Journal of Educational and Behavioral Statistics
2014, Vol. 39, No. 6, pp. 426–451
DOI: 10.3102/1076998614559412

Testing for Aberrant Behavior in Response Time Modeling

Sukaesi Marianti
Jean-Paul Fox
Marianna Avetisyan
Bernard P. Veldkamp
University of Twente

Jesper Tijmstra
Tilburg University

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. This 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 RT 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 RTs 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 this study, Bayesian tests of significance for detecting various types of aberrant RT patterns are proposed and evaluated. In a simulation study, the tests were successful in identifying aberrant response patterns. A real data example is given to illustrate the use of the proposed person-fit tests for RTs.

Keywords: *response times; aberrant behavior; person fit*

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 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; in other words, RTs are modeled independently of the correctness of the response. Examples of this approach are as follows: Maris (1993) modeled RTs exclusively, whereas accuracy scores are not taken into consideration. Schnipke and Scrams (1997) estimated rapid guessing with assumption that accuracy and RTs are independent given speed and ability. (b) Test models that describe the distribution of RTs as well as responses. This approach takes correctness of the response and RTs into account; the correct responses reflect both speed and accuracy. 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.

Recently, there is another approach introduced by van der Linden (2006, 2007) who 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.

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 lognormal RT model to deal with various types of aberrant behavior. Based on this lognormal 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. 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.

After introducing the lognormal RT model, an estimation procedure is described to estimate simultaneously all model parameters. Then, person-fit statistics are defined under the lognormal RT model, which differ with respect to their null distribution. It will be shown that given all information, each RT pattern can be flagged as aberrant with a specific posterior probability, to quantify the extremeness of each pattern under the model. In a simulation study, the power to detect the aberrancies is investigated by simulating various types of aberrant response behavior. Finally, the results from a real data example and several directions for future research are presented.

RT Modeling

van der Linden (2006) proposed a lognormal 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 lognormal 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 lognormal RT models, each test taker is assumed to have a constant working speed during the test. Let $p = 1, \ldots, N$ be an index for the test takers, $i = 1, \ldots, I$ 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 is the mean. In the same way, the faster a test taker operates, the lower is 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 $\mu_{pi} = \phi_i(\lambda_i - \zeta_p)$, such that the reduction in RT by operating faster is not constant over items. The higher the time discrimination of an item, the higher is the reduction in the mean when operating faster. For example, when a test taker operates a constant C faster, the mean equals $\mu_{pi} = \phi_i(\lambda_i - (\zeta_p + C)) = \phi_i(\lambda_i - \zeta_p) - \phi_i C$, 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 lognormal RT model, σ_i^2 can 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 lognormal model for the distribution of T_{ip} can be represented by

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

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.

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 2 items with the same time intensity, where 1 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. Note that this value of zero corresponds to the population-average total test time, which corresponds to the sum of all time intensities. 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 Lognormal 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 γ 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 γ 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 lognormal model leads to the following random effects model to model the logarithm of RTs:

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

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 Lognormal 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. To apply Gibbs sampling, the full conditional

distributions of the model parameters need to be specified. For the lognormal 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 $t^*_{ip} = \ln(t_{ip})$ 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,

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

The l_o statistic can be evaluated over all items in the test, but it is also possible to consider a subpart of the test. An unusually large value 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 $\mathbf{t}^*_p = \left(t^*_{1p}, \ldots, t^*_{Ip}\right)$ is given by:

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

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

The test statistic l_o 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 & Sijtsma, 1995; 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 the Appendix, 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:

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

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

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

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 $\boldsymbol{T}_p^*$. 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 lognormal 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 χ^2 distributed with I degrees of freedom. Box, Hunter, and Hunter (1978, p. 118) showed that a χ^2 distributed variable T with I degrees of freedom, the distribution of $(T-I)/\sqrt{2I}$ 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 χ^2 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

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

Summarized, three person-fit statistics for RTs are considered that differ in the way the null distribution is derived. An overview of the tests is given in Table 1.

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,

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

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	χ^2	Exact	$P\left(l^t\left(\mathbf{T}_p^*\right) > C\right) = P\left(\chi_I^2 > C\right)$
l_s^t	Normal	Approximation	$P\left(l_s^t\left(\mathbf{T}_p^*\right) > C\right) \approx \Phi\left(l_s^t\left(\mathbf{T}_p^*\right) > C\right)$

since the null distribution is a χ^2 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 lognormal model. When the posterior probability is close to zero, it can be concluded that the pattern is unlikely under the posited lognormal 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:

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

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 Equations (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 1 when the RT pattern of test taker P is flagged, or a value of 0 otherwise. Thus,

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

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

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

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 lognormal 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), 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.

Results

Through simulation studies, the performance of the person-fit statistics for RT patterns is evaluated. A comparison is made between three different programs for estimating the model parameters. 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. The MCMC method for estimating the model parameters of the lognormal model was implemented in *R* and is referred to as log normal response times (LNRT).[1]

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 lognormal model. The specification of the lognormal 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 lognormal distribution with the mean equal to the average item time and 3 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 lognormal model. For half of the test items, working speed of (aberrant) test takers with a variant working speed was 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 lognormal 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 lognormal 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 LNRT program. Note that the posterior uncertainty in the model parameters was 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 lognormal 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.

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. Subsequently, patterns can be marked as aberrant with a specific posterior probability, which represents the accuracy of making the right decision. However, increasing the accuracy of correctly identifying an aberrant pattern is accompanied with a decrease in the probability of identifying all aberrant patterns.

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 2, 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.

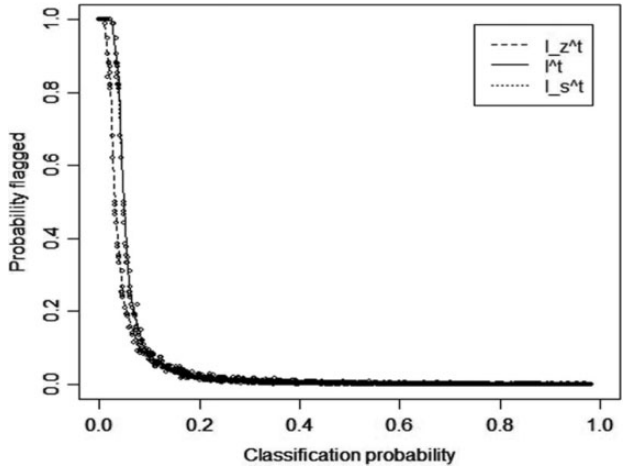

FIGURE 1. *Classification probability versus probability of being flagged for the three different statistics (*N = *1,000,* I = *10).*

With respect to aberrant response types, the aberrant patterns were detected in all cases under all classification probabilities (under the heading Aberrant in Table 2). 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.

Test Speededness

In Table 3, 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 2

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					
			5%		10%		20%	
	Posterior Classification	Model Fit Population	Aberrant	Population	Aberrant	Population	Aberrant	Population
	No	.044	1.000	.052	1.000	.102	1.000	.201
$N = 500$, $I = 10$	.80	.025	1.000	.050	1.000	.100	1.000	.200
	.95	.021	0.999	.050	1.000	.100	1.000	.200
$N = 1{,}000$, $I = 10$	No	.056	1.000	.051	1.000	.101	1.000	.201
	.80	.035	1.000	.050	1.000	.100	0.999	.200
	.95	.030	1.000	.050	0.999	.100	0.999	.200
$N = 500$, $I = 20$	No	.035	1.000	.050	1.000	.100	1.000	.200
	.80	.024	1.000	.050	1.000	.100	1.000	.200
	.95	.019	1.000	.050	1.000	.100	1.000	.200
$N = 1{,}000$, $I = 20$	No	.047	1.000	.050	1.000	.100	1.000	.200
	0.80	.033	1.000	.050	1.000	.100	1.000	.200
	0.95	.029	1.000	.050	1.000	.100	1.000	.200

TABLE 3

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
	No	.888	.078	.885	.116	.850	.192
$N = 500$, $I = 10$	.80	.859	.060	.855	.097	.800	.166
	.95	.848	.056	.836	.093	.771	.159
$N = 1{,}000$, $I = 10$	No	.929	.093	.917	.131	.878	.205
	.80	.910	.073	.894	.110	.836	.176
	.95	.899	.068	.880	.105	.816	.170
$N = 500$, $I = 20$	No	.991	.074	.990	.121	.979	.213
	.80	.987	.063	.986	.110	.813	.167
	.95	.986	.060	.982	.107	.807	.164
$N = 1{,}000$, $I = 20$	No	.995	.085	.994	.131	.988	.224
	.80	.993	.072	.992	.117	.981	.205
	.95	.991	.069	.990	.114	.978	.202

One Extreme Response

In Table 4, 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.

In Figure 2, the receiver operating characteristic (ROC) curves of the l^t test illustrate the performance for artificial data generated for 1,000 students and 10 items, where 10% of the students show aberrant behavior on 5 items. The x-axis, referred to as the false alarm rate, represents the percentage of incorrectly identified aberrant RT patterns and the y-axis, referred to as the hit rate (i.e., sensitivity), represents the percentage of correctly identified aberrant RT patterns.

For the left plot, random RTs were generated for 5 items, where the variability in random RTs was equal to or 1½ or 2 times larger than the variability in RTs generated under the lognormal model. It follows that for small threshold values, accurate decisions can be made when the variance of random RTs is larger than the variance of the model generated RTs. In that case, with a significance level of .1, more than 80% of the patterns can be correctly classified.

For the speededness condition, the performance of the l^t test was less good. In this condition, 10% of the students worked slower on the first 5 items. Their speed levels were one, two, or three standard deviations lower compared to the last 5 items of the test. An increase of one standard deviation in working speed means for a student who was working with a population average speed level increased his or her speed level to work faster than 84% of the students in the population. It follows from the ROC curves that even in the extreme situation, only 75% of the RT patterns of students who increased their speed levels with three standard deviations were detected, given a false positive rate of less than 10%.

The difference in test performance between the two conditions can be explained by the fact that the two conditions, random RTs and speededness, induce misfit at different levels of analysis. The condition random RTs induce a misfit at the level of observations, and the l^t test is designed to detect misfits at this level. The speededness condition implies a violation at the level of students, since students were assumed to work with a constant speed level. Thus, the l^t test can only pick up the implied residual deviations due to a change in working speed at Level 2, and this decreased the performance of the test.

TABLE 4
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
	No	.830	.072	.732	.101	.314	.088
$N = 500$, $I = 10$	.80	.782	.055	.664	.081	.251	.064
	.95	.738	.049	.604	.072	.219	.055
$N = 1{,}000$, $I = 10$	No	.858	.083	.741	.111	.380	.108
	.80	.824	.065	.688	.090	.320	.083
	.95	.788	.06	.636	.081	.288	.073
$N = 500$, $I = 20$	No	.676	.057	.473	.072	.137	.048
	.80	.606	.044	.396	.056	.105	.034
	.95	.554	.039	.352	.049	.089	.028
$N = 1{,}000$, $I = 20$	No	.811	.077	.555	.090	.175	.064
	.80	.766	.063	.490	.073	.141	.047
	.95	.715	.058	.446	.065	.127	.042

Note. RT = response time.

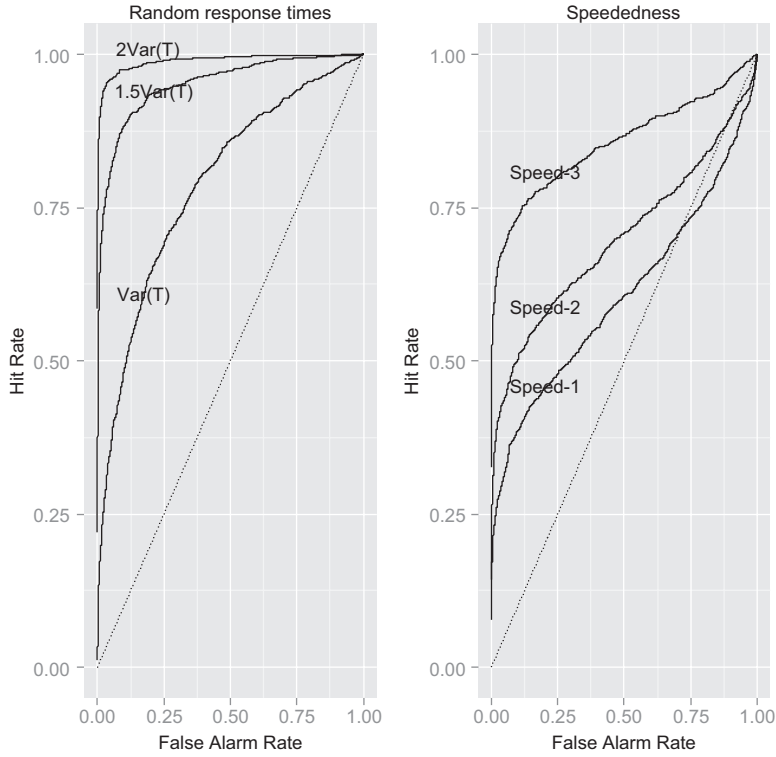

FIGURE 2. *The receiver operating characteristic (ROC) curve of the* l^t *test for simulated data (1,000 persons, 10 items) with 10% aberrant (response times) RTs according to degrees of random RTs (left subplot) and speededness (right subplot).*

Real Data Example

The data of Wise, Pastor, and Kong (2009) was investigated using the l^t person-fit statistic. The data set included 329 test takers who each answered 65 items of a computer-based version of the *Natural World Assessment* test (NAW-8). This test is used to assess the quantitative and scientific reasoning proficiencies of college students. Wise et al. tried to identify item and examinee characteristics to identify rapid guessing behavior of test takers with motivation problems. van der Linden (2009) investigated the data for a possible collusion between RT patterns of test takers. However, the main purpose of this study is to investigate the extremeness of RT patterns under the general lognormal RT model using the proposed person-fit statistic. This example will illustrate the ease of computing person-fit statistics, given RT patterns and further relevant quantities.

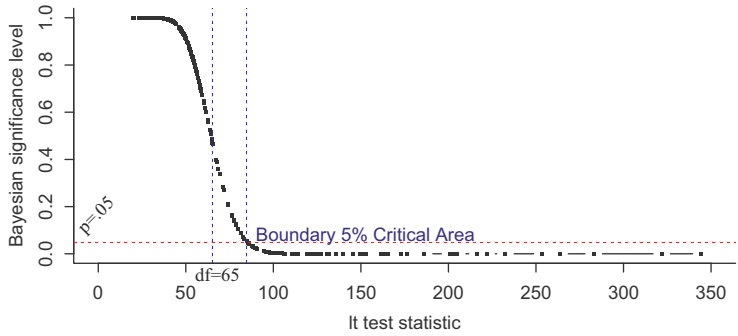

FIGURE 3. *Natural world assessment (NAW)-8 test; estimated statistic values and corresponding posterior significance levels.*

First, the lognormal RT model was fitted using 15,000 MCMC iterations and 5,000 iterations as the burn-in. The average time intensity was around 2.70 on a logarithmic scale (around 14.88 seconds), with a posterior standard deviation of .08. The variability in time intensities was around .39. The variability in test takers' working speed was around .08, where the average level of working speed in the population was fixed at 0. So most of the variability between RTs was explained by the differences in time intensities.

For each test taker, a person-fit statistic value of l^t and posterior probability of the extremeness of the RT pattern was computed. In Figure 3, the estimated statistic values (x-axis) are plotted against the posterior probability of significance. The statistic values were assumed to be χ^2 distributed with 65 degrees of freedom, which marks the point of an observing statistic value with 50% probability. When considering a significance level of .05, estimated statistic values higher than 84.82 were located in the critical region. The estimated number of aberrant patterns was around 20%, which means that one of the five patterns was flagged as aberrant, and when including variable time discriminations around 34% was identified as aberrant. The students in the test had no stakes whatsoever in the test and were not motivated to give their best effort. Wise et al. (2009) estimated the proportion of rapid guessing to be around 10%. Around 25% of the students showed rapid-guessing behavior on more than 10% of the items. However, only 7% of the rapid-guessers were also flagged by the l^t test. Our test accounts for variable working speed and item characteristics, where the diagnostic of Wise et al. is based on a known item time threshold to identify rapid guessing. Furthermore, other types of aberrant response behavior might be responsible for the 20% RT patterns that were flagged in this study.

This study stresses the importance to identify noneffortful responses, which would otherwise undermine the success of low-stakes achievements tests. A good test is of little value when students are not willing to cooperate and to put

effort in their work. Therefore, it is important to have a person-fit test for RTs that can be used to check patterns and to identify aberrant response behavior.

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 χ^2 distribution.

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).

Appendix

The marginal distribution of the RT data is used to evaluate the fit of an RT pattern. This l_o 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_o 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:

$$\begin{aligned} E(l_0) &= E\left(\sum_i \frac{\left(T_{pi}^* - \mu_{pi}\right)^2}{\sigma_i^2} + \log\left(2\pi\sigma_i^2\right)\right) = \sum_i \left(E\left(\frac{T_{pi} - \mu_{pi}}{\sigma_i}\right)^2 + \log\left(2\pi\sigma_i^2\right)\right) \\ &= \sum_i \left(E\left(Z_{pi}^2\right) + \log\left(2\pi\sigma_i^2\right)\right) = \sum_i \left(1 + 1og\left(2\pi\sigma_i^2\right)\right) = I + \sum_i 1og\left(2\pi\sigma_i^2\right), \end{aligned} \tag{A1}$$

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:

$$\begin{aligned} Var(l_0) &= \sum_i Var\left(\left(\frac{T_{pi}^* - \mu_{pi}}{\sigma_i}\right)^2\right) \\ &= \sum_i E\left(\left(\frac{T_{pi}^* - \mu_{pi}}{\sigma_i}\right)^2\right)^2 - \left(E\left(\left(\frac{T_{pi}^* - \mu_{pi}}{\sigma_i}\right)^2\right)\right)^2 \\ &= \sum_i E\left(Z_{pi}^4\right) - \left(E\left(Z_{pi}^2\right)\right)^2 = \sum_i (3-1) = 2I. \end{aligned} \tag{A2}$$

The expected value of the fourth power of a standard normally distributed variable follows from a variable transformation. Let $y = Z_{pi}^2$. Then, $E\left(Z_{pi}^4\right) = E(y^2)$, which can be expressed as a γ distribution with shape Parameter 5/2 and scale Parameter 2. The value 3 follows from the fact that the γ density integrates to 1 over the range of positive numbers.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. The LNRT program written in *R* will be made available by the authors.

References

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*, 383–430.

Box, G. E., Hunter, J. S., & Hunter, W. G. (1978). *Statistics for experimenters*. New York, NY: 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*, 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*, 1–14.

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*, 21–48.

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

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and application as psychometic models for response times. *Psychometrika*, *58*(3), 445–469.

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

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*, 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*, 79–80.

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*, 213–232.

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, NY: Academic Press.

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. (2009). A bivariate log-normal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, *34*, 378–394.

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

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*, 185–201.

Authors

SUKAESI MARIANTI is a PhD student at the Department of Research Methodology, Measurement and Data Analysis at University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; e-mail: s.marianti@utwente.nl. Her research interests are in response time modeling and detection of aberrant behavior on test items, especially in educational and psychological research.

JEAN-PAUL FOX is a professor at the Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; e-mail: j.p.fox@utwente.nl. His research interest is in Bayesian response modeling particularly in the context of large-scale surveys.

MARIANNA AVETISYAN is an assistant professor at the Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500

AE Enschede, The Netherlands; e-mail: m.avetisyan-1@utwente.nl. Her research interests are in applying Bayesian techniques and developing statistical models, in particular using Markov chain Monte Carlo (MCMC) methods in various scientific disciplines including test theory and epidemiology.

BERNARD P. VELDKAMP is a professor at the Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; e-mail: b.p.veldkamp@utwente.nl. His research interest are in computerized assessment and data mining.

JESPER TIJMSTRA is an assistant professor in psychometrics at the Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Tilburg University, 5000 LE, Tilburg, The Netherlands; e-mail: j.tijmstra@uvt.nl. His primary research interests include parametric and nonparametric item response modeling, with an emphasis on dealing with model assumptions and issues of robustness.

Manuscript received April 30, 2014
Revision received September 22, 2014
Accepted October 10, 2014