Journal of Educational and Behavioral Statistics
Vol. XX, No. X, pp. 1–26
DOI: 10.3102/1076998620957240
Article reuse guidelines: sagepub.com/journals-permissions

Testing the Within-State Distribution in Mixture Models for Responses and Response Times

Renske E. Kuijpers
Cito, Netherlands Institute for Educational Measurement

Ingmar Visser
Dylan Molenaar
University of Amsterdam

Mixture models have been developed to enable detection of within-subject differences in responses and response times to psychometric test items. To enable mixture modeling of both responses and response times, a distributional assumption is needed for the within-state response time distribution. Since violations of the assumed response time distribution may bias the modeling results, choosing an appropriate within-state distribution is important. However, testing this distributional assumption is challenging as the latent within-state response time distribution is by definition different from the observed distribution. Therefore, existing tests on the observed distribution cannot be used. In this article, we propose statistical tests on the within-state response time distribution in a mixture modeling framework for responses and response times. We investigate the viability of the newly proposed tests in a simulation study, and we apply the test to a real data set.

Keywords: *mixture models; response times; response time distribution; item response theory*

1. Introduction

Recently, research interest has grown in modeling response times next to the item responses in order to investigate individual differences in ability and speed. Focusing on the item response times in addition to the item responses has facilitated various aspects of psychological testing including, for instance, item selection in computerized adaptive testing (van der Linden et al., 1999; Veldkamp, 2016), test design (van der Linden, 2007), and item calibration (T. Wang & Hanson, 2005). In addition, response times have been shown useful in detecting item preknowledge (McLeod et al., 2003), aberrant response patterns (Marianti et al., 2014; van der Linden & Guo, 2008; C. Wang, Xu, Shang, & Kuncel, 2018), and individual differences in the use of solution strategies. For instance, van der Maas and Jansen (2003) showed that response times can give detailed information on the type and duration of different solution strategies children use to solve a balance scale task. Suitable models to enable these inferences concerning individual differences in responses and response times include the model by Roskam (1987) and more recently the hierarchical model by van der Linden (2007, 2009), which was elaborated by Molenaar et al. (2015a).

Besides these applications of response times to individual differences research, response times have been used to facilitate the study of within-subject differences in solution strategies or psychological processes that underlie the responses to psychometric tests and questionnaires. For instance, response times have been used to identify fast guessing (e.g., Schnipke & Scrams, 1997) and within-subject differences in solution strategies (e.g., Molenaar et al., 2016). Other applications include the study of within-subject differences in motivation (Wise & Kong, 2005) and faking on personality test items (Holden & Kroner, 1992).

To facilitate the detection of within-subject differences in responses and response times, various approaches based on mixture modeling have been proposed. For instance, the earliest contribution by Schnipke and Scrams (1997) focused on a two-state within-subject mixture model for the response times only. Here, one state represented rapid-guessing behavior of examinees and the other state modeled the responses of examinees who actually tried to solve the item (i.e., a regular response process). The model by Schnipke and Scrams did not include a latent speed variable but can be seen as one of the first models within this framework. In their model, the mean and variance of the response times are estimated freely for each item in the regular process state, while in the rapid-guessing state, a common mean and variance parameter is assumed to underly the items. On the basis of this model, C. Wang and Xu (2015) and C. Wang, Xu and Shang (2018) proposed a mixture model in which separate measurement models are proposed for modeling the responses and response times in the slower response state, whereas for the faster guessing state, only a guessing parameter is estimated for the responses, and a mean and variance parameter is estimated for the response times. Molenaar et al. (2016) generalized this approach by proposing a mixture model that specifies a measurement model for the responses and a measurement model for the response times separately in each state.

Inspired by the aforementioned models, the hierarchical mixture modeling approach (Molenaar et al., 2016; Schnipke & Scrams, 1997; C. Wang & Xu, 2015; C. Wang, Xu, & Shang, 2018; C. Wang, Xu, Shang, & Kuncel, 2018) that we focus on in this article is a mixture extension of the hierarchical model by van der Linden (2007, 2009) to allow for within-subject differences in ability and speed. In the van der Linden model, between-subject differences in ability level are captured by means of a continuous random latent ability variable θ_p, which underlies the item responses of respondent $p = 1, \ldots, N$ to item $i = 1, \ldots, J$.

Individual differences with respect to the speed with which the responses to the test items are given are modeled by a continuous random latent speed variable τ_p, which underlies the response times. In contrast to the van der Linden model, which assumes speed and ability to be constant within subjects, the hierarchical mixture modeling approach allows speed and ability to differ within subjects. Therefore, an item-specific latent class variable C_{pi} is assumed to underlie the response and the response time of person p to item i. Although the latent class variable C_{pi} can have K states in principle, it is commonly assumed to have two states, $K = 2$: In one state, the item properties of the faster responses are modeled and in the other state the item properties of the slower responses are modeled. Each respondent is allowed to switch between the slow and fast response state from item to item, such that within-subject differences are captured by the latent class variables.

To enable mixture modeling of both the responses and response times, a distributional assumption is needed for the within-state response time distribution. Correct specification of this distribution is important as it has been shown that violations of the assumed response time distribution may bias modeling results for mixture models in general (Vermunt, 2011), for growth mixture models (Bauer & Curran, 2003), and for the hierarchical mixture modeling framework for responses and response times as discussed above (Molenaar et al., 2018). More specifically, Molenaar et al. (2018) showed that if the observed response time distribution differs from the assumed distribution, within-state parameter estimates and information criteria like Akaike's Information Criterion (AIC; Akaike, 1974), the Bayesian Information Criterion (BIC; Schwarz, 1978), and the consistent AIC (CAIC; Bozdogan, 1987) will be biased, and spurious states will be detected even if there are no states underlying the data. Thus, specifying an appropriate within-state response time distribution is important. In practice, however, it is commonly unknown which type of distribution would fit the within-state response times best. Often, statistically convenient distributions are chosen, for example, the log-normal, the exponential, or the chi-square distribution. These distributions are considered convenient as respectively the logarithmic, the reciprocal, and the square root transformation will result in normally distributed response times. Once a distribution is chosen, this assumed distribution should ideally be tested to ensure that the mixture modeling results are valid. However, testing this distributional assumption is challenging as the within-state distribution is by definition different from the observed distribution since the latter is aggregated over states. That is, if the within-state log-transformed response time distribution is normal, the observed log-response time distribution will be skewed (assuming that the two states differ in their expected log-response time and their log-response time variance). As a result, it is not clear whether skewness in the observed log-transformed distribution reflects a mixture of two states or a misspecification of the response time distribution (Molenaar et al., 2018). Therefore, traditional statistical tests (e.g., the Shapiro–Wilk test [SW test], 1965) on the observed response time distribution cannot be used.

In this article, we propose statistical tests on the within-state response time distribution in the hierarchical mixture modeling framework for responses and response times. Specifically, we propose tests on normality of the transformed response time distribution using the SW test (Shapiro & Wilk, 1965; see also Royston, 1982a, 1982b, 1992) and the Kolmogorov–Smirnov test (KS test; Kolmogorov, 1933; Smirnov, 1948). Although the SW test and the KS test are well-established methods to test a hypothesized distribution to hold for an observed variable, the innovative aspect of the present study is that we apply these tests to investigate a hypothesized distribution to hold for the latent within-state distribution of the hierarchical mixture modeling framework for responses and response times. We focus on the log-transformation for the response times, making our approach a test on log-normality. We prefer to focus on log-normality as this is the most commonly used assumption in mixture modeling of response times. However, the proposed methodology can readily be used to test for other distributions by using a different response time transformation. That is, one can consider for instance the square root transformation to test for a chi-square distribution and the reciprocal transformation to test for an exponential distribution. In addition, for the KS test, it is straightforward to accommodate any other distribution (e.g., the Weibull, ex-Gaussian, or Wald distribution) as long as its cumulative distribution function exists and can be evaluated.

The proposed normality tests can be used for various types of (response time) mixture models; however, in this study, we apply the tests to the Markov-dependent item states model and the independent item states model (Molenaar et al., 2016). This article is organized as follows: First, we discuss the hierarchical mixture modeling approach with log-response times. Next, we present the normality tests for the within-state response time distribution. We then present a simulation study to investigate the performance of the different normality tests. In addition, we illustrate the use of the tests by means of a real data application, and we end with a general discussion.

2. The Hierarchical Mixture Modeling Approach

In the hierarchical mixture modeling approach (Molenaar et al., 2016; Schnipke & Scrams, 1997; C. Wang & Xu, 2015; C. Wang, Xu, & Shang, 2018), an item-specific latent class variable C_{pi} is assumed to underlie the response X_{pi} and the response time T_{pi} of person p to item i. The levels of this latent class variable are referred to as *item states*. The item states for respondent p on the items of a test can be collected in vector $\mathbf{c}_p = [C_{p1}, C_{p2}, \ldots, C_{pJ}]$. In addition, the item responses are collected in vector $\mathbf{x}_p = [X_{p1}, X_{p2}, \ldots, X_{pJ}]$, and the log-transformed response times are collected in vector $\mathbf{t}_p = [T_{p1}, T_{p2}, \ldots, T_{pJ}]$.

Although the latent class variable C_{pi} can have multiple states in principle, it is commonly assumed to have two states: a state capturing the measurement properties of the slower responses and a state capturing the measurement properties of the faster responses. The slower state, with a larger expected response time, is denoted by $C_{pi} = 0$, the faster state by $C_{pi} = 1$. As discussed above, the response times are assumed to follow a log-normal distribution in each state, such that the log-response times follow a normal distribution (e.g., C. Wang & Xu, 2015). In addition, the item responses and the log-response times are assumed to be independent conditional on the overall latent ability θ_p and the overall speed τ_p within the states $\mathbf{c}_p$.

In order to separate the effects of the item, the person and the latent class variable C_{pi}, a measurement model for the responses and a measurement model for the response times, is specified. Various models can be considered for modeling the responses, like the Rasch model (Loeys et al., 2014), the graded response model (e.g., Molenaar et al., 2015b; Ranger & Ortner, 2011), the two-parameter logistic model (e.g., Molenaar et al., 2015a, 2015b; Ranger & Ortner, 2012), or the three-parameter model (e.g., van der Linden, 2007). Here, we specify a two-parameter logistic model to model the item responses, that is

$$P(\mathbf{x}_p|\theta_p, \mathbf{c}_p) = \prod_{i=1}^{N} \omega(\alpha_{ki} \times \theta_p + \beta_{ki})^{x_{pi}} \times \omega\big(-[\alpha_{ki} \times \theta_p + \beta_{ki}]\big)^{1-x_{pi}}, \tag{1}$$

where $\omega(.)$ denotes the logistic function, and α_{ki} and β_{ki} denote the discrimination and easiness parameters for item i and state k. Note that the item response parameters are allowed to differ across states; however, they are treated as fixed parameters within each state.

Although other models have been proposed as well, a normal one-factor model is commonly assumed for modeling the log-response times (van der Linden, 2007). Here, we thus assume the log-response times to follow a conditional multivariate normal distribution. As follows, $f(\mathbf{t}_p|\tau_p, \theta_p, \mathbf{c}_p)$ models a person's response times, given their ability, latent states, and overall speed by means of

$$f(\mathbf{t}_p|\tau_p, \theta_p, \mathbf{c}_p) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2_{\varepsilon i}}} \times \exp\left(-\frac{1}{2}\frac{(T_{pi} - \mu_{pi})^2}{\sigma^2_{\varepsilon i}}\right), \tag{2}$$

with

$$\mu_{pi} = E(T_{pi}|\tau_p, \theta_p, \mathbf{c}_p) = v_i - \delta C_{pi} - \tau_p, \qquad \delta > 0, \tag{3}$$

where $\sigma^2_{\varepsilon i}$ is the residual log-response time variance, v_i is the time intensity parameter, and δ denotes the difference in expected speed between the slower and faster state. Here, the constraint $\delta > 0$ ensures that $C_{pi} = 1$ corresponds to the faster state, which is the state with the smaller response times.

Next, by assuming a bivariate normal distribution for the latent ability variable θ_p and the latent speed variable τ_p, the general log marginal likelihood of item response vector $\mathbf{x}_p$ and the log-response time vector $\mathbf{t}_p$, given the model parameter vector $\boldsymbol{\eta}$, is given by

$$\ell(\mathbf{x}_p, \mathbf{t}_p; \boldsymbol{\eta}) = \ln \iint_{-\infty} \sum_{C_{p1}}^{K-1} \sum_{C_{p2}}^{K-1} \cdots \sum_{C_{pJ}}^{K-1} P(\mathbf{x}_p|\theta_p, \mathbf{c}_p) \times f(\mathbf{t}_p|\tau_p, \theta_p, \mathbf{c}_p) P(\mathbf{c}_p) \times g(\theta_p, \tau_p) \mathbf{d}\theta \mathbf{d}\tau, \tag{4}$$

where $g(.)$ denotes the bivariate normal density function. $P(\mathbf{c}_p)$ denotes the state probabilities for a set of items, which, for each particular item, models the probability that the response of person p to item i belongs to a given state. As follows, $P(\mathbf{c}_p) = \left[P(C_{p1}), P(C_{p2}), \ldots, P(C_{pJ})\right]$. $P(\mathbf{x}_p|\theta_p, \mathbf{c}_p)$ thus models the probability of obtaining a particular response pattern, given a person's ability and their latent states $\mathbf{c}_p$ for the different items, and $f(\mathbf{t}_p|\tau_p, \theta_p, \mathbf{c}_p)$ models a person's response times.

In the hierarchical mixture modeling framework, the item-specific latent class variables are commonly assumed to be independent from item to item. However, in practice, these latent class variables may be dependent, for instance, if a respondent guesses on one item, they may be more likely to guess on the next item. To account for such a possible dependency between the latent class variables in the hierarchical mixture framework, Molenaar et al. (2016) also considered a model with a time homogenous first-order Markov-structure on the item-specific latent class variables. As a result, in this model, $P(C_{pi})$, the state a person is in regarding a certain item, depends on the state of the previous item. That is, the state probability $P(C_{pi})$ is decomposed as follows:

$$P(C_{pi}) = P(C_{p1}) \times P(C_{pi}|C_{p(i-1)}), \tag{5}$$

where $P(C_{p1})$ is referred to as the initial state probability and $P(C_{pi}|C_{p(i-1)})$ is the so-called transition probability. Since the number of states in our model equals two, there are two initial state probabilities. In addition, as the Markov dependency is assumed to be time homogenous, the transition probabilities are assumed equal for all subsequent items, resulting in four transition probabilities. In model estimation, only one initial state probability and two transition probabilities have to be estimated as the others follow from these estimates (see Figure 1 for a graphical representation of the model).

3. Normality Tests

In the model discussed above, the log-response times within each state are assumed to follow a normal distribution. To test this assumption, we use the SW test and the KS test. Specifically, we propose the following procedure: First, the hierarchical mixture model is fit to the responses and response times. Next, the resulting posterior state probabilities are obtained for each response which in turn are used to draw posterior state assignments to state 0 or 1 for each person's

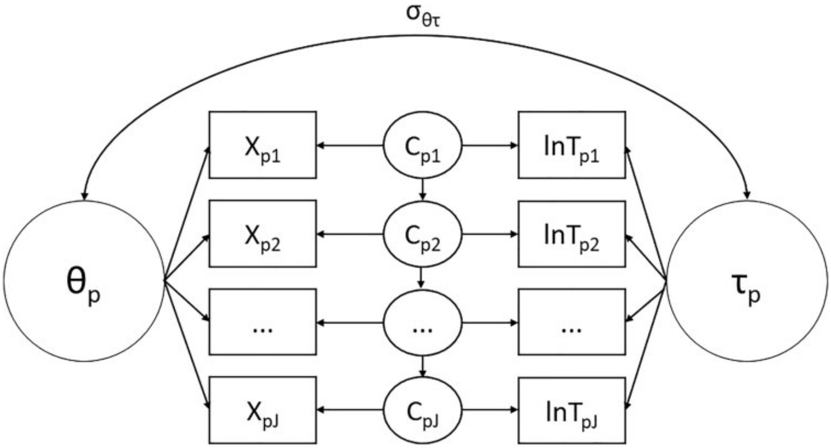

FIGURE 1. *Markov-dependent item states model.*

response to each item. Then the normality tests are conducted on either (1) the response times in state 0 and state 1 according to the posterior state assignment or (2) on the response times weighted by the posterior state probabilities.

The rationale for the above procedure is that if the within-state log-response time distribution is correctly specified, the resulting posterior state probabilities and posterior state assignments are correct. As a result, the SW and KS test statistics will follow their theoretical distribution under the null hypothesis of normality. However, if the within-state log-response time distribution is incorrectly specified, the resulting posterior state probabilities and posterior state assignments are wrong, and the SW and KS test statistics will not follow their null distributions. Below we discuss the SW and KS tests and apply them to the mixture modeling framework.

3.1. SW Test

The SW test, also called the W test for normality, tests the null hypothesis that an observed variable comes from a normally distributed population. Since the original test as proposed by Shapiro and Wilk (1965) could not be used for sample sizes larger than 50, Royston (1982b, 1992) extended it to sample sizes up to 2,000. Then, suppose that $x_1 < x_2 < \ .... < x_N$ is an ordered sample of size N on which the normality test is carried out. The W statistic, as calculated for an item $i = 1, \ldots, J$, is defined as

$$W_i = \frac{\left(\sum a_p x_p\right)^2}{\sum \left(x_p - \bar{x}\right)^2}, \tag{6}$$

where $a = a_1, a_2, ..., a_N$ are the normalized best linear unbiased coefficients, and $p = 1, \ldots, N$. Vector **a** is defined by

$$\mathbf{a} = (\mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m})^{-\frac{1}{2}}\mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1}, \tag{7}$$

where $\mathbf{m} = m_1, \ldots, m_N$ is a vector of expected values of standard normal order statistics, and $\mathbf{V}$ is the corresponding covariance matrix. Note that vector **a** is antisymmetric, that is, $a_N = -a_1$, and for odd N, $a_{(N/2)+1} = 0$ (Royston, 1992). More details on the calculation of the W_i statistic can be found in the Appendix.

3.2. KS Test

The KS test is a nonparametric goodness of fit test, which measures the distance between an empirical distribution function of a sample and a hypothesized cumulative distribution function (the one-sample KS test) or which compares the distribution functions of two samples (the two-sample KS test). Here, we will focus on the one-sample KS test for testing the within-state response time distribution of item i for normality, which is defined by

$$D_{Ni} = \sup_x |F_N(x) - F(x)|, \tag{8}$$

where $F_N(x)$ is the empirical distribution function, and $F(x)$ the reference distribution. In this article, we assume a normal distribution for the log-response times; however, other distributions like the Wald distribution can be used as a reference distribution as well. Therefore, the KS test can not only be used to test the within-state response time distribution for normality but can be used to test for all kinds of distributions. As Monahan (2011) noted, the Kolmogorov–Smirnov statistic is distribution free for continuous random variables. As a result, the distribution function is evaluated at the observations, and then sorted, such that

$$D_{Ni} = \max_p \{p/N - F(X_{(p)}), F(X_{(p)}) - (p-1)/N\}, \tag{9}$$

where X_p denotes the response time of respondent p on item i, with $p = 1, \ldots, N$; and $X_{(1)} \leq \ldots \leq X_{(p)} \leq \ldots \leq X_{(N)}$ denote the ordered response times of a sample of respondents. The accompanying p value for the statistic needs to be bootstrapped, since the resulting sampling distribution is unknown, and therefore, the mean and standard deviation are unknown and thus must be estimated from the data when testing for normality (Lilliefors, 1967).[1] When testing for other types of distributions, estimating only the mean and standard deviation might not be sufficient, and additional parameters might need to be estimated (e.g., a skewness parameter).

Since the posterior state probabilities differ for each respondent on each item they are taken into account by including them as weights in the KS test, Equation 9 needs to be modified to account for the weighted response times (Monahan, 2011). The empirical distribution function can be estimated by

$$F_N(x) = \frac{1}{N}\sum_{p=1}^{N} P(C_{pi}|\,\mathbf{x}_p,\,\mathbf{t}_p) I(X_p \leq x) \Big/ \frac{1}{N}\sum_{p=1}^{N} P(C_{pi}|\,\mathbf{x}_p,\,\mathbf{t}_p) \tag{10}$$

where $P(C_{pi}|\mathbf{x}_p, \mathbf{t}_p)$ are the posterior state probabilities, and $I(X_p \leq x)$ is an indicator function, which equals 1 if $X_p \leq x$ and 0 otherwise. Then, the X_p must be sorted together with the corresponding posterior state probabilities, the increments can be computed by

$$F_k = \sum_{p=1}^{k} P(C_{pi}|\,\mathbf{x}_p,\,\mathbf{t}_p) \Big/ \sum_{p=1}^{N} P(C_{pi}|\,\mathbf{x}_p,\,\mathbf{t}_p). \tag{11}$$

As follows, $F_N(x) = F_k$ for $X_{k-1} < x \leq X_k$, which can be calculated by estimating quantiles for the underlying normal distribution, weighted by the posterior state probabilities, that correspond to N ordered probabilities. Using F_k from Equation 11, the Kolmogorov–Smirnov statistic for item i modifies to

$$D_{Ni} = \sup_x |F_N(x) - F(x)| = \max_k \left\{F_k - F(X_{(k)}), F(X_{(k)}) - F_{k-1}\right\}. \tag{12}$$

Like in the unweighted case, the p value of D_{Ni} needs to be bootstrapped, since the mean and standard deviation of the log-response time distribution are estimated from the data (Lilliefors, 1967).

4. Simulation Study

4.1. Method

In the simulation study, we compared the performance of the unweighted SW test, the unweighted KS test, and the weighted KS test. Data are simulated according to nine different scenarios, mostly based on Molenaar et al. (2018) who found biased modeling results of the hierarchical mixture model in the case of nonnormality. The first three scenarios concern Markov mixture models that include Markov-dependent item states, the next three scenarios concern mixture models with independent item states, and the final three scenarios are generated according to a baseline model that does not include item states (i.e., a static model without mixtures). The scenarios differ in the distribution that is used for the log-response times, which is a normal, a truncated, or a skewed distribution. The responses are modeled using the two-parameter logistic model, with item parameters α_{ki} and β_{ki}, which denote the discrimination and easiness of an item i in state k, respectively. For each scenario, we conducted 50 replications with 20 items, the sample size was equal to 500 or 1,000.

The three Markov mixture model scenarios are the following:

Normal Markov mixture: In this scenario, we use the Markov-dependent item states model with normal log-response times to simulate the data. We use $\delta = .1$. That is, the expected log-response times differ by 0.1 between the fast and slow state.

The item parameters for the two different states are set as follows: the discrimination parameters are set to $\alpha_{0i} = 1.5$ for State 0 (slow state) and to $\alpha_{1i} = 1$ for State 1 (fast state). We set the easiness parameters to increasing, equally spaced values: for the slower state, β_{0i} is between -2 and 0, for the faster state β_{1i} is between 0 and 2. The response time parameters for all items i are chosen as $v_i = 2$, and the residual response time variances as $\sigma^2_{\varepsilon i} = 0.13$. Furthermore, $\sigma^2_\tau = 0.13$ and $\sigma^2_{\theta\tau} = 0.14$, so that the correlation between θ_p and τ_p equals 0.4. Finally, the initial state probability, $P(C_{p1} = 1)$, is fixed to 0.5, and the transition probabilities, $P(C_{pi} = 1|C_{p(i-1)} = 0)$ and $P(C_{pi} = 0|C_{p(i-1)} = 1)$, are fixed to .231, which corresponds to mildly instable states according to Bacci et al. (2014).

Truncated Markov mixture: In this scenario, the data are generated using the same parameter values as in the normal Markov mixture scenario above. However, now we use a right-truncated normal distribution for the log-response times, with truncation at $\ln(12)$ such that the response times are cut off at 12 seconds. This scenario mirrors data from time-pressured tests, where respondents only have a limited amount of time to answer the items.

Skewed mixture: Here, the data are generated using the same parameter values as in the normal Markov mixture scenario. However, the normally distributed log-response times are transformed using a Box–Cox transformation (Box & Cox, 1964). In general, the transformation is used to transform skewed variables in such a way that they are closer to a normal distribution. Here, we use the transformation the other way around, so that we transform the normally distributed log-response times into skewed variables using the Box–Cox transformation. That is, we transform the normally distributed log-response times using $\ln(T_{pi})' = (\lambda \ln(T_{pi}) + 1)^\lambda$ with $\lambda = .3$.

Figure 2 shows the resulting log-response time distribution of an arbitrary item from an example run of the three different Markov mixture model scenarios. Note that in the truncated scenario, the log-response times are negatively skewed, while in the skewed scenario, the log-response times are positively skewed.

In the three independent mixture scenarios, we used the same parameter values and setup as for the Markov mixture scenarios above; however, the item states are assumed to be independent, that is, the Markov structure is omitted (i.e., $P(C_{pi} = 1|C_{p(i-1)} = 0) = P(C_{pi} = 1)$ and $P(C_{pi} = 0|C_{p(i-1)} = 1) = P(C_{pi} = 0)$). Next, for the three baseline scenarios, the data do not include item states, and a baseline model (i.e., the traditional hierarchical model without mixtures by van der Linden, 2007) is used to generate the data. Like the mixture models, the scenarios differ in the distribution used for the log-response times, which again are either a normal, a truncated or a skewed distribution:

Normal baseline: In this scenario, the log-response times are normally distributed. The item parameters are as follows: The discrimination parameters for all items are set to $\alpha_i = 1$, the easiness parameters β_i are set to increasing, equally spaced

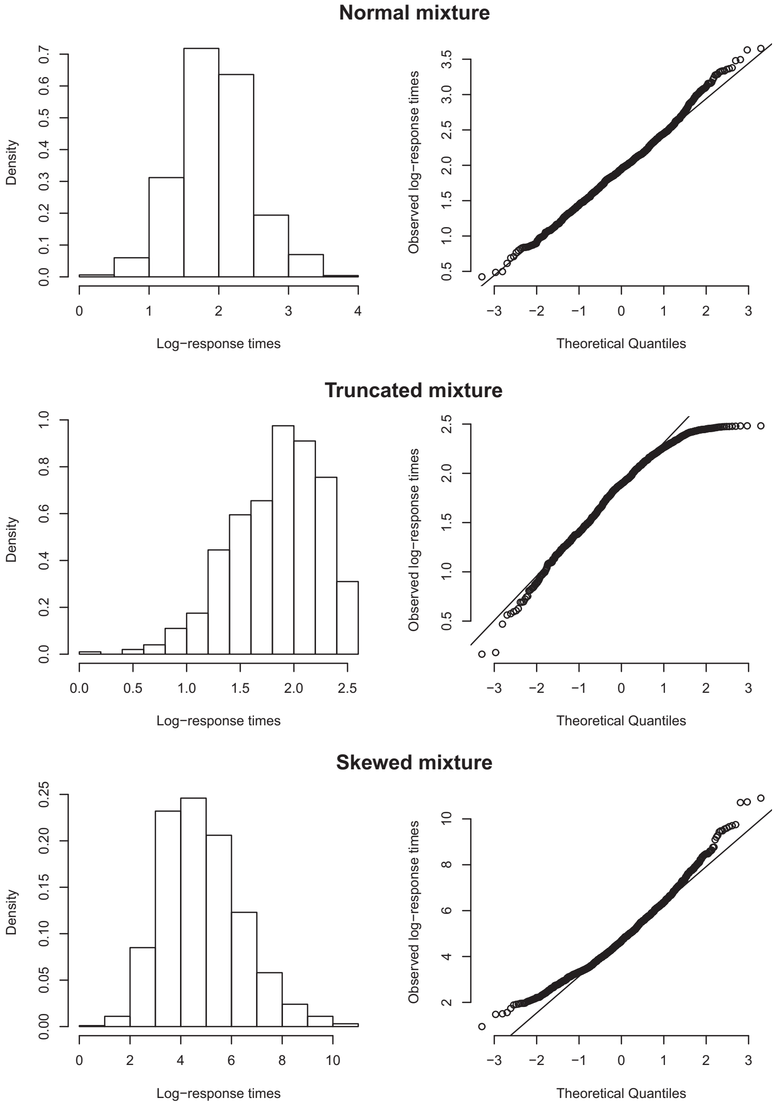

FIGURE 2. *Example run for a random item for normal, truncated, and skewed mixture model scenarios.*

values between −2 and 2. The response time parameters v_i, $\sigma^2_{\varepsilon i}$, σ^2_{τ}, and $\sigma^2_{\theta\tau}$ are set to the same values as in the normal mixture model scenario.

Truncated baseline: Here, the data are generated using the same parameter values as in the normal baseline scenario. Like in the truncated mixture model scenario, we

TABLE 1.
Mean Number of Times That the Normality Hypothesis Is Rejected Over All Items Within a State for the Normal Scenarios

		SW Test		KS Unweighted		KS Weighted	
N		State 0	State 1	State 0	State 1	State 0	State 1
500	Markov mixture	.054	.057	.048	.055	.093	.122
	Independent mixture	.047	.049	.050	.044	.102	.105
	Baseline model	.051	.054	.044	.046	.081	.127
1,000	Markov mixture	.045	.050	.063	.065	.091	.092
	Independent mixture	.048	.057	.052	.044	.072	.089
	Baseline model	.054	.045	.053	.043	.075	.093

use a right-truncated normal distribution for the log-response times, with truncation at ln(12), such that the response times are cut off at 12 seconds.

Skewed baseline: Here, the data are generated using the same parameter values as in the normal baseline scenario. However, like the skewed mixture model scenario, the normally distributed log-response times are transformed using a reverse Box–Cox transformation (Box & Cox, 1964), with the same value for λ.

5. Results

For the truncated scenarios, convergence problems occurred in 17 replications (4 in the truncated Markov mixture, 9 in the truncated independent mixture, and 4 in the truncated baseline scenario). In the results below, those replications are excluded. Tables 1, 2, and 3 contain the results for the normal, truncated, and skewed scenarios, respectively. Specifically, the tables depict the proportion of items for which the null hypothesis of normality is rejected at a .05 level of significance, averaged over all items within a state. That is, a proportion of .900 for a given state indicates that—averaged over the replications—the normality hypothesis is rejected for 90% of all items within that state.

First, for the normal scenarios in Table 1, these proportions indicate the Type I error rate of our approach. Ideally, these rates are close to the level of significance for the tests to be viable. As can be seen from the table, for the Markov mixture, the independent mixture, and the baseline scenarios, the results are similar. That is, the SW test, and the unweighted KS test have acceptable Type I error rates. In addition, the weighted KS test is associated with an inflated Type I error rate.

For the truncated and skewed scenarios, the mean proportion of normality rejections reflects the power to detect within-state departures from normality. Here, we used Cohen's (1988, p. 56) rule of thumb, considering a power coefficient of .80 or higher to be acceptable. As can be seen from Tables 2 and 3, the

TABLE 2.
Mean Number of Times That the Normality Hypothesis Is Rejected Over All Items Within a State for the Truncated Scenarios

		SW Test		KS Unweighted		KS Weighted	
N		State 0	State 1	State 0	State 1	State 0	State 1
500	Markov mixture	1.000	.163	0.999	.110	0.998	.157
	Independent mixture	1.000	.159	1.000	.107	0.999	.169
	Baseline model	1.000	.194	1.000	.122	1.000	.163
1,000	Markov mixture	1.000	.252	1.000	.175	1.000	.211
	Independent mixture	1.000	.279	1.000	.180	1.000	.212
	Baseline model	1.000	.315	1.000	.209	1.000	.232

TABLE 3.
Mean Number of Times That the Normality Hypothesis Is Rejected Over All Items Within a State for the Skewed Scenarios

		SW Test		KS Unweighted		KS Weighted	
N		State 0	State 1	State 0	State 1	State 0	State 1
500	Markov mixture	.191	0.935	.144	.631	.239	.681
	Independent mixture	.165	0.952	.098	.662	.212	.706
	Baseline model	.143	0.946	.079	.658	.191	.687
1,000	Markov mixture	.256	0.998	.150	.911	.254	.913
	Independent mixture	.255	1.000	.140	.922	.267	.918
	Baseline model	.246	0.996	.160	.917	.271	.925

results indicate that generally the power is acceptable in one state and substantially smaller in the other state. In the truncated scenario, State 0 is associated with larger power, while in the skewed scenario, State 1 is associated with larger power as compared to State 0. This can be explained from Table 4 which contains the average initial state parameter estimates for State 1, together with the transition parameters in the different scenarios (i.e., the table reflects the proportions of persons in the different states). As can be seen from the mean initial state parameter estimates for State 1, in the truncated scenario, State 0 (slower state; larger log-response times) is the larger state, and in the skewed scenario, State 1 (faster state; smaller log-response times) is the larger state. This is due to the log-response times being positively skewed in the skewed scenario and negatively skewed in the truncated scenario (as mentioned above). As a result, due to these larger sample sizes in State 0 for the truncated scenarios and State 1 for the skewed scenarios, power differs between the two states. That is, when fitting a

TABLE 4.
*Mean Estimates (*SD*) of the Initial State Probabilities and the Transition Probabilities*

		Initial State	Transition 0–1	Transition 1–0
Normal scenarios				
$N = 500$	Markov mixture	.454 (.169)	.242 (.149)	.209 (.169)
	Independent mixture	.462 (.137)	.467 (.169)	.491 (.163)
	Baseline	.463 (.197)	.385 (.277)	.405 (.260)
$N = 1{,}000$	Markov mixture	.418 (.154)	.202 (.045)	.200 (.054)
	Independent mixture	.492 (.135)	.474 (.132)	.477 (.132)
	Baseline	.461 (.157)	.325 (.237)	.414 (.260)
Truncated scenarios				
$N = 500$	Markov mixture	.144 (.050)	.196 (.017)	.703 (.023)
	Independent mixture	.145 (.052)	.198 (.017)	.719 (.034)
	Baseline	.219 (.050)	.191 (.018)	.735 (.028)
$N = 1{,}000$	Markov mixture	.150 (.034)	.193 (.012)	.704 (.018)
	Independent mixture	.147 (.038)	.199 (.014)	.721 (.023)
	Baseline	.213 (.041)	.186 (.013)	.724 (.018)
Skewed scenarios				
$N = 500$	Markov mixture	.793 (.266)	.788 (.167)	.091 (.030)
	Independent mixture	.865 (.181)	.828 (.028)	.096 (.015)
	Baseline	.845 (.218)	.845 (.036)	.087 (.014)
$N = 1{,}000$	Markov mixture	.856 (.182)	.804 (.085)	.091 (.018)
	Independent mixture	.893 (.056)	.815 (.115)	.091 (.014)
	Baseline	.827 (.234)	.820 (.110)	.097 (.044)

normal mixture to nonnormal data, the nonnormality is best detected in the largest state. Furthermore, even though class sizes are comparable, power tends to be larger in the larger state of the truncated scenarios when compared to the larger state of the skewed scenarios. This is due to the fact that the truncated distribution departs more from normality than the skewed distribution (see Figure 2). In general, the power of the KS tests (weighted and unweighted) is smaller as compared to the SW test. We return to this point in the discussion. Furthermore, the weighted KS test has slightly more power as compared to the unweighted KS test (but is also associated with an increased Type I error rate, see above).

5.1. Conclusion

Taken together the above, Type I error rate and the power of the proposed tests seem acceptable for the SW test and the unweighted KS test with more power for the SW test. The weighted KS test is associated with an inflated Type I error rate.

There are no important differences between the Markov mixtures, independent mixtures, and baseline model. It turns out that, generally, violations of normality are only detected in one of the states. We note that of course the power of our approach depends on the severity of the normality violations (this is why the power seems somewhat larger in the truncated scenario as compared to the skewed scenario: the data in the truncated scenario is heavier skewed). In that sense, we consider our simulation study as a prove of principle (i.e., given the effect size we have chosen, we demonstrated that the approach is viable).

Our results indicate that nonnormality is detected if the data contain nonnormal mixtures (i.e., the Markov mixture and independent mixture scenarios) or if the data are nonnormal without mixtures (i.e., the baseline scenarios). In practice, where one does not know the data generating model, a significant normality test thus indicates that (1) the data follow a mixture model with a nonnormal within-state distribution or (2) the data are nonnormal but do not contain mixtures. For the present purposes, the distinction between (1) and (2) is not of importance as the implications are the same: In both cases, there is no mixture of normal distributions in the data, so the results of a normal (Markov-)mixture model should not be trusted. If our proposed tests are insignificant in both states, it can safely be concluded that (1) the data follow a mixture model with a normal within-state distribution or (2) the data are normal without mixtures (i.e., the baseline scenario's). As in both cases, the (within-state) data are normal, (1) and (2) can be distinguished by comparing the baseline model and the mixture models using common information criteria (e.g., BIC and CAIC) as demonstrated by Molenaar et al. (2018). Therefore, we propose the procedure summarized in the flow chart in Figure 3. That is, first, the fit of a normal (Markov-)mixture model is compared to that of a normal baseline model. If the baseline model fits better, it can be concluded that the transformed response times are normally distributed and that there is no mixture of normal distributions underlying the data. If the mixture model fits better, one can consult the statistics proposed in this article. If these statistics are insignificant in both states, it can be concluded that there is a true mixture of normal distributions underlying the response time data, and the results of the mixture model can be validly interpreted. However, if the proposed statistics are significant, it can be concluded that there is no mixture of normal distributions underlying the data, and the results of the normal (Markov-)mixture model cannot be trusted.

6. Application

The within-state normality tests are illustrated by means of a real-data application. The data consist of the responses and response times of 389 psychology freshman of the University of Amsterdam to 28 items of the knowledge subtest of the Dutch version of the Intelligence Structure Test (Amthauer et al., 2001). The knowledge subtest measures essential types of knowledge, which people acquire

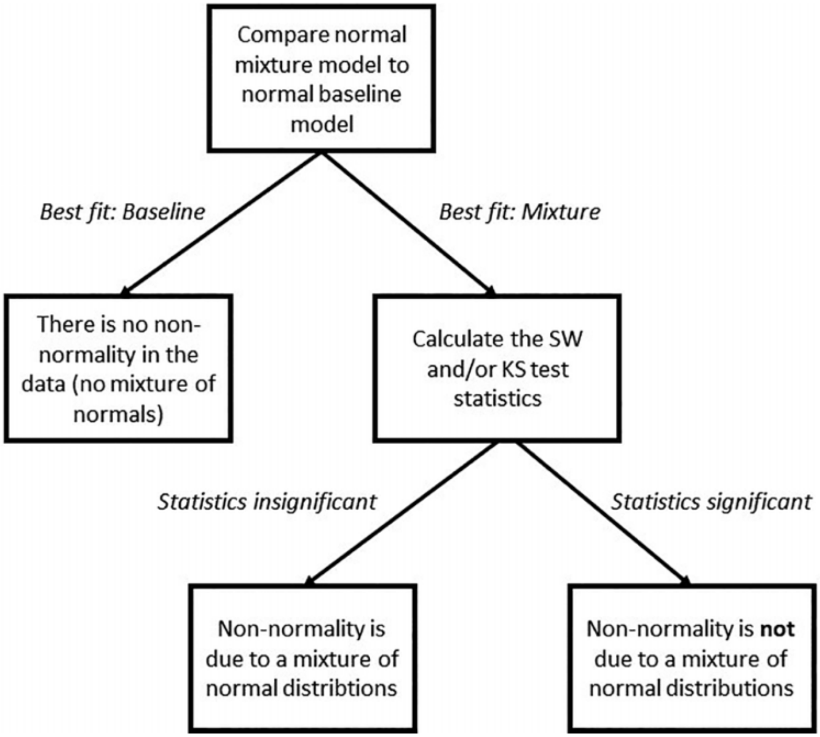
Compare normal mixture model to normal baseline model
Best fit: Baseline
Best fit: Mixture
There is no non-normality in the data (no mixture of normals)
Calculate the SW and/or KS test statistics
Statistics insignificant
Statistics significant
Non-normality is due to a mixture of normal distribtions
Non-normality is **not** due to a mixture of normal distributions

FIGURE 3. *A flowchart of the proposed procedure using the Shapiro–Wilk or Kolmogorov–Smirnov statistics.*

in schools, higher education and other educational institutions, as well as daily life knowledge acquired from within their culture (Hogrefe Ltd., 2016). The items in the subtest cover a broad range of topics, like economics, geography, mathematics, history, art and culture, natural sciences, and daily life facts (de Vries, 2017). The items are dichotomously scored, with 0 indicating a false response and 1 indicating a correct response. Looking more closely at the response time distributions, Figures 4 and 5 show for a selection of items that they are not normally distributed. The observed response time distributions seem skewed, and the question is whether this nonnormality can be explained by a mixture of a fast and a slow state or whether there is an alternative explanation.

In the paper by Molenaar et al. (2016) mixture models with various types of Markov dependencies are fitted to the data and are shown to fit better than a baseline model without item states. However, as the modeling results of a fitting mixture model with a Markov dependency are only interpretable if the assumed normal distribution for the log-response times holds, we test this assumption using the proposed methodology.

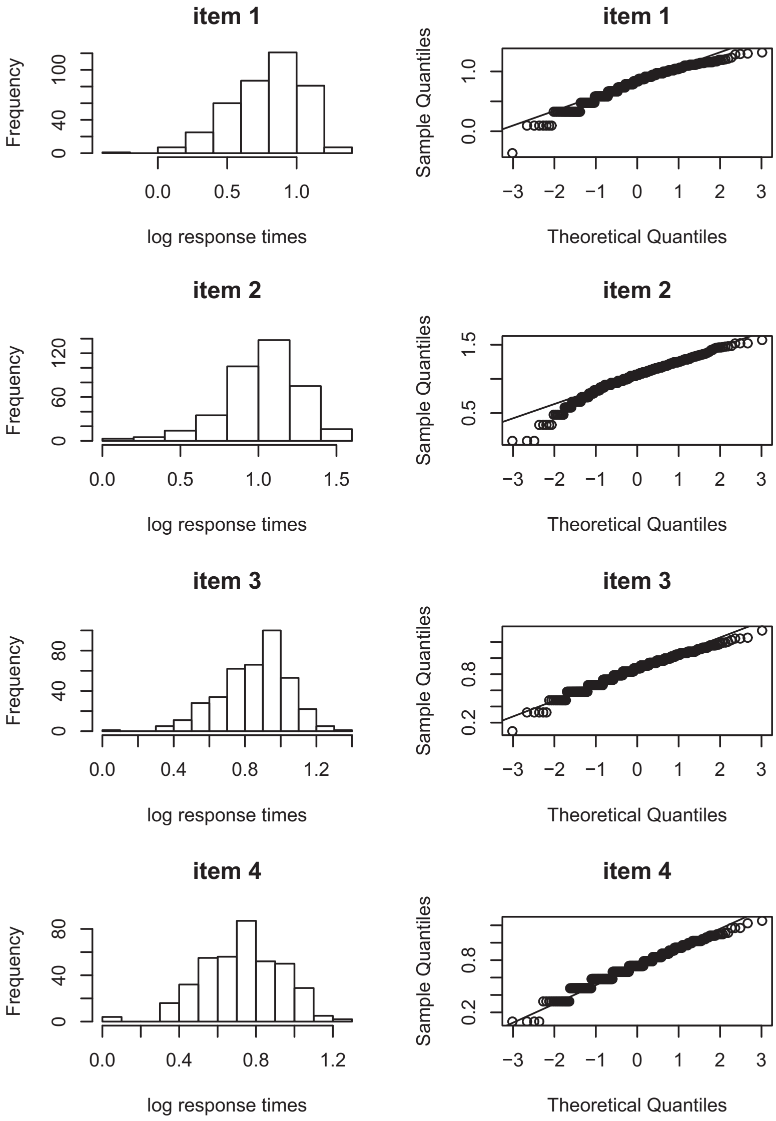

FIGURE 4. *Item score distributions for items 1–4.*

Using a significance level of $\alpha = .05$, Table 5 shows that especially in State 1, the fast state, the continuous response times for most of the 28 items do not follow a normal distribution. Furthermore, all three tests indicate that for State 0, the slow state, the response times for the majority of the items are normally

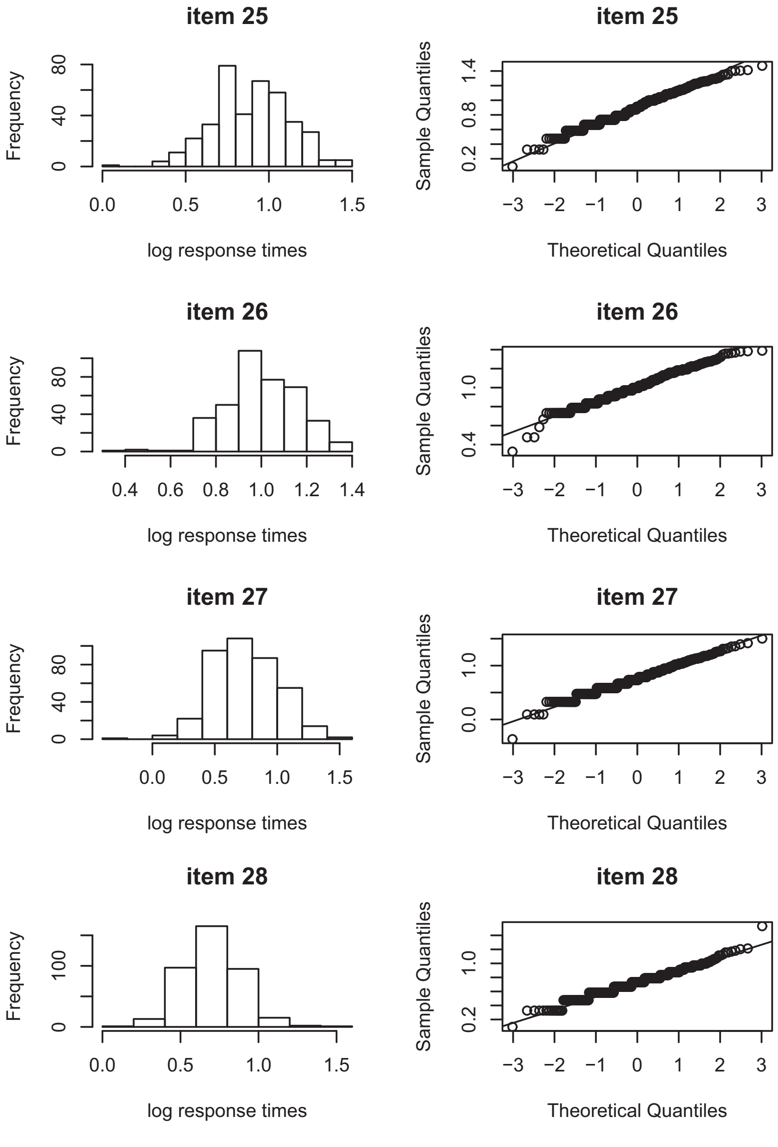

FIGURE 5. *Item score distributions for items 25–28.*

distributed. However, since normality does not hold for the majority of the items in State 1, the parameter estimates are biased and states detected in the data can be spurious. Therefore, we cannot use a mixture model with a Markov dependency to interpret the responses and response times, and we need to fit an alternative model in order to explain the data.

TABLE 5.
Number of Nonnormal Items Within Each State

	State 0	State 1
SW test	8	26
KS unweighted	4	27
KS weighted	5	26

Note. SW = Shapiro–Wilk; KS = Kolmogorov–Smirnov.

7. Discussion

If the within-state response time distribution in the hierarchical mixture modeling framework is misspecified, parameter estimates and model fit indices will be biased and spurious states can be detected in the data as a result. Therefore, in this article, we proposed statistical tests for normality of the within-state log-response time distribution.

In a simulation study, we found that violations of nonnormality can successfully be detected using our tests based on the SW and KS tests. Most importantly, our test has demonstrated an acceptable Type I error rate, which indicates that our test can be used to successfully identify situations where normality holds, and where model fit indices like BIC and CAIC can successfully be used to test between models that include and do not include normal mixtures. If normality is violated, it cannot be concluded whether the violations of normality are due to within-state nonnormality or due to observed nonnormality. However, in such cases, normal mixture modeling should not be adopted anyway, and our test is shown to be a good indicator for such situations.

We also found that the weighted and unweighted KS tests had a slightly increased Type I error rate. In addition, the SW test was associated with larger power as compared to the KS tests. This is line with for instance Razali and Wah (2011), Stephens (1974), and Yap and Sim (2011) who all noted that the KS test in general tends to have smaller power than the SW test. In addition, Shapiro et al. (1968) furthermore showed that in the case of misspecifying the parameters of the hypothesized null distribution, the power and Type I error can be influenced. Type I error rates at the 5% significance level can increase to 61% for a sample size of 50, and the effect becomes more pronounced when sample size increases. Monahan (2011) on the other hand noted that since the KS tests are less powerful, it should not be used in small samples but could be used in larger samples. Although the SW test thus seems preferable over the KS test in the present study, the KS test is more flexible as it can be used to test any assumed distribution, where the SW test can only be used on distributions that can be transformed to a normal one.

As noted above, if in practice, normality is rejected, it is advisable to not interpret the modeling results since they are unreliable. An alternative in that case may be to categorize the response times and use a model for categorical variables as is shown in the application. As Molenaar et al. (2018) showed, such an approach hardly produces parameter bias and false positives with respect to states underlying the data. Furthermore, the approach is comparable to the parametric within-subject mixture modeling approach regarding power. A second solution would be to use a nonparametric or semi-parameter alternative for modeling the responses and response times. However, such an approach is yet to be developed. The semi-parametric models by C. Wang, Fan, et al. (2013) and C. Wang, Chang, and Douglas (2013) can possibly provide a point of departure.

In this article, we have assumed a Markov structure for the dependency between the states. However, the present approach to test for normality is equally amenable to mixture models without Markov structure (e.g., the model by C. Wang & Xu, 2015). In the simulation study, we did not find any important differences between the scenario's in which the data included a Markov structure and scenario's in which the data did not included a Markov structure. However, if the Markov dependency becomes stronger (i.e., higher transition probabilities), power for a Markov model may be larger.

Another aspect of the present approach is that of the fit of the mixture model under consideration. That is, we tested for normality by assuming a certain mixture model (in this case, a Markov-dependent item states model). We showed that minor departures from normality can be detected if the model is otherwise correctly specified. If the initial mixture model is misspecified, false positives may arise. The severity of this inflation will depend on the size of the misfit. We therefore think that, first, in practice, were—hopefully—a researcher has a well (theoretically) motivated model that is not severely misspecified, the consequences will not be large. Second, the consequences of this inflation do not have serious consequences as also discussed with respect to the baseline model above. That is, if the tests proposed in the present article are significant, the conclusion should be that the results of the mixture model cannot be trusted.

The normality tests presented in this article are all tests for univariate normality. Testing normality thus needs to be conducted on an item-by-item basis, so in practice, a correction for multiple testing is appropriate (e.g., a Bonferroni or Bonferroni–Holm correction). A focus of future research may be the development of an overall test for multivariate normality, which considers all items at once. Then, the present univariate tests can be used as post hoc tests to investigate whether individual items are responsible for violations of the assumed distribution.

Appendix

Even though **m** and **V** in vector **a** can be computed using various algorithms (see Royston, 1982a, 1982b; also see Davies & Stephens, 1978; Shea &

Scallan, 1988), Royston (1992, also see Verrill & Johnson, 1988) offers a close approximation for vector **a** which is based on the Weisberg and Bingham (1975) statistic $\left(\sum b_p x_p\right)^2 / \sum \left(x_p - \bar{x}\right)^2$, where $b_p = (\tilde{\mathbf{m}}^T \tilde{\mathbf{m}})^{-\frac{1}{2}} \tilde{m}_p$, $\tilde{m}_p = \Phi^{-1}\{(p - 3/8)/(N + 1/4)\}$, and Φ is the normal cumulative distribution function. Only for the first two components, and the last two since **a** is antisymmetric, **b** differs from **a**. The value for the last two (and first two) components from **a** can, for $4 \leq N \leq 1,000$, be approximated by

$$\tilde{a}_N = b_N + 0.221157y - 0.147981y^2 - 2.071190y^3 + 4.434685y^4 - 2.706056y^5, \quad \text{(A1)}$$

$$\tilde{a}_{N-1} = b_{N-1} + 0.042981y - 0.293762y^2 - 1.752461y^3 + 5.682633y^4 - 3.582663y^5, \quad \text{(A2)}$$

where $y = N^{-\frac{1}{2}}$. The remaining a_p are approximated by

$$\tilde{a}_p = \phi^{-\frac{1}{2}} \tilde{m}_p, \quad \text{(A3)}$$

for $p = 2, \ldots, N - 1$ when $N \leq 5$ and $p = 3, \ldots, N - 2$ when $N > 5$. In Equation A3, the $\tilde{m}_p$ are normalized by using

$$\phi = (\tilde{\mathbf{m}}^T \tilde{\mathbf{m}} - 2\tilde{m}_N^2)/(1 - 2\tilde{a}_N^2) \quad \text{if} \quad N \leq 5, \quad \text{(A4)}$$

$$= (\tilde{\mathbf{m}}^T \tilde{\mathbf{m}} - 2\tilde{m}_N^2 - 2\tilde{m}_{N-1}^2)/(1 - 2\tilde{a}_N^2 - 2\tilde{a}_{N-1}^2) \quad \text{if} \quad N > 5. \quad \text{(A5)}$$

After calculation of the approximated **a**, the W_i statistic can be computed using Equation 6. Royston (1992) showed that small values of W_i indicated nonnormality, therefore, W_i needs to be normalized. For $4 \leq N \leq 11$, the transformed W_i (denoted by w), the mean μ and standard deviation σ are defined by

$$w = -\ln, [\gamma - \ln(1 - W_i)],$$

$$\gamma = -2.273 + 0.459N,$$

$$\mu = 0.5440 - 0.39978N + 0.025054N^2 - 0.0006714N^3,$$

$$\sigma = \exp(1.3822 - 0.77857N + 0.062767N^2 - 0.0020322N^3).$$

For $12 \leq N \leq 2{,}000$, they equal

$$w = \ln(1 - W_i),$$

$$x = \ln N,$$

$$\mu = -1.5861 - 0.31082x - 0.083751x^2 + 0.0038915x^3,$$

$$\sigma = \exp(-0.4803 - 0.082676x + 0.0030302x^2).$$

As follows, the p value of W_i can be found by using $z = (w - \mu)/\sigma$, which corresponds to the upper tail of $N(0, 1)$ if $z > 0$ and to the lower tail if $z < 0$.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was supported by a grant from the Dutch Research Council (NWO): VENI-451-15-008.

ORCID iD

Renske E. Kuijpers https://orcid.org/0000-0002-2268-3995
Ingmar Visser https://orcid.org/0000-0003-3855-2778

Note

[1] In the bootstrap procedure, we obtained the p value of the Kolmogorov–Smirnov (KS) statistic ($\mathrm{KS}_{\mathrm{data}}$) by randomly drawing 100 normal variables, standardizing these, and calculating the KS statistics ($\mathrm{KS}_{\mathrm{sampling}}$). The p value is than obtained by calculating the proportion of samples in which the $\mathrm{KS}_{\mathrm{sample}}$ exceeds $\mathrm{KS}_{\mathrm{data}}$.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-struktur-test 2000 R* [Intelligence-structure-test 2000 R]. Hogrefe.

Bacci, S., Pandolfi, S., & Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, *8*, 125–145.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*(3), 338–363.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, *26*, 211–252.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370. https://doi.org/10.1007/BF02294361

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Davies, C. S., & Stephens, M. A. (1978). Algorithm AS128. Approximating the covariance matrix of normal order statistics. *Applied Statistics*, *27*, 206–212.

de Vries, J. (2017). *IST: Intelligentie Structuur Test* [Intelligence Structure Test] (HTS Report no. 5105-7035). Retrieved from Hogrefe Uitgevers B.V. website: https://

www.hogrefe.nl/shop/media/downloads/sample-reports/5700601_ISTbasisrapport_mr.pdf

Hogrefe Ltd. (2016). *IST: Intelligence Structure Test: English version of the Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)* (HTS Report no.13216-189). Retrieved from Hogrefe Ltd website: https://www.hogrefe.co.uk/shop/media/downloads/sample-reports/5535902_mr.pdf

Holden, R. R., & Kroner, D. G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment*, *4*(2), 170–173. http://doi.org/10.1037/1040-3590.4.2.170

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distributione [On the empirical determination of a class of distribution]. *Giornale dell' Istituto Italiano degli Attuari*, *4*, 83–91.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399–402.

Loeys, T., Legrand, C., Schettino, A., & Pourtois, G. (2014). Semi-parametric proportional hazards models with crossed random effects for psychometric response times. *British Journal of Mathematical and Statistical Psychology*, *67*, 304–327. https://doi.org/10.1111/bmsp.12020

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*, 426–451. https://doi.org/10.3102/1076998614559412

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*, 121–137. https://doi.org/10.1177/0146621602250534

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 205–228.

Molenaar, D., Oberski, D., Vermunt, J. K., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*(5), 606–626.

Molenaar, D., Tuerlinckx, F., & Van der Maas, H. L. J. (2015a). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197–219. https://doi.org/10.1111/bmsp.12042

Molenaar, D., Tuerlinckx, F., & Van der Maas, H. L. J. (2015b). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74. https://doi.org/10.1080/00273171.2014.962684

Monahan, J. F. (2011). *Numerical methods of statistics* (2nd ed.). Cambridge University Press.

Ranger, J., & Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement*, *71*(2), 389–406.

Ranger, J., & Ortner, T. M. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*, 128–148.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151–171). North Holland.

Royston, J. P. (1982a). Algorithm AS177. Expected normal order statistics (exact and approximate). *Applied Statistics*, *31*, 161–165.

Royston, J. P. (1982b). An extension of Shapiro and Wilk's *W* test for normality to large samples. *Applied Statistics*, *31*, 115–124.

Royston, J. P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, *2*, 117–119.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611.

Shapiro, S. S., Wilk, M. B., & Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, *63*(324), 1343–1372.

Shea, B. L., & Scallan, A. J. (1988). AS R72. A remark on Algorithm AS128: Approximating the covariance matrix of normal order statistics. *Applied Statistics*, *37*, 151–155.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, *19*, 279–281. https://doi.org/10.1214/aoms/1177730256

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, *69*(347), 730–737.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384. https://doi.org/10.1007/s11336-0079046-8

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210. http://doi.org/10.1177/01466219922031329

van der Maas, H. L. J., & Jansen, B. R. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*(2), 141–177.

Veldkamp, B. P. (2016). On the issue of item selection in computerized adaptive testing with response times. *Journal of Educational Measurement*, *53*, 212–228.

Vermunt, J. K. (2011). K-means may perform as well as mixture model clustering but may also be much worse: Comment on Steinley and Brusco (2011). *Psychological Methods*, *16*(1), 82–88.

Verrill, S., & Johnson, A. (1988). Tables and large-sample distribution theory for censored-data correlation statistics for testing normality. *Journal of the American Statistical Association*, *83*, 1192–1197.

Wang, C., Chang, H., & Douglas, J. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*, 148–168.

Wang, C., Fan, Z., Chang, H., & Douglas, J. (2013). A semi-parametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*, 381–417.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477.

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*, 223–254. https://doi.org/10.1007/s11336-016-9525-x

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*, 469–501. https://doi.org/10.3102/1076998618767123

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*(5), 323–339.

Weisberg, S., & Bingham, C. (1975). An approximate analysis of variance test for non-normality suitable for machine calculation. *Technometries*, *17*, 133–134.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), 2141–2155. https://doi.org/10.1080/00949655.2010.520163

Authors

RENSKE E. KUIJPERS is a scientific researcher at the Department of Psychometrics and Research at Cito, PO Box 1034, 6801 MG Arnhem, the Netherlands; email: renske.kuijpers@cito.nl. Her research interests are response time modeling, item response theory, person fit, categorical marginal models, and Mokken scale analysis.

INGMAR VISSER is an associate professor at the Programme Group Developmental Psychology of the University of Amsterdam, PO Box 15906, 1001 NK Amsterdam, the Netherlands; email: i.visser@uva.nl. His research interests are cognitive development, implicit and explicit learning, categorization learning, hidden Markov models and latent variable models for learning, and developmental processes.

DYLAN MOLENAAR is an assistant professor at the Programme Group Psychological Methods of the University of Amsterdam, PO Box 15906, 1001 NK Amsterdam, the Netherlands; email: d.molenaar@uva.nl. His research interests are item response theory, factor analysis, response time modeling, intelligence, and statistical modeling of genotype by environment interactions.

Manuscript received January 23, 2019
First revision received February 14, 2020
Second revision received July 9, 2020
Third revision received July 29, 2020
Accepted August 7, 2020