned} P\left(F_p = 1 \middle| \mathbf{t}_p^*\right) &= \int_{\boldsymbol{\lambda}} \int_{\zeta_p} I\left(F_p = 1 \middle| \mathbf{t}_p^*, \zeta_p, \boldsymbol{\lambda}\right) p\left(\zeta_p, \boldsymbol{\lambda}\right) d\zeta_p d\boldsymbol{\lambda} \\ &\approx \sum_{m=1}^{M} I\left(F_p^{(m)} = 1 \middle| \zeta_p^{(m)}, \boldsymbol{\lambda}^{(m)}\right) / M, \end{aligned} \tag{16}$$

where in MCMC iteration m, $F_p^{(m)} = 1$ when $P\left(\chi^2 > l^t\left(\mathbf{t}_p^*\right) \middle| \zeta_p^{(m)}, \boldsymbol{\lambda}^{(m)}\right) < \alpha$. So, the probability that a pattern will be flagged is evaluated in each iteration. The average probability over iterations approximates the marginal probability of a flagged RT pattern. The extremeness of the pattern can be quantified, since the posterior probability in Equation (16) states how likely it is that the pattern will be flagged under the log-normal model. It can be decided that only patterns that have a posterior probability of .95 or higher will be flagged under the model. This reduces the probability of making a Type I error, since the posterior probability quantifies the extremeness of each RT pattern, instead of classifying the pattern based on a chosen significance level α.

The posterior probability of the extremeness of the response pattern in Equation (14) can also be defined from a posterior predictive perspective. Given the model parameters, the posterior probability of the test statistic is evaluated given its sampling distribution. When the distribution of the statistic is unknown, the posterior predictive distribution of the data can be used to assess the distribution of the test statistic. In that case, the extremeness of the estimated test statistic is evaluated using the posterior predictive distribution of the data. This is shown by

$$P\left(l^t\left(\mathbf{T}_p^{*rep}\right) > l^t\left(\mathbf{t}_p^*\right)\right) = \int_{\mathbf{t}_p^{*rep}} P\left(l^t\left(\mathbf{T}_p^{*rep}\right) > l^t\left(\mathbf{t}_p^*\right)\right) p\left(\mathbf{T}_p^{*rep} \middle| \zeta_p, \boldsymbol{\lambda}\right) d\mathbf{T}_p^{*rep}, \tag{17}$$

where $\mathbf{T}_p^{*rep}$ denotes the replicated data under the model and the left-hand side of Equation (17) represents the posterior predictive probability of observing a statistic value that is greater than the statistic value based on the observed data.

Posterior predictive tests have been suggested in many different applications to evaluate the fit of models. Rubin (1984) and Gelman, Meng, and Stern (1996), among others, advocated the use of posterior predictive assessment to evaluate the compatibility of the model to the data. Box (1980) recommended the use of the marginal predictive distribution of the data to evaluate the fit of the model, which is also known as prior predictive assessment.

Van der Linden and Guo (2008) also suggested using a predictive distribution to evaluate RTs. In their approach, a cross-validation predictive residual distribution is used to evaluate the extremeness of the remaining RTs. Furthermore, the predicted response is compared to the observed response in an adaptive test application. The normal distribution of the logarithm of RTs is used to calculate the power of identifying aberrant RTs. They also used a less accurate method, which was based on classifying estimated residuals. Ignoring the uncertainty of the estimates, RTs were flagged as aberrant when the corresponding estimated standardized residuals were larger than 1.96 or smaller than −1.96. In the present approach, the posterior

uncertainty is taken into account, and RTs are flagged to be aberrant with a certain posterior probability.

A Mixture Log-Normal RT Model

Although more accurate decisions can be made when the model parameters are known, the data are often needed to estimate the model parameters and to evaluate the fit of the model. When the data contain a relatively large percentage of RT patterns not fitting the model, these patterns will bias the parameter estimates. For example, in the log-normal model in Equation (1) it is assumed that the working speed of test takers is normally distributed and is constant throughout the entire test. When test takers show aberrant response behavior, working with a relatively higher speed at the end of the test (compared to the other part of the test) will lead to underestimating the time intensities of the last test items. These test items appear to take less time due to the behavior of the test takers.

To improve the quality of the parameter estimates, flagged RT patterns should not be used in the test calibration. Therefore, a two-component mixture distribution can be defined in which one class defines the set of aberrant RT patterns and the other class the set of nonaberrant RT patterns. The object is to use all RT patterns classified as nonaberrant but to use only a (significance-level) percentage of randomly selected RT patterns of the class of aberrant patterns for item parameter estimation. The flagged patterns located in the class of aberrancies (or misfits), which are not selected, are not used in the estimation of the item parameters to avoid a distortion in item parameter estimates.

This is how the procedure works. In each iteration of the MCMC method, each RT pattern is evaluated according to our test statistic in Equation (10). When the RT pattern is flagged as aberrant with a posterior probability of .975 or higher, the RT pattern is assigned to the class of misfits. However, this class of flagged RT patterns also includes patterns that are extreme but still fit the log-normal model. That is, tail-area events are excluded, which are needed to obtain a correct distribution of the RTs. Therefore, in each MCMC iteration, the set of patterns that are not excluded and 2.5% (of the total sample) of randomly selected aberrant RT patterns are used to estimate the model parameters. In the case where 20% of the data consist of RT patterns flagged as aberrant, 2.5% will be used in the estimation procedure. Since it is unknown which of the 20% of RT patterns represents correct tail events under the log-normal model, in each iteration of the estimation method a new set of 2.5% RT patterns is sampled from the class of aberrant RTs. Let A_0 denote the class of nonaberrant RT patterns and A_1 the class of aberrant RT patterns. The RT patterns are assumed to follow a log-normal distribution according to Equation (1) for patterns assigned to class A_0. The distribution of the patterns assigned to class A_1 are not specified, although this option will be useful when a specific type of aberrant response behavior is considered. According to the specifications of the mixture distribution, the distribution of the data is given by

$$p\left(\mathbf{t}_p \middle| \zeta_p, \boldsymbol{\lambda}, \phi, \boldsymbol{\sigma}^2\right) = p\left(\mathbf{t}_p \middle| \boldsymbol{\zeta}_p, \boldsymbol{\lambda}, \phi, \boldsymbol{\sigma}^2, A_0\right) P\left(\mathbf{t}_p \in A_0\right) + p\left(\mathbf{t}_p \middle| A_1\right) P\left(\mathbf{t}_p \in A_1\right). \qquad (18)$$

The posterior probability of assigning an RT pattern to class A_1 equals

$$P\left(\mathbf{t}_p \in A_1\right) = P\left(l^t\left(\mathbf{T}_p^*\right) > l^t\left(\mathbf{t}_p^*\right)\right), \quad (19)$$

which is the posterior probability of obtaining an even greater test statistic (of a more extreme pattern) than the estimated statistic for the observed RT pattern under the log-normal model. When this posterior probability is less than .025, the decision is made to assign the pattern to the class. Classes A_0 and A_1 are complementary, which means that each pattern is assigned to one of the classes.

This mixture modeling approach enables the computation of posterior classification probabilities of RT patterns. Furthermore, a set of RTs will be defined that are not extreme under the model with a posterior probability of at least .975, which can be used to estimate the model parameters. It will be shown that in the MCMC estimation method, both events can be estimated simultaneously.

Results

Through simulation studies, the performance of the person-fit statistics for RT patterns is evaluated. Study 1 concerns a parameter recovery study to evaluate the performance of the estimation method. A comparison is made between three different programs for estimating the model parameters. In Study 2, the detection rates of the l^t statistic are evaluated for different types of misfit. Different conditions are simulated to investigate the performance of the statistic.

Study 1: Investigation of Parameter Recovery

The MCMC method for estimating the model parameters of the log-normal model was implemented in R and is referred to as LNRT. This general program for RT modeling and checks on aberrances can be compared with two other programs, when considering the log-normal model specification of van der Linden (2006). The above-mentioned log-normal model was defined in WinBUGS (Appendix B), with the restriction that the time discriminations were fixed to one. Furthermore, the CIRT software of Fox et al. (2007) was used. They modeled item responses and RTs using a hierarchical RT item response model to measure speed of working and accuracy. In this modeling framework, speed of working and accuracy are assumed to be correlated, since the speed of working is assumed to influence the accuracy of responses. In this parameter recovery study, item responses and RTs were simulated with zero correlation between the latent variables' speed of working and accuracy. Therefore, a comparison can be made between the parameter estimates of the LNRT, the WinBUGS program, and the CIRT program, since the influence of the item responses on the log-normal model estimates was negligible. In this way the performance of the LNRT program can be evaluated.

A test length of 10 and sample sizes of 500 and 1,000 test takers were considered. Normally distributed RTs were simulated on a logarithmic scale. The working speed was generated from a standard normal distribution. The time intensities were generated from a normal distribution with a mean of zero and a standard deviation of one, respectively.

In Tables 2 and 3, the simulated (true) parameters and expected a posteriori (EAP) estimates are given for the three different programs. For both sample sizes, the time-intensity parameter estimates are comparable for the different programs and are close to the true parameter values. The estimated standard deviations of the time-intensity parameters are slightly higher for the WinBUGS program than for other programs, which might be caused by the slightly less informative prior specifications.

The population variance of the time intensities is slightly overestimated by the CIRT program for both sets. Although the true value of σ_λ^2 was set to one, the empirical variance of the estimated time intensities was around .33. This value corresponds with the EAP estimates from the LNRT and WinBUGS program. The CIRT program computes the covariance matrix of all item characteristics (Fox et al., 2007). In CIRT, the default prior for the covariance matrix is an inverse Wishart distribution, which often leads to an overestimation of the covariance parameters when they are relatively small. The other programs used an inverse-gamma distribution as a prior for the variance parameter. The variance parameter of the population distribution of working speed was correctly estimated by all models.

TABLE 2
Parameter estimates from LNRT, WinBUGS, and CIRT for N = 500 and K = 10

		Ten Items (I = 10)						
		True Values	LNRT		WinBUGS		CIRT	
	Parameter	Mean	Mean	*SD*	Mean	*SD*	Mean	*SD*
N = 500	λ_1	−0.366	−0.333	0.033	−0.332	0.051	−0.33	0.03
	λ_2	0.539	0.5	0.051	0.496	0.066	0.5	0.05
	λ_3	0.735	0.671	0.051	0.662	0.066	0.67	0.05
	λ_4	0.104	0.125	0.059	0.123	0.075	0.13	0.06
	λ_5	−0.623	−0.734	0.058	−0.725	0.073	−0.73	0.06
	λ_6	0.917	0.932	0.044	0.927	0.06	0.93	0.04
	λ_7	−0.414	−0.373	0.051	−0.369	0.067	−0.37	0.05
	λ_8	−0.436	−0.478	0.054	−0.474	0.07	−0.48	0.05
	λ_9	−0.014	−0.014	0.045	−0.012	0.06	−0.01	0.04
	λ_{10}	−0.443	−0.448	0.024	−0.447	0.05	−0.45	0.02
	Population Parameters							
	σ_ζ^2	1	1.011	0.070	0.802	0.059	1.022	0.071
	σ_λ^2	1	0.331	0.185	0.333	0.184	1.439	0.773
	μ_λ	0	−0.010	0.120	−0.018	0.188	−0.015	0.381

TABLE 3
Parameter estimates from LNRT, WinBUGS, and CIRT for N = 1,000 and K = 10

		Ten Items (I = 10)						
		True Values	LNRT		WinBUGS		CIRT	
	Parameter	Mean	Mean	*SD*	Mean	*SD*	Mean	*SD*
N = 1,000	λ_1	0.385	0.370	0.027	0.372	0.041	0.370	0.030
	λ_2	−0.104	−0.119	0.032	−0.117	0.044	−0.120	0.030
	λ_3	−0.754	−0.636	0.042	−0.633	0.054	−0.630	0.040
	λ_4	0.414	0.417	0.036	0.418	0.048	0.420	0.040
	λ_5	2.093	2.024	0.030	2.023	0.043	2.020	0.030
	λ_6	0.105	0.098	0.026	0.099	0.041	0.100	0.030
	λ_7	−0.131	−0.106	0.030	−0.102	0.043	−0.100	0.030
	λ_8	0.351	0.347	0.031	0.348	0.045	0.350	0.030
	λ_9	−0.808	−0.770	0.035	−0.768	0.047	−0.770	0.030
	λ_{10}	−1.551	−1.532	0.034	−1.526	0.046	−1.530	0.030
	Population Parameters							
	σ^2_ζ	1	0.974	0.049	0.916	0.046	0.975	0.049
	σ^2_λ	1	0.898	0.478	0.903	0.504	1.980	1.065
	μ_λ	0	0.006	0.140	0.017	0.301	0.010	0.448

Study 2: Investigation of Detection Rates

Data sets were generated under different types of response behavior to simulate aberrant responses. Different data specifications were considered: sample sizes of 500 and 1,000 test takers, and test lengths of 10 and 20 items. For each type of aberrant response behavior, 5%, 10%, or 20% of the test takers responded in this way. The remaining response patterns were generated according to the log-normal model. The specification of the log-normal model was equal to the setting in the parameter recovery study, except that time-discrimination parameters were generated from a normal distribution with mean = 1 and variance = .17. Three types of aberrant behavior were simulated:

Random response behavior. The first type of aberrant RTs represented test takers who responded to the test items with random RTs on a subset of items. The simulated aberrant RTs did not correspond with the time intensities of the items. Much faster or slower times were simulated given the time intensities of the items. For half of the test items, aberrant RTs were generated from a log-normal distribution with the mean equal to the average item time and three times the average standard deviation of the RTs. The average test times for the aberrant RT patterns were similar to those for the nonaberrant RT patterns. This

corresponds to the strategy that a test taker might know the average time to complete the test but not the average time to complete each item.

Test speededness or variant working speed. Test takers with an invariant working speed will work with a constant level of speed. The assumption of conditionally independently distributed RTs given working speed is violated when the working speed is variant. This can occur when, for example, the test taker is not concentrating, has preknowledge of some items, or operates under higher time pressure than others. In this second type of aberrant pattern, half of the test items were answered much faster than expected under the log-normal model. For half of the test items, working speed of (aberrant) test takers with a variant working speed were simulated to be 1.5 standard deviations faster than the population average working speed.

One extreme RT. Test takers are assumed to work with a constant speed such that the total test time is assumed to reflect the total amount of time required to produce all answers. The total test time will be biased when test takers are interrupted or distracted while taking the test. When a test taker is taking a break (e.g., getting coffee) and is not working on the test, the next observed RT will not reflect the time spent on producing an answer. This will also bias the total test time. In this third condition, extreme RTs were simulated from a log-normal distribution with a mean equal to at least twice the maximum time intensity of the items in the test. Each aberrant RT pattern consisted of only one extreme RT.

The detection and false-alarm rates were investigated under the log-normal model for the different types of violations. In this study, item parameters were assumed to be known, but the working speed and other model parameters were estimated from the data using the LNTR program. Note that the posterior uncertainty in the model parameters were taken into account in the estimation of the test statistics and the flagging of RT patterns. RT patterns were flagged to be aberrant in different ways. First, following Equation (16), each test taker's probability of a flagged pattern was computed. Subsequently, the average posterior probability was computed from the individual posterior probabilities of a flagged pattern, thus representing the average posterior probability of flagged patterns in the population. Under the model, this average probability of flagged patterns represents the Type I error. Furthermore, for RTs generated under the model, patterns were approximately flagged to be aberrant with probability .05, when using the significance level $\alpha = .05$. Second, patterns were flagged to be aberrant when the posterior probability of an aberrant pattern was at least .80 or .90 (according to Equation (16)), which will be referred to as the classification probability.

Comparing Three Statistics

Before looking into detail at the false alarm rates and detection for the various conditions, the three statistics in Table 1 were compared. For data simulated under the log-normal model, the classification probability of being assigned to the class of patterns included in the estimation of item parameters (according to Equation (19)) and the probability of a flagged pattern (according to Equation (16)) were computed for the three statistics. In Figure 1, for each statistic the probabilities of each pattern are plotted against each other and a smoothing curve is drawn through the points to

represent the relationship. For the curve of l^t and l_s^t, patterns with a classification probability less than 5% are most likely to be flagged as aberrant, since a significance level of 5% was used. Both statistics give a similar picture, and the curves are almost equal. Therefore, it can be concluded that the approximate null distribution of l_s^t is nearly as accurate as the exact null distribution of l^t.

The curve of the approximate null distribution of l_z^t shows a shift to the left for low classification probabilities. These posterior classification probabilities are too conservative, which leads to lower probabilities of being flagged for l_z^t compared to l^t. This makes l_z^t not very useful for the detection of aberrant patterns.

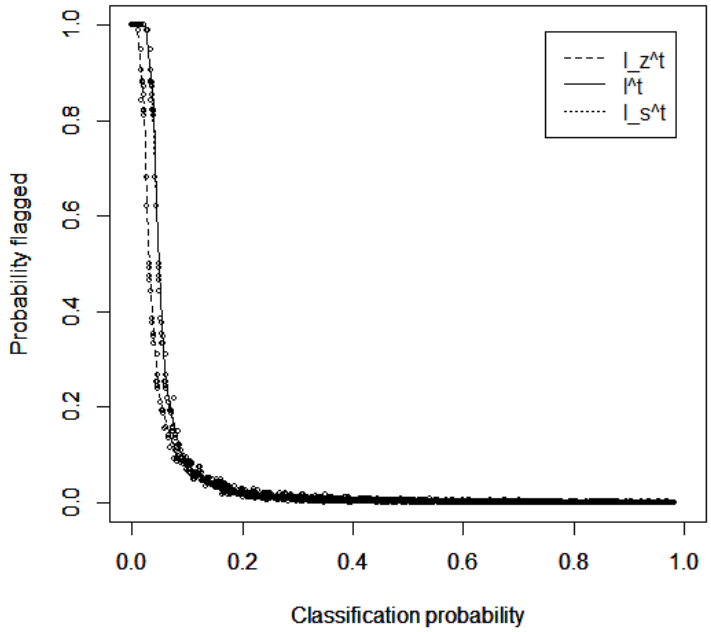

FIGURE 1. *Classification probability versus probability of being flagged for the three different statistics (N = 1,000, I = 10)*

For each RT pattern, a probability of being flagged and a classification probability are computed. In Figure 1, each point of the curve represents an RT pattern. The location of the point in the curve shows whether it is a regular or a suspicious pattern. The Type I error is equal to the expected probability of being flagged in the population. Patterns can be marked as aberrant with a posterior probability of at least .80.

Since l_z^t is not very useful for the detection of aberrant patterns and the approximate null distribution of l_s^t is nearly as accurate as the exact null distribution of l^t, attention will be focused on l^t in the simulation study.

Model-Fitting Responses and Random Response Behavior

In Table 4, the false-alarm rates and detection rates, averaged over 50 replicated data sets, are given for the l^t statistic for different sample sizes and for model-fitting responses and responses with 5%, 10%, and 20% of the RT patterns generated under random response behavior.

In the model-fitting condition, differences in false-alarm rates were found. The false-alarm rate is slightly lower for a population size of 500 compared to a size of 1,000. When flagging patterns with a posterior classification probability of at least .80, the false-alarm rate is much lower than the results for the average posterior probability flagging and decreases slightly more for a classification probability of .95. In that case, only the most extreme patterns are classified.

With respect to aberrant response types, the aberrant patterns were detected in all cases under all classification probabilities (under the heading “Aberrant” in Table 4). Given the specifications of random response behavior, the patterns were detected as significantly different from patterns that can be expected under the model. When 5% was simulated to be aberrant, then this 5% was also identified in the population (under the heading “Aberrant”). Under the different percentages, the percentage of aberrant patterns was still detected in the population.

TABLE 4

False alarm rates and detection rates of l^t for a 10- and 20-item test and 500 and 1,000 examinees using a significance level of .05 (50 replications)

			Random Response Behavior					
		Model Fit	5%		10%		20%	
	Posterior Classification	Population	Aberrant	Population	Aberrant	Population	Aberrant	Population
N = 500 I = 10	No	0.044	1.000	0.052	1.000	0.102	1.000	0.201
	.80	0.025	1.000	0.050	1.000	0.100	1.000	0.200
	.95	0.021	0.999	0.050	1.000	0.100	1.000	0.200
N = 1,000 I = 10	No	0.056	1.000	0.051	1.000	0.101	1.000	0.201
	.80	0.035	1.000	0.050	1.000	0.100	0.999	0.200
	.95	0.030	1.000	0.050	0.999	0.100	0.999	0.200
N = 500 I = 20	No	0.035	1.000	0.050	1.000	0.100	1.000	0.200
	.80	0.024	1.000	0.050	1.000	0.100	1.000	0.200
	.95	0.019	1.000	0.050	1.000	0.100	1.000	0.200
N = 1,000 I=20	No	0.047	1.000	0.050	1.000	0.100	1.000	0.200
	0.80	0.033	1.000	0.050	1.000	0.100	1.000	0.200
	0.95	0.029	1.000	0.050	1.000	0.100	1.000	0.200

Test Speededness

In Table 5, detection rates are given for the l^t statistic for different sample sizes and responses simulated under test speededness or variant working speed. In the same way, data sets were simulated with 5%, 10%, and 20% of the RT patterns generated under test speededness, and patterns were flagged to be aberrant with a significance level of .05.

For different percentages, with patterns showing test speededness, the detection rate is around .90 for a test of 10 items and approximately .99 for a longer test of 20 items. The detection rates are only somewhat smaller when they are computed using a classification probability of at least .80 or .90. In the worst case of 20% aberrant

patterns, the detection rate is around 77% of the simulated aberrant patterns. When looking at the percentage of detections in the population, slightly more patterns are flagged than the simulated percentage of aberrant patterns.

TABLE 5

Detection rates of l^t for a 10- and 20-item test and 500 and 1,000 examinees using a significance level of .05 (50 replications)

		Test Speededness					
		5%		10%		20%	
	Posterior Classification	Aberrant	Population	Aberrant	Population	Aberrant	Population
N = 500 I = 10	No	0.888	0.078	0.885	0.116	0.850	0.192
	.80	0.859	0.060	0.855	0.097	0.800	0.166
	.95	0.848	0.056	0.836	0.093	0.771	0.159
N = 1,000 I = 10	No	0.929	0.093	0.917	0.131	0.878	0.205
	.80	0.910	0.073	0.894	0.110	0.836	0.176
	.95	0.899	0.068	0.880	0.105	0.816	0.170
N = 500 I = 20	No	0.991	0.074	0.990	0.121	0.979	0.213
	.80	0.987	0.063	0.986	0.110	0.813	0.167
	.95	0.986	0.060	0.982	0.107	0.807	0.164
N = 1,000 I = 20	No	0.995	0.085	0.994	0.131	0.988	0.224
	.80	0.993	0.072	0.992	0.117	0.981	0.205
	.95	0.991	0.069	0.990	0.114	0.978	0.202

One Extreme Response

In Table 6, averaged over 50 replicated data sets, detection rates are given for the l^t statistic for different sample sizes and RT patterns including an extreme response for the first item. The detection rates are somewhat acceptable, when only 5% of the patterns include an extreme response. When the test length increases, the detection rates decrease, since it becomes more difficult to identify the longer RT patterns with just one extreme RT. When the sample size increases, the detection rates also increase. A distortion in detection rates became visible when the percentage of aberrant patterns increased. In that case, the measurement error variance increased, which simply adjusted the range of possible RTs. Thus, the variability in RTs for the first item was increased by an increase in the estimated measurement error variance for the first item. The detection rates were much better when the extreme response was randomly assigned across patterns to one of the test items.

TABLE 6

Detection rates of l^t for a 10- and 20-item test and 500 and 1,000 examinees using a significance level of .05 (50 replications).

		An Extreme RT					
		5%		10%		20%	
	Posterior Classification	Aberrant	Population	Aberrant	Population	Aberrant	Population
N = 500 I = 10	No	0.830	0.072	0.732	0.101	0.314	0.088
	.80	0.782	0.055	0.664	0.081	0.251	0.064
	.95	0.738	0.049	0.604	0.072	0.219	0.055
N = 1,000 I = 10	No	0.858	0.083	0.741	0.111	0.380	0.108
	.80	0.824	0.065	0.688	0.090	0.320	0.083
	.95	0.788	0.06	0.636	0.081	0.288	0.073
N = 500 I = 20	No	0.676	0.057	0.473	0.072	0.137	0.048
	.80	0.606	0.044	0.396	0.056	0.105	0.034
	.95	0.554	0.039	0.352	0.049	0.089	0.028
N = 1,000 I = 20	No	0.811	0.077	0.555	0.090	0.175	0.064
	.80	0.766	0.063	0.490	0.073	0.141	0.047
	.95	0.715	0.058	0.446	0.065	0.127	0.042

Mixture Modeling

The mixture modeling approach was used to avoid the distortion in parameter estimates due to the aberrant RT patterns. In Table 7, the false-alarm and detection rates are presented for the different types of aberrant response behavior. For the RTs that fit the model, the false-alarm rate of 3.9% is only slightly smaller than the significance level of 5%. The computation and evaluation of the test statistic leads to quite accurate Type I errors. For each type of aberrant response behavior, results comparable to those in Table 4 are obtained.

For test speededness, results similar to those shown in Table 5 are obtained when 5% or 10% of the simulated patterns are aberrant. When the percentage of aberrant patterns increases to 20%, the detection rates are much lower. In that case, the item parameters are biased due to the aberrant RT patterns, which are not classified as aberrant. A biased proportion of flagged patterns is obtained, and around 10% of the aberrant patterns are not detected. For the last type, results comparable to those shown in Table 6 are obtained. When the item parameters are known, slightly higher detection rates are obtained. However, as in Table 6, the detection rates are acceptable for 5% aberrant RT patterns. For higher percentages, the detection rates are low, since the measurement error variance of the first item accommodates extreme RTs for the first item.

TABLE 7

Detection rates of l^t for a 10-item test and 1,000 examinees using a significance level of .05 for different types of aberrant response behavior

		Aberrant Response Behavior					
	Model Fit	5%		10%		20%	
Posterior Classification	Population	Aberrant	Population	Aberrant	Population	Aberrant	Population
		Random Response Behavior					
.80	0.023	1.000	0.050	1.000	0.100	0.999	0.200
.95	0.019	1.000	0.050	0.999	0.100	0.999	0.200
No	0.039	1.000	0.051	1.000	0.101	0.999	0.200
		Test Speededness					
.80		0.818	0.057	0.759	0.088	0.388	0.085
.95		0.796	0.052	0.730	0.083	0.352	0.076
No		0.851	0.072	0.799	0.104	0.452	0.106
		An Extreme RT					
.80		0.707	0.051	0.462	0.060	0.147	0.041
.95		0.668	0.045	0.424	0.053	0.130	0.035
No		0.757	0.065	0.525	0.077	0.187	0.059

Discussion

The response behavior of test takers needs to be checked in order to assess the quality of tests. Aberrant response behavior will bias the test results, represented by biased parameter estimates and incorrect statistical inferences. RT patterns can be checked by evaluating the residuals given a model that explains the variability of patterns of a population of regular test takers. As an analogue to the likelihood-based statistic in person-fit testing to evaluate response patterns, usually denoted as l_z, a likelihood-based person-fit statistic for RT patterns was proposed, denoted as l^t. In total, three versions of this statistic were considered: l_z^t and l_s^t have approximately normal sampling distributions, and l^t has an exact chi-squared distribution.

Various statistical techniques have been proposed in the literature to check response patterns. Residual analysis and checks on aberrant response patterns have been proposed, and extensive literature reviews have been done by Meijer and Sijtsma (1995, 2001) and Karabatsos (2003). A check for RT patterns has been discussed by van der Linden and van Krimpen-Stoop (2003) and van der Linden and Guo (2008), who have mainly been interested in detecting cheating behavior. Their method is based on evaluating the posterior probability that an observed RT is lower or higher than the posterior predicted RT under the model. In this report, the actual size of each residual is also taken into account, which makes it possible to assess the extremeness of a single RT. Furthermore, the null distribution is known, which is used to quantify the extremeness of each pattern and to compute the posterior probability of an aberrant pattern under the null model.

Different types of aberrant response behavior were considered. The best results were obtained for random response behavior using l^t or l_s^t, where we found detection rates close to one under different conditions. When test takers manipulate their RTs to match the total test time (e.g., in case of cheating), aberrant RT patterns can still be accurately identified given the discrepancy for each item between the observed RT and the expected RT under the model. It was remarkable that accurate

detection rates were obtained for a 10-item test. The continuous nature of the RTs increase the amount of information significantly compared to the categorical nature of item responses.

In the case of test speededness, acceptable detection rates were obtained, which also increased when assuming a known or calibrated working speed. When the test time is limited, violations of a constant working speed can occur at the end of the test. The proposed test can facilitate in evaluating the occurrence of speededness effects, where test takers are confronted with time limits. When tests are given without the objective being to measure the working speed, which is also unrelated to the construct that is supposed to be measured, the purpose of the test is negatively affected by effects of speededness. The performance of the test takers will be affected by effects of test speededness, which can lead to more guessing behavior and inaccurate item parameter estimates (Bolt et al., 2002)

For the last type of aberrant behavior, one extreme RT, the percentage of aberrant patterns in the sample highly influenced the results. The detection rates were acceptable when 5% of the RT patterns were simulated to be aberrant, but for higher percentages the detection rates were much lower. In that case, the estimate of the measurement error variance increased to adapt the model to account for additional variability in an item vector of RTs. By restricting the variance to be constant across items, patterns with an extreme RT can also be identified.

Overall it can be concluded that the two-component mixture modeling approach worked well to calibrate the item parameters using a selected set of RT patterns, and to flag aberrant RT patterns. The results of the test statistics based on estimated and true item parameters did not differ much. Only for test speededness, with 20% of simulated aberrant RT patterns, the detection rates were much lower due to biased item parameter and biased classification estimates.

RT checks are meant to identify aberrant patterns, which can appear for several reasons. The proposed checks can be used to flag patterns, and adjustments can be made to flag items as well. Further investigations are required to analyze flagged patterns more thoroughly using possibly additional information. Other types of residual checks can be defined. For example, statistics based on residuals can be used to investigate RT differences between groups of test takers. Item-specific between-group differences in RTs can indicate differential item functioning; that is, an item's time intensity differs across groups. Between-group differences in RTs can also indicate group-specific distributions of working speed.

More research is needed to include response information in the detection of aberrant response behavior. The connection of RT patterns with patterns of accuracy (correct/incorrect) will certainly increase the power of detecting aberrant behavior (Van der Linden & Guo, 2008).

References

Bergstrom, B. A., Gershon, R. C., & Lunz, M. E. (1994, April). *Computer adaptive testing: Exploring examinee response time using hierarchical linear modeling*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331–348.

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, *143*(4), 383–430.

Box, G. E., Hunter, J. S., & Hunter, W. G. (1978). *Statistics for experimenters*. New York: Wiley.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86.

Fox, J. P., Klein Entink, R., & Linden, W. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software*, *20*(7), 1–14.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2004). *Bayesian data analysis* (2nd Ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Hornke, L. F. (1997). Untersuchung von Itembearbeitungszeiten beim computergestützten adaptiven Testen. *Diagnostica, 43*, 27–39.

Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicológica: Revista de metodología y psicología experimental*, *21*(1), 175–190.

Hornke, L. F. (2005). Response time in computer-aided testing: A “verbal memory” test for routes and maps. *Psychology Science*, *47*(2), 280.

Jansen, M. G. (2007). Testing for local dependence in Rasch’s multiplicative gamma model for speed tests. *Journal of Educational and Behavioral Statistics*, *32*(1), 24–38.

Jansen, M. G., & Glas, C. A. (2001). Statistical tests for differential test functioning in Rasch’s model for speed tests. In *Essays on item response theory* (pp. 149–162). New York: Springer.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298.

Klein Entink, R. H., Fox, J. P., & Van Der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74*(1), 21–48.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational and Behavioral Statistics*, *4*(4), 269–290.

Masters, G., & Keeves, J. (1999). *Advances in measurement in educational research and assessment.* Amsterdam: Elsevier Science.

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*(4), 321–336.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8*(3), 261–272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*(1), 75–106.

Rubin, D. B. (1984). Comment: assessing the fit of logistic regressions using the implied discriminant analysis. *Journal of the American Statistical Association, 79*(385), 79–80.

Schnipke, D. L., & Pashley, P. J. (1997, March). *Assessing subgroup differences in item response times.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232.

Schnipke, D. L., & Scrams, D. J. (1999a). *Exploring issues of test taker behaviour: Insights gained from response time analyses.* Princeton, NJ: Law School Admission Council.

Schnipke, D. L., & Scrams, D. J. (1999b). *Representing response-time information in item banks* (Vol. 97, No. 9). Law School Admission Council.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building thefoundation for future assessments* (pp. 237–266). Mahwah, NJ: LawrenceErlbaum Associates, Inc

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66,* 331–342.

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

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365–384.

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika, 68*(2), 251–265.

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

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer et al. (Eds.), *Computer adaptive testing: A primer* (2nd ed.) (pp. 271–299). Hillsdale, NJ: Laurence Erlbaum Associates.

Weiss, D. J., & Schleisman, J. L. (1999). Adaptive testing. In *Advances in measurement in educational research and assessment* (pp.129–137). Amsterdam: Elsevier Science.

Appendix A

The marginal distribution of the RT data is used to evaluate the fit of an RT pattern. This l_0 statistic as defined in Equation (5) can be standardized to derive the null distribution. The standardized version is denoted as l_z^t, which requires the computation of the expected value and the variance.

The l_0 follows from the independently normally distributed logarithm of RTs as stated in Equation (6). Then the expected statistic value as a function of the RT is given by,

$$\begin{aligned} E(l_0) &= E\left(\sum_i \frac{\left(T_{pi}^* - \mu_{pi}\right)^2}{\sigma_i^2} + \log\left(2\pi\sigma_i^2\right)\right) = \sum_i \left(E\left(\frac{T_{pi} - \mu_{pi}}{\sigma_i}\right)^2 + \log\left(2\pi\sigma_i^2\right)\right) \\ &= \sum_i \left(E\left(Z_{pi}^2\right) + \log\left(2\pi\sigma_i^2\right)\right) = \sum_i \left(1 + 1\log\left(2\pi\sigma_i^2\right)\right) = I + \sum_i 1\log\left(2\pi\sigma_i^2\right) \end{aligned} \quad \text{(A-1)}$$

since the Z_{pi} is standard normally distributed and the expected value of a squared standard normally distributed variable equals one $\left(E\left(Z_{pi}^2\right) = Var\left(Z_{pi}\right) = 1\right)$.

The variance of the statistic value as a function of the RT is given by

$$\begin{aligned} Var(l_0) &= \sum_i Var\left(\left(\frac{T_{pi}^* - \mu_{pi}}{\sigma_i}\right)^2\right) \\ &= \sum_i E\left(\left(\frac{T_{pi}^* - \mu_{pi}}{\sigma_i}\right)^2\right)^2 - \left(E\left(\left(\frac{T_{pi}^* - \mu_{pi}}{\sigma_i}\right)^2\right)\right)^2 \\ &= \sum_i E\left(Z_{pi}^4\right) - \left(E\left(Z_{pi}^2\right)\right)^2 = \sum_i (3-1) = 2I. \end{aligned} \quad \text{(A-2)}$$

The expected value of the fourth power of a standard normally distributed variable follows from a variable transformation. Let $y = Z_{pi}^2$. Then, $E\left(Z_{pi}^4\right) = E\left(y^2\right)$, which can be expressed as a Gamma distribution with shape parameter 5/2 and scale parameter 2. The value three follows from the fact that the Gamma density integrates to one over the range of positive numbers.

Appendix B

WinBUGS code for the one-parameter log-normal RT model. This also includes the computation of the l^t statistic and the standardized versions.

```
model
{
   for (p in 1:N) {
     for (i in 1:I) {
     T[p,i] ~ dlnorm(mu[p,i],sigmam[i])
          mu[p,i] <- intensity[i] - speed[p]
          lZd[p,i] <- pow(log(T[p,i]) - mu[p,i],2)*sigmam[i]
      }
 speed[p] ~ dnorm(0,sigmasp)
 lZ[p] <- (sum(lZd[p,1:I]) - I)/sqrt(2*I)
dum[p] <- (pow(mean(lZd[p,1:I]),1/3) - (1-2/(9*I)))/sqrt(2/(9*I))
lZP1[p] <- 1 - phi(dum[p])    #tail-area probability statistic
lZP2[p] <- 1 - phi(lZ[p])  #tail-area probability standardized statistic
  }

#Priors
 for (i in 1:I) {
     intensity[i] ~ dnorm(mub,sigmab)
          sigmam[i] ~ dgamma(.1,.1)
     sigmamn[i] <- 1/sigmam[i] #Measurement error variance
}

#Hyper prior

mub ~ dnorm(0,1.0E-6)
sigmab ~ dgamma(.1,.1)
sigmabn <- 1/sigmab    #Population variance time intensity
sigmasp ~ dgamma(1,.1)
sigmaspn <- 1/sigmasp #Population variance speed test takers
}
```