PSYCHOMETRIKA—VOL. 84, NO. 3, 649–672
SEPTEMBER 2019
https://doi.org/10.1007/s11336-019-09671-8

MODELING DEPENDENCE STRUCTURES FOR RESPONSE TIMES IN A BAYESIAN FRAMEWORK

KONRAD KLOTZKE AND JEAN-PAUL FOX

UNIVERSITY OF TWENTE

A multivariate generalization of the log-normal model for response times is proposed within an innovative Bayesian modeling framework. A novel Bayesian Covariance Structure Model (BCSM) is proposed, where the inclusion of random-effect variables is avoided, while their implied dependencies are modeled directly through an additive covariance structure. This makes it possible to jointly model complex dependencies due to for instance the test format (e.g., testlets, complex constructs), time limits, or features of digitally based assessments. A class of conjugate priors is proposed for the random-effect variance parameters in the BCSM framework. They give support to testing the presence of random effects, reduce boundary effects by allowing non-positive (co)variance parameters, and support accurate estimation even for very small true variance parameters. The conjugate priors under the BCSM lead to efficient posterior computation. Bayes factors and the Bayesian Information Criterion are discussed for the purpose of model selection in the new framework. In two simulation studies, a satisfying performance of the MCMC algorithm and of the Bayes factor is shown. In comparison with parameter expansion through a half-Cauchy prior, estimates of variance parameters close to zero show no bias and undercoverage of credible intervals is avoided. An empirical example showcases the utility of the BCSM for response times to test the influence of item presentation formats on the test performance of students in a Latin square experimental design.

Key words: response time modeling, conditional independence, local dependence, testlets, Bayesian marginal modeling, non-informative prior distribution.

1. Introduction

In various research settings, it is of interest to make inferences about the effect of a treatment or experimental condition on a certain population. For example, two randomly sampled groups from the same population of students may be presented the same literacy test in different forms: The first group takes a traditional pencil-and-paper test, and the second group takes the computer-based counterpart. In that context, a researcher may want to gain insight into the differential functioning of items, or the test as a whole, across the two test forms. In other words, the focus lies on making inferences about the difference in performance between the two testing conditions, not on assessing the individuals' proficiency in reading and writing. A marginal model is appropriate when inferences about population-averages (e.g., comparing means or (co)variances across groups) are the goal of research (Diggle, Heagerty, Liang, & Zeger 2013). Unlike in their conditional counterpart, in a marginal framework the person effects are not modeled; they are integrated out. The interdependency between a person's observations is then not implied by a random-effect structure but is explicitly modeled in a covariance matrix. As discussed below, if inferences about population-averages are the focus of research, a marginal approach greatly favors the parsimony of the model at hand and can offer several advantages in the context of parameter estimation and model selection.

Correspondence should be made to Konrad Klotzke, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: k.klotzke@utwente.nl

© 2019 The Author(s)

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

A novel Bayesian Covariance Structure Model (BCSM) is proposed for clustered response times that is partly built on properties of a marginal modeling approach, but also explicitly accounts for the clustered structure of the data by modeling a structured covariance matrix. In the BCSM, the implied covariance structure of each random effect is separately modeled in the same additive covariance matrix, whereby each layer in the additive structure corresponds to one random effect. Therefore, the BCSM is a marginal modeling approach in which the dependence structure is explicitly modeled and hence preserved.

The BCSM differs from existing marginal modeling approaches, since the complete joint distribution of the observations is specified (and hence the complete likelihood of the model parameters). Thus, the BCSM preserves likelihood-based methods, which makes it possible to accommodate missing at random by default, the likelihoods usually give support to a unique maximum and can be used as the building blocks for a Bayesian modeling approach. This is not possible when using generalized estimating equations (GEE) to estimate a marginal model (Diggle et al., 2013; Liang & Zeger, 1986). In GEE, the covariance structure is treated as nuisance parameters and the focus lies solely on modeling the mean response. This avoids having to specify the conditional structure and therefore a possible misspecification of the same. A major downside of the GEE approach is that marginalization of different conditional structures can lead to inferentially identical models (Lee & Neider, 2004). This is the direct consequence of treating the covariance structures as nuisance parameters which do not have to be explicitly modeled to obtain consistent estimates. In other words, with an arbitrary covariance structure certain model assumptions cannot be checked for. Finally, contrary to the proposed framework, GEEs can be seen purely as an estimation procedure and do not allow common likelihood-based methods to assess the goodness-of-fit of a model, to compare models, to accommodate for missing at random, and to make inferences about model parameters.

To differentiate the proposed approach from existing marginal modeling methods, models constructed under the proposed framework are referred to as Bayesian Covariance Structure Models (BCSMs). BCSM offers three key advantages over a corresponding (conditional) random-effects model:

1. Tests for random-effect variances in mixed-effects models (e.g. Goldhammer & Kroehne, 2014) are complicated, as they require testing at the edge of the parameter space (Wood, 2013). These so-called boundary effects can lead to an underestimation of the statistical power of the corresponding tests and thus can bias the inferences made about the random-effect variance parameters of interest (Baguley, 2012, pp. 737–740). In a Bayesian framework, this problem is commonly tackled by choosing a more sophisticated prior distribution (e.g. Gelman, 2006; Gustafson, Hossain, & MacNab, 2006). The proposed BCSM, however, treats these parameters as covariances, which do not underlie the restriction of a lower or upper limit, as long as the positive definiteness of the covariance matrix is ensured. In line with that, boundary effects are reduced with truncated shifted inverse-gamma priors that allow the parameter space to cover negative values while enforcing sufficient rules for the positive definiteness of the covariance matrix. These priors are not as sharply peaked near zero as the default inverse-gamma priors and thus carry less information. Furthermore, in contrast to, for example, the half-Cauchy prior proposed by Gelman (2006), conjugacy is preserved. As a result, the hypothesis space is expanded to cover all likely parameter values and the availability of expressions of known forms for the conditional posterior distributions allows efficient Gibbs-sampling. In addition, given the proposed vague prior specification, more accurate estimates of very small random-effect variance parameters, respectively the corresponding covariances, can be obtained.

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

2. Specifying the effective number of parameters is trivial in the proposed framework, whereas in the random-effects model this forms an obstacle when applying model selection techniques such as the Bayesian Information Criterion (BIC) (Schwarz, 1978).
3. Estimation of random-effect variances is more likely to suffer from convergence issues with small sample sizes when compared to corresponding marginal models (Bell, John, & Jeffrey, 2008; Muth et al., 2016). This means that if the individual random effects themselves are not of interest and instead variance and covariance parameters are to be investigated, the proposed framework is of utility even when only limited data are available.

The BCSM for response times represents a multivariate generalization of the log-normal latent variable model (Klein Entink, Kuhn, Hornke, & Fox, 2009; van der Linden, 2006). A logarithmic transformation is applied to the naturally positively skewed distribution of response times, whereby the transformed response times of a person can be modeled with a normal distribution. In the conditional random-effect response time model, the observed response times are treated as realizations of a random variable and the corresponding probability distribution is determined by the items' time intensity and the person's speed. In the proposed BCSM for response times, the random effects themselves are not modeled. Instead, the implied interdependence between a person's response times is modeled in an additive covariance structure. On the lowest level of the additive covariance structure, the interdependence between a person's response times as implied by the person speed variable is modeled in a heterogeneous compound symmetric structure, where the measurement error variance parameters are free to vary across items. Therefore, in BCSM the random-effect variances are parameterized as covariance parameters. Latent variables such as time pressure, motivation, or the impact of testlet structures are not modeled but can cause local dependence within blocks of items. To take the additional sources of variation in a person's response times into account, as illustrated by Fig. 1, the contribution of each latent variable on the interdependence of response times is explicitly modeled in its own layer in the additive covariance structure. This allows the estimation of the individual (co)variance parameters and makes it possible to evaluate hypotheses about the parameters. Therefore, a statement can be made about whether or not a certain latent variable or factor has an impact on the interdependence between a person's response times (i.e., in the form of a test for local dependence within a block of items). As argued above, tests about

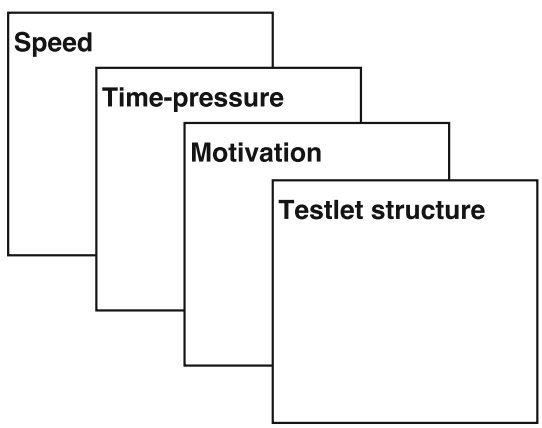


FIGURE 1.
In an additive covariance structure, each explicitly modeled layer represents the influence of a random-effect variable on the interdependence between a person's response times.

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

the random-effect variances offer a more exhaustive hypothesis space and are satisfied with a smaller sample size when compared to a corresponding random-effects model. Finally, the random effects themselves are not modeled, but their values can be recovered from the model's residuals.

The setup of the remaining text is as follows: a multivariate generalization of the log-normal response time model is specified within the BCSM framework. Extensions to include multidimensionality and factor loadings are discussed. Conjugate truncated shifted inverse-gamma priors are proposed that take into account the additive structure and positive definiteness of the covariance matrix, and resulting posteriors are derived. A Gibbs-sampling algorithm is defined with which samples from the full joint posterior can be obtained. A Bayes factor based on importance sampling and the BIC are discussed for the purpose of model selection in BCSM. Simulation studies are utilized to evaluate the proposed response time model's performance in parameter recovery and model selection. The proposed response time model is applied to an empirical example in an educational measurement setting. Finally, the results, limitations, and future prospects of the BCSM framework are discussed.

2. BCSM for Response Times

Before we define the response time model within the BCSM framework, we explain the notation as follows. The subscript i refers to the i-th person, g to the g-th group, and k to the k-th item. The number of persons in group g is denoted as n_g, and N stands for the total number of persons across all groups. Furthermore, the total number of groups and items is denoted as G and p, respectively. A bar over a data structure indicates the arithmetic mean over one or more dimensions that are specified by a dot in the subscript. For example, $\bar{T}_{.gk}$ denotes the mean log-response time over all persons in group g to item k. Finally, $\boldsymbol{I}_p$ and $\boldsymbol{J}_p$ are the identity matrix and a matrix of ones, each of dimension $p \times p$. The $p \times n_g$ data matrix $\boldsymbol{T}_g$ contains the logarithmic transformation of the measured time that it took persons in group g to give a response to the respective items.

In the log-normal model for response times, the response times of a person are explained by a person parameter and an item parameter. The item parameter λ_{gk} is the population-average log-response time for item k in group g. The person parameter ζ_{ig} represents the constant speed of person i in group g across all items and is assumed to follow a normal population distribution: $\zeta_{ig} \sim N(\mu_{\zeta_g}, \delta_g)$. It thus expresses the deviation of the person's speed from the population-average. This leads to the following equation for the log-response time of person i in group g to item k:

$$T_{igk} = \lambda_{gk} - \zeta_{ig} + \varepsilon_{igk},\ \varepsilon_{igk} \sim N(0, \sigma^2_{gk}). \tag{1}$$

The person speed parameter ζ_{ig} in Eq. (1) can be replaced with the sum of the average population speed of group g (μ_{ζ_g}), and the error of the group's population speed distribution $\varepsilon_{\zeta_{ig}}$:

$$\begin{aligned} T_{igk} &= \lambda_{gk} - (\mu_{\zeta_g} + \varepsilon_{\zeta_{ig}}) + \varepsilon_{igk} \\ &= \lambda_{gk} - \mu_{\zeta_g} + (\varepsilon_{\zeta_{ig}} + \varepsilon_{igk}) \\ &= \lambda_{gk} - \mu_{\zeta_g} + \tilde{\varepsilon}_{igk}. \end{aligned} \tag{2}$$

The error ε_{igk} in the distribution of response times and the error of the population distribution of speed $\varepsilon_{\zeta_{ig}}$ are conditionally independent. From that, it follows that the sum of the error terms $\tilde{\varepsilon}_{igk}$

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

is normally distributed with a mean of zero and a variance of $\delta_g + \sigma^2_{gk}$. As illustrated by Eq. (3), due to the marginalization, the response times of a person to different items are correlated with the covariance parameter δ_g. Given the above-mentioned marginalization, the covariance between the response times for two persons i and j of the same group g to items k and l is the following:

$$\begin{aligned}
\text{Cov}(T_{igk}, T_{jgl}) &= \text{Cov}(\lambda_{gk} - \mu_{\zeta_g} + \varepsilon_{\zeta_{ig}} + \varepsilon_{igk}, \lambda_{gl} - \mu_{\zeta_g} + \varepsilon_{\zeta_{jg}} + \varepsilon_{jgl}) \\
&= \text{Cov}(\varepsilon_{\zeta_{ig}} + \varepsilon_{igk}, \varepsilon_{\zeta_{jg}} + \varepsilon_{jgl}) \\
&= \text{Cov}(\varepsilon_{\zeta_{ig}}, \varepsilon_{\zeta_{jg}}) + \text{Cov}(\varepsilon_{igk}, \varepsilon_{jgl}) \\
&= \begin{cases} \delta_g + \sigma^2_{gk} & \text{if } i = j, k = l \\ \delta_g & \text{if } i = j, k \neq l \\ 0 & \text{if } i \neq j \end{cases}.
\end{aligned} \tag{3}$$

Consequently, the response times of each person are multivariate log-normally distributed with a p-dimensional mean vector

$$\boldsymbol{\mu}_{T_g} = [\lambda_{g1} - \mu_{\zeta_g}, \ldots, \lambda_{gp} - \mu_{\zeta_g}] \tag{4}$$

and the compound symmetry covariance matrix

$$\begin{aligned}
\boldsymbol{\Sigma}_{T_g} &= \text{diag}(\boldsymbol{\sigma}^2_g) + \delta_g \boldsymbol{J}_p \\
&= \begin{bmatrix} \delta_g + \sigma^2_{g1} & \delta_g & \ldots & \delta_g \\ \delta_g & \delta_g + \sigma^2_{g2} & \ldots & \delta_g \\ \vdots & \vdots & \ddots & \vdots \\ \delta_g & \delta_g & \ldots & \delta_g + \sigma^2_{gp} \end{bmatrix}.
\end{aligned} \tag{5}$$

Note that due to the marginalization, the mean and covariance structure is the same for all members of a group.

In the BCSM framework, the model specified in Eq. (5) describes the base layer of the additive covariance structure. Additional layers are modeled without modifying the mean structure specified in Eq. (4). As a result, multidimensionality in the interdependency of the response times can be introduced without including additional latent variables. Note that in the proposed model, each additional layer is explicitly modeled. This stands in contrast to an arbitrary covariance structure of a marginal model that is ambiguous about the corresponding conditional model. In the example illustrated by Fig. 2, persons are assumed to experience time pressure during the last part of the test. In a random-effects model, the time pressure effect would be represented by the latent variable γ_{ig}. That means that the variance of the random effects, i.e., $\text{Var}(\gamma_{ig}) = \Delta_g$, implies the dependence structure of a person's response times. In the BCSM approach, only the dependence structure is modeled; γ_{ig} itself is not modeled but would explain the specific dependence among response times to the affected (testlet) items of person i in the mean component. Note that Δ_g is parametrized as a covariance parameter in the BCSM. Furthermore, let $\boldsymbol{u}_g$ be a p-dimensional design vector of 0's and 1's where a 1 indicates that the response times to an item are affected by γ_{ig}. Then, an additive covariance structure is obtained, which is a straightforward extension of Eq. (5):

$$\boldsymbol{\Sigma}^*_{T_g} = \boldsymbol{\Sigma}_{T_g} + \Delta_g \boldsymbol{u}_g \boldsymbol{u}_g^T. \tag{6}$$

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

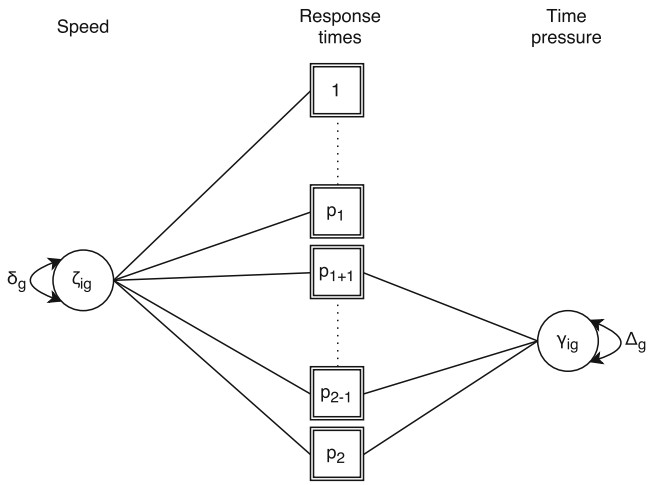


FIGURE 2.
Multidimensionality in the interdependence between response times is realized through the additional covariance parameter Δ_g. In a setting where the individual latent effects γ_{ig} are not of interest for hypothesis testing and model selection, modeling the implied local dependence in the response time data is sufficient.

Note that this extension is realized by modifying the covariance structure of the model with the addition of the covariance parameter Δ_g. In other words, instead of modeling the individual effect of a time pressure (γ_{ig}) on a person's response times, the implied covariance of a time pressure effect (Δ_g) on the errors is modeled. Furthermore, note that no additional identification rules are required, as long as the design vectors are mutually distinct (i.e., no two $\boldsymbol{u}_g$'s are the same). This holds for any pattern of an arbitrary number of additional layers.

BCSM is not limited to modeling the dependence structure implied by the single factor random intercepts model defined in Eq. (1). In fact, the described modeling approach generalizes to any covariance structure that can be expressed in the form of Eq. (6). This includes modeling the implied dependences of a random intercept and slope model [conditional: Eq. (7); BCSM: Eq. (8)], and of a testlet structure [conditional: Eq. (30); BCSM: Eqs. (31) and (32)]. Finally, dependences that are implied by correlated random effects are modeled directly in the additive covariance structure by specifying additional design vectors. Consequently, correlations between random effects are handled the same way as any other dependences in the data and do not require a modification of the described modeling approach.

As an illustration of modeling the dependence structure implied by correlated random effects, Table 1 contains the design vectors of a testlet RT BCSM for six items and three testlets. The first design vector specifies the dependences in the data that follow from the latent person speed variable. The next three rows specify the testlet structure, i.e., item 1 and 2, item 3 and 4, and item 5 and 6 each form a testlet. RTs to items in the same testlet are locally dependent. This dependence is explicitly modeled through the covariance parameter on the respective layer [i.e., Δ_g in Eq. (6)].

Following the same reasoning, the final three rows of Table 1 specify dependences between testlets. The corresponding covariance parameters can be interpreted as the covariances between testlet random effects in a random effects model. It is, however, important to note that BCSM is not limited to modeling dependences that are implied by random-effect structures. In particular, modeling negative interdependences (e.g., negative within-cluster correlations) poses a challenge in the random-effects modeling approach (e.g. El Leithy, Abdel Wahed, & Abdallah, 2016; Pryseley, Tchonlafi, Verbeke, & Molenberghs, 2011), but is straightforward and unambiguous in

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

TABLE 1.
The dependences implied by correlated random effects are directly modeled in the additive covariance structure without modeling the random effects themselves.

Covariance layer	Design vector					
Speed	1	1	1	1	1	1
Testlet 1	1	1	0	0	0	0
Testlet 2	0	0	1	1	0	0
Testlet 3	0	0	0	0	1	1
Cross testlets 1, 2	1	1	1	1	0	0
Cross testlets 1, 3	1	1	0	0	1	1
Cross testlets 2, 3	0	0	1	1	1	1

This is realized through the specification of cross-covariances between testlets through additional design vectors. Each row corresponds to the design vector of one covariance layer.

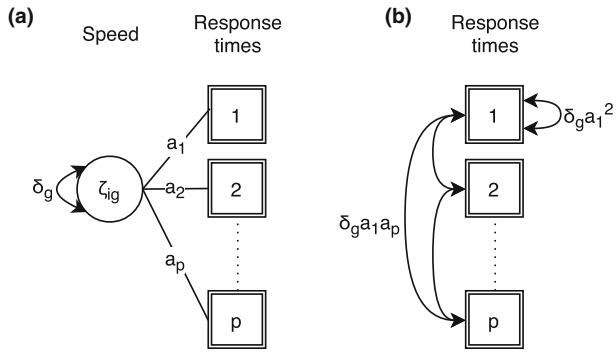


FIGURE 3.
a In a random-effects model, time-discrimination parameters can be interpreted as item-specific factor loadings for the latent person speed variable. **b** In BCSM, the dependence structure implied by time-discrimination parameters is directly modeled without the inclusion of random effects. Measurement error variances are not shown.

BCSM where dependences are modeled through covariance instead of variance parameters. Negative interdependences can furthermore naturally occur when jointly modeling different sorts of data, e.g., responses and response times (e.g. Klein Entink, Fox, & van der Linden, 2008; van der Linden, 2007).

Finally, factor loadings can be modeled in the proposed framework. An example is the time-discrimination parameter, which represents the quality of an item to discriminate between distributions of persons with a different level of speed (Klein Entink et al., 2008). The factor loading is included in the conditional response time model as an item-specific slope parameter a_{gk}:

$$T_{igk} = \lambda_{gk} - a_{gk}\zeta_{ig} + \varepsilon_{igk}. \tag{7}$$

Again, from this follows an additive covariance structure in the BCSM framework:

$$\boldsymbol{\Sigma}_{T_g} = \text{diag}(\boldsymbol{\sigma}_g^2) + \delta_g \boldsymbol{a}_g \boldsymbol{a}_g^T. \tag{8}$$

The corresponding random-effects model and its BCSM counterpart are shown in Fig. 3. Note that the resulting covariance matrix is not compound symmetric, but the properties necessary to build an additive structure are preserved. In fact, Eq. (8) removes the restriction of $u_{gk} \in \{0, 1\}$ in Eq. (6) and allows $a_{gk} \in \mathbb{R}$.

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

3. Priors for Additive Covariance Matrices

In the proposed BCSM framework, the random-effect variance parameters are represented by covariance parameters. While covariance parameters do not underlie the restriction of being greater or equal to zero, to keep the covariance matrix positive definite certain lower bounds may not be crossed. The lower bounds are obtained through applying the Sherman–Morrison formula to the given problem (Lange, 2010, pp. 260–261) and are enforced by truncating the prior at hand.

A sufficient condition for the positive definiteness is defined for any additive layer d of a $p \times p$-dimensional covariance matrix $\boldsymbol{A}$ of form

$$\boldsymbol{A}_{d+1} = \boldsymbol{A}_d + \psi \boldsymbol{v}\boldsymbol{v}^T, \tag{9}$$

where ψ is a scalar and $\boldsymbol{v}$ is a vector of length p. From the Sherman–Morrison formula, it follows that

$$1 + \psi \boldsymbol{v}^T \boldsymbol{A}_d^{-1} \boldsymbol{v} > 0 \tag{10}$$

is a sufficient condition for the positive definiteness of $\boldsymbol{A}_{d+1}$, under the presumption that $\boldsymbol{A}_d$ is also positive definite. The base layer $\boldsymbol{A}_1$ follows a heterogenous compound symmetry structure:

$$\boldsymbol{A}_1 = \text{diag}(\boldsymbol{\sigma}^2) + \delta \mathbf{1}_p \mathbf{1}_p^T. \tag{11}$$

From the condition defined in Eq. (10), it follows that $\min(\boldsymbol{\sigma}^2) > 0$ and $\delta > -1/\mathbf{1}_p^T \text{diag}(\boldsymbol{\sigma}^2)^{-1}\mathbf{1}_p$ together ensure that $\boldsymbol{A}_1$ is positive definite. If the base layer $\boldsymbol{A}_1$ is positive definite, then the following condition is thus sufficient to ensure the positive definiteness of any additional layer:

$$\psi > -1/\boldsymbol{v}^T \boldsymbol{A}_d^{-1} \boldsymbol{v}. \tag{12}$$

Note that a closed-form expression for $\boldsymbol{A}_d^{-1}$ can be derived from the Sherman–Morrison formula.

In line with the approach suggested by Fox, Mulder, and Sinharay (2017), shifted inverse-gamma priors are defined for the variance and covariance parameters. To ensure the positive definiteness of the covariance matrix, the condition defined in Eq. (12) is implemented through the indicator function $\mathbb{1}_t$. From this follows an extended inverse-gamma distribution with four parameters, where υ is the shift parameter and τ is the truncation point:

$$IG(x, \alpha, \beta, \upsilon, \tau) = \left[\frac{\beta^\alpha}{\Gamma(\alpha)} (x + \upsilon)^{-\alpha - 1} \exp\left(-\frac{\beta}{x + \upsilon} \right) \right] \cdot \mathbb{1}_t(x > \tau). \tag{13}$$

Note that $\tau = -\upsilon$ equals an untruncated shifted inverse-gamma distribution and $\tau = \upsilon = 0$ equals a default inverse-gamma distribution.

Consequently, the priors for the covariance and variance parameters can be written as

$$\pi(\delta_g | \boldsymbol{\sigma}_g^2) = IG(\delta_g, \alpha_0, \beta_0, \bar{\sigma}_g^2/p, -1/\mathbf{1}_p^T \text{diag}(\boldsymbol{\sigma}_g^2)^{-1}\mathbf{1}_p) \tag{14}$$

and

$$\pi(\boldsymbol{\sigma}_g^2 | \delta_g) = \prod_{k=1}^{p} IG(\sigma_{gk}^2, \alpha_0, \beta_0, \delta_g, 0). \tag{15}$$

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

For covariance parameters in additional layers, the truncation point changes according to Eq. (12). Note that the priors are defined in a conditional form, e.g., $\pi(\delta_g|\boldsymbol{\sigma}_g^2)$ and $\pi(\boldsymbol{\sigma}_g^2|\delta_g)$. This is sufficient for the Markov chain Monte Carlo (MCMC) algorithm. For Bayes factor testing, the joint prior, e.g., $\pi(\delta_g, \boldsymbol{\sigma}_g^2)$, can be constructed as the product of the (estimated) marginal priors.

4. Posterior Distributions

Given Eq. (5), the covariance between two responses times of a person i in group g for the k-th and/or l-th item is the following:

$$\text{Cov}(T_{igk}, T_{igl}) = \delta_g + \sigma_{gk}^2 \cdot \mathbb{1}\,(k = l)\,, \tag{16}$$

where $\mathbb{1}$ is the indicator function. Note that the total variance of a person's response time consists of a between-subject part (δ_g) and a within-subject part (σ_{gk}^2). The terms between-subject and within-subject follow from the assumption that all persons within a group share a common covariance structure.

The between sum of squares

$$\text{SSB} = \sum_{i=1}^{n_g} \left(\bar{T}_{ig.} - \bar{T}_{.g.}\right)^2, \tag{17}$$

is a sufficient statistic for the covariance parameter δ_g. In fact, multiplying the likelihood of the person means

$$p(\bar{T}_{1g.} \ldots \bar{T}_{ng.}|\bar{\sigma}_g^2, \delta_g) = (2p\pi)^{-\frac{n_g}{2}} (\bar{\sigma}_g^2/p + \delta_g)^{-\frac{n_g}{2}} \exp\left(-\frac{\text{SSB}/2}{\bar{\sigma}_g^2/p + \delta_g}\right), \tag{18}$$

with the conjugate truncated shifted inverse-gamma prior specified in Eq. (14) leads to the conditional posterior of δ_g:

$$\begin{aligned} &p(\delta_g|\bar{T}_{1g.} \ldots \bar{T}_{ng.}\bar{\sigma}_g^2) \\ &\quad = \left[\frac{(\beta_0 + \text{SSB}/2)^{(\alpha_0 + n_g/2)}}{\Gamma(\alpha_0 + n_g/2)} (\delta_g + \bar{\sigma}_g^2/p)^{-(\alpha_0 + n_g/2)-1} \exp\left(-\frac{\beta_0 + \text{SSB}/2}{\delta_g + \bar{\sigma}_g^2/p}\right)\right] \\ &\quad \cdot \mathbb{1}_t(\delta_g > -1/\mathbf{1}_p^T \text{diag}(\boldsymbol{\sigma}_g^2)^{-1}\mathbf{1}_p). \end{aligned} \tag{19}$$

Similarly, the within sum of squares of component k, $\text{SSW}_k = \sum_{i=1}^{n_g}(T_{igk} - \bar{T}_{.gk})^2$, is a sufficient statistic for the corresponding measurement error variance parameter. Given the prior specified in Eq. (15), the posterior is a truncated shifted inverse-gamma distribution with shift parameter δ_g and a truncation that ensures that $\sigma_{gk}^2 > 0$:

$$\sigma_{gk}^2 \sim IG(\alpha_0 + n_g/2, \beta_0 + \text{SSW}_k/2, \delta_g, 0). \tag{20}$$

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

To ease the derivation of Bayes factors about the invariance of measurement error variance parameters within or across groups, it is useful to sample the mean variance $\bar{\sigma}_g^2$ directly as an auxiliary parameter. As proved in "Appendix B", the posterior is also truncated shifted inverse-gamma:

$$\bar{\sigma}_g^2 \sim IG(\alpha_0 + n_g/2, \beta_0 + \text{SSW}/(2p), \delta_g, 0), \tag{21}$$

where $\text{SSW} = \sum_{i=1}^{n_g} \sum_{k=1}^{p} \left(T_{igk} - \bar{T}_{.gk}\right)^2$.

Like the covariance parameter δ_g in the base layer, the posterior of a covariance parameter Δ_{gd} in any additional layer d is shifted inverse-gamma distributed with a truncation to ensure the positive definiteness of the resulting covariance matrix. For example, if $d = 2$,

$$\Delta_{g2} \sim IG(\alpha_0 + n_g/2, \beta_0 + \text{SSB}_{\Delta_{g2}}/2, \bar{\sigma}_{g2}^2/p_2 + \delta_g, t_{\text{PSD}_{\Delta_{g2}}}), \tag{22}$$

where $\bar{\sigma}_{g2}^2/p_2$ is the average measurement error variance across the items that are affected by the additional covariance layer (i.e., items selected in the corresponding design vector $\boldsymbol{u}_{\Delta_{g2}}$ divided by the number of affected items p_2). Furthermore,

$$\text{SSB}_{\Delta_{g2}}/2 = \sum_{i=1}^{n_g} \left(\frac{1}{p_2} \sum_{k \in \boldsymbol{u}_{\Delta_{g2}}}^{p_2} (T_{igk}) - \frac{1}{n_g p_2} \sum_{i=1}^{n_g} \sum_{k \in \boldsymbol{u}_{\Delta_{g2}}}^{p_2} (T_{igk}) \right)^2$$

and $t_{\text{PSD}_{\Delta_{g2}}}$ is the truncation point following from Eq. (12). Note that Eq. (22) can be generalized to any number of additive layers by recursively computing the shift parameter and truncation point based on the layers below the current layer in the resulting covariance matrix.

5. Bayesian Inference

A Gibbs-sampling algorithm is specified with which samples from the full joint posterior distribution of the BCSM for response times can be drawn. As outlined in Algorithm 1, after the initialization phase, the item parameters, group parameters, measurement error variance parameters, and covariance parameters are sampled iteratively from their respective conditional posterior distribution. Finally, posterior mean estimates of the respective parameters are computed as the arithmetic mean of the MCMC samples while taking a burn-in phase into account.

To identify the model, the mean of the item parameters is assumed to be equal across groups; that is, $\bar{\lambda}_g = \bar{\lambda}_h$ for groups g and h. Furthermore, the group speed mean is fixed to zero in the first group ($\mu_{\zeta_1} = 0$). This rescaling is done via the (posterior) MCMC samples. Thereby, a distinction is made between the (untransformed) freely estimated parameters, for which a prior is specified, and the constrained (rescaled) parameters that are used for further computations (e.g. Fox, Klein Entink, & van der Linden, 2007; Luo & Jiao, 2018). For the fixed item and group effects, a locally uniform prior is defined. Finally, data missing at random $\boldsymbol{\omega}_g$ is properly imputed by drawing samples from the posterior predictive distribution of the data in each iteration. See "Appendix A" for details on the sampling steps.

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

```
/* Initialize */
Initialize chains with starting values
for 1:Chains do
    for 2:Iterations do
        for 1:Groups do
            /* Gibbs-sampling */
            Sample missing data ω_g|T_g, λ_g^{m-1}, μ_{ζ_g}^{m-1}, δ_g^{m-1}, σ_g^{2 m-1}
            Sample item parameters λ_g|T_g, ω_g, μ_{ζ_g}^{m-1}, δ_g^{m-1}, σ_g^{2 m-1}
            Sample group parameter μ_{ζ_g}|T_g, ω_g, λ_g, δ_g^{m-1}, σ_g^{2 m-1}
            Sample covariance parameter δ_g|T_g, ω_g, λ_g, σ_g^{2 m-1}
            Sample measurement error variance parameters σ_g^2|T_g, ω_g, λ_g, δ_g
        end
    end
end
/* Summarize */
for 1:Groups do
    Compute posterior mean estimates from MCMC samples
end
```

Algorithm 1: Sampling scheme of the BCSM for response times

6. Bayes Factor Testing

A Bayes factor quantifies the relative evidence of two competing models. More specifically, it is the ratio of evidence for each model times the a priori assumptions about the evidence, that is, the prior odds (Kass & Raftery, 1995):

$$\mathrm{BF}_{01} = \frac{m\left(\boldsymbol{T}; M_0\right)}{m\left(\boldsymbol{T}; M_1\right)} \cdot \frac{\pi_0}{\pi_1}. \tag{23}$$

Being a priori by nature, the prior odds $\frac{\pi_0}{\pi_1}$ incorporate information such as former research results or expert opinions and are not derived in the process of computing the Bayes factor. Thus, Eq. (23) simplifies to a ratio of marginal likelihoods. The marginal likelihood of the data under a model M_b is obtained by integrating the probability density function of the data with respect to the prior density:

$$m\left(\boldsymbol{T}|M_b\right) = \int \cdots \int p(\boldsymbol{T}|\phi_1, \ldots, \phi_z, M_b)\pi(\phi_1, \ldots, \phi_z|M_b)\, \mathrm{d}\phi_1 \ldots \mathrm{d}\phi_z, \tag{24}$$

where $\phi_1, \ldots, \phi_z$ are the model parameters of interest for the given Bayes factor. An estimator for the marginal likelihood is constructed based on the importance sampling technique proposed by Perrakis, Ntzoufras, and Tsionas (2014). In importance sampling, instead of integrating with respect to the prior density as in Eq. (24), the integration is applied with respect to an importance sampling density $g(\phi_1, \ldots, \phi_z|M_b)$. As illustrated by Perrakis et al. (2014), using the product of the marginal posterior distributions of the parameters of interest as the importance sampling density, that is, $g(\phi_1, \ldots, \phi_z|\boldsymbol{T}, M_b) = \prod_{u=1}^{z} p(\phi_u|\boldsymbol{T}, M_b)$, leads to an estimator with desirable properties: first, it is unbiased; second, it has a finite variance; and third, it handles any unknown

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

constants in the prior distributions as long as the corresponding marginal posteriors are included in the importance sampling density. The resulting integral

$$m\left(\boldsymbol{T}|M_b\right) = \int \cdots \int \frac{p\left(\boldsymbol{T}|\phi_1, \ldots, \phi_z, M_b\right) \pi\left(\phi_1, \ldots, \phi_z|M_b\right)}{\prod_{u=1}^{z} p\left(\phi_u|\boldsymbol{T}, M_b\right)} \prod_{u=1}^{z} p\left(\phi_u|\boldsymbol{T}, M_b\right) \mathrm{d}\phi_u \quad (25)$$

is estimated by

$$\hat{m}\left(\boldsymbol{T}|M_b\right) = \frac{1}{J} \sum_{j=1}^{J} \frac{p\left(\boldsymbol{T}|\phi_1^{(j)}, \ldots, \phi_z^{(j)}, M_b\right) \pi\left(\phi_1^{(j)}, \ldots, \phi_z^{(j)}|M_b\right)}{\prod_{u=1}^{z} p\left(\phi_u^{(j)}|\boldsymbol{T}, M_b\right)}, \quad (26)$$

where $\phi_1^{(j)}, \ldots, \phi_z^{(j)}$ are draws from the respective marginal posterior distributions and J is the number of MCMC samples utilized to estimate the marginal likelihood. Draws from the marginal posterior distributions are obtained by permuting the samples from the full joint posterior distribution (Perrakis et al., 2014, pp. 5–6): before randomly reordering each column (corresponding to the posterior sample of one model parameter) of the MCMC chain, the draws within each row (corresponding to one MCMC iteration) are naturally correlated draws from the conditional posterior distributions. After re-ordering, each row represents decorrelated draws from the marginal posterior distributions. The marginal posterior probabilities in the denominator and the marginal prior probabilities in the numerator of Eq. (26) are estimated through Rao-Blackwellization (Gelfand & Smith, 1990). In the case of data missing at random, the missing data parameters $\boldsymbol{\omega}$ do not provide additional information about the model evidence. Therefore, the marginal likelihood estimation is based solely on the observed data.

A straightforward example of the estimator specified in Eq. (26) is the evaluation of evidence in favor of the hypothesis that the covariance parameter is unrestricted ($H_1 : \delta \neq 0$) against evidence supporting the complementary hypothesis ($H_2 : \delta = 0$):

$$\mathrm{BF}_{12} = \frac{\frac{1}{J}\sum_{j=1}^{J} \frac{p\left(\boldsymbol{T}|\delta^{(j)}, \boldsymbol{\sigma}^{2(j)}, \boldsymbol{\xi}^{(j)}, M_1\right) \pi\left(\delta^{(j)}, \boldsymbol{\sigma}^{2(j)}, \boldsymbol{\xi}^{(j)}|M_1\right)}{p\left(\delta^{(j)}|\boldsymbol{T}, M_1\right) p\left(\boldsymbol{\sigma}^{2(j)}|\boldsymbol{T}, M_1\right) p\left(\boldsymbol{\xi}^{(j)}|\boldsymbol{T}, M_1\right)}}{\frac{1}{J}\sum_{j=1}^{J} \frac{p\left(\boldsymbol{T}|\delta^{(j)}, \boldsymbol{\sigma}^{2(j)}, \boldsymbol{\xi}^{(j)}, M_2\right) \pi\left(\boldsymbol{\sigma}^{2(j)}, \boldsymbol{\xi}^{(j)}|M_2\right)}{p\left(\boldsymbol{\sigma}^{2(j)}|\boldsymbol{T}, M_2\right) p\left(\boldsymbol{\xi}^{(j)}|\boldsymbol{T}, M_2\right)}}, \quad (27)$$

where $\pi(\delta|M_2)$ has a point mass at $\delta = 0$, and $\boldsymbol{\xi}$ is a block of nuisance parameters (e.g., item and group intercepts). If necessary, multiple blocks of nuisance parameters can be specified. Note that possible unknown normalizing constants of the priors also appear in the corresponding marginal posterior densities, thus canceling out. The estimator specified in Eq. (26) is furthermore suited to obtain evidence under order-restricted hypotheses such as $H3 : \delta_1 < \delta_2$, in which case the parameter space is constrained in some way (e.g. Gelfand, Smith, & Lee, 1992; Klugkist, Laudy, & Hoijtink 2005; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010).

In a setting where priors are deemed to be purely objective, an efficient approach to obtaining an estimate of the marginal likelihood of the data is the Laplace approximation (Bruijn, 1970, Chapter 4). Therefore, in this situation an appropriate method for comparing model evidence is the BIC:

$$\mathrm{BIC}_m = -2\log(\hat{L}_m) + d_m log(\hat{N}^*), \quad (28)$$

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

where m refers to the m-th model, $\hat{L}_m$ is the likelihood of the data given the posterior mean estimates of the model parameters, d_m is the number of free parameters under the model, and $\hat{N}^*$ is the sample size. Note that in BCSM, random effects are not estimated. Therefore, compared to linear mixed-effects models, shrinkage is greatly reduced in BCSMs and a better approximation of the effective number of parameters is achieved. Under a vague prior specification (i.e., small shape and large scale parameters), asymptotically, the effective number of parameters in the BCSM is equal to the number of fixed effects plus the parameters in the covariance matrix (e.g. Overholser & Xu, 2014). A safe choice for the effective sample size is the total number of observations $\hat{N}^* = pN$ (Faes, Molenberghs, Aerts, Verbeke, & Kenward, 2009). A Bayes factor for two competing models 0 and 1 can be approximated given the respective BICs:

$$\text{BF}_{01} = \frac{m\left(\boldsymbol{T}; M_0\right)}{m\left(\boldsymbol{T}; M_1\right)} \approx \exp\left(\frac{-\Delta\text{BIC}_{01}}{2}\right), \tag{29}$$

where $\Delta\text{BIC}_{01} = \text{BIC}_0 - \text{BIC}_1$.

7. Simulation Studies

Two simulation studies are conducted. The first simulation study aims at evaluating the estimation of testlet (co)variance parameters close to zero and the coverage rate of the relevant credible intervals. In that context, a BCSM testlet model for response times is compared to a random-effect testlet model. In the second simulation study, a Bayes factor for the local independence of response times within testlets is evaluated under different sample sizes and population values of the testlet (co)variance parameters. Both simulation studies are based on a test that consists of testlets: blocks of items that relate to a common content area (Wainer & Kiely 1987).

A testlet structure implies that a person's response times can be more alike within a testlet than across testlets. In other words, the grouping of similar items introduces dependence between a person's response times within a testlet. This dependence is not accounted for by merely controlling the persons' constant working speed across the test. Consequently, random-effects models commonly introduce a person–testlet interaction effect into the model equation (e.g. Hecht, Siegle, & Weirich, 2017; Wang & Wilson, 2005):

$$T_{igk} = \lambda_{gk} - (\zeta_{ig} + \theta_{igj(k)}) + \varepsilon_{igk},\ \varepsilon_{igk} \sim N(0, \sigma^2_{gk}), \tag{30}$$

where $j(k)$ denotes an item k in testlet j and $\theta_{igj(k)}$ is the corresponding person–testlet interaction effect. The random speed and person–testlet interaction effects are normally distributed, with $\zeta_{ig} \sim N(\mu_{\zeta_g}, \delta_g)$ and $\theta_{igj(k)} \sim N(\mu_{\theta_{gj}}, \Delta_{gj})$. To identify the model, the variance of the random speed effects is fixed. In the BCSM approach, the person–testlet interaction effects are not modeled in the mean term; that is,

$$\boldsymbol{T}_{ig} = \boldsymbol{\lambda}_g - (\mu_{\zeta_g} + \boldsymbol{\mu}_{\theta_{g(k)}}) + \boldsymbol{\varepsilon}_{ig},\ \boldsymbol{\varepsilon}_{ig} \sim N(\mathbf{0}_p, \boldsymbol{\Sigma}_g). \tag{31}$$

Instead, the additive covariance structure is extended with an additional layer and covariance parameter for each testlet. For a test consisting of N_t testlets, this results in an additive covariance structure with $N_t + 1$ layers: The first layer follows from the influence of the latent speed on the response times, and the remaining layers represent the contribution of each testlet effect on the dependence of a person's response times. Therefore, like the person speed parameters, the person–testlet interaction effects themselves are not modeled. Instead, the dependence between a

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

person's response times within a testlet is explicitly modeled in the covariance structure of the error term. The part of the dependence between response times that is assumed to be explained by the latent person speed is operationalized as the covariance parameter δ_g. The part of the dependence between response times that is assumed to be explained by the testlet structure, while keeping the latent speed constant, is operationalized as Δ_{gj}. The additive layer structure is represented by the following covariance matrix:

$$\boldsymbol{\Sigma}_g = \text{diag}(\boldsymbol{\sigma}_g^2) + \delta_g \boldsymbol{J}_p + \sum_{j=1}^{N_t} \Delta_{gj} \boldsymbol{u}_{gj} \boldsymbol{u}_{gj}^T, \tag{32}$$

where $\boldsymbol{u}_{gj}$ is a p-dimensional design vector specifying which items belong to testlet j in group g.

7.1. Parameter Estimation and Credible Intervals

To measure the precision and bias of testlet (co)variance parameter estimates and the coverage rate of the corresponding credible intervals a simulation experiment is conducted. The number of test-takers ($N = 300$), the length of the test ($p = 30$), and the number of testlets ($N_t = 3$) are fixed across the 1000 replications. All test-takers are part of the same group. The population values of the first three testlet (co)variance parameters are $\Delta_{g1} = 0$, $\Delta_{g2} = .01$ and $\Delta_{g3} = .05$. The remaining population parameters are $\delta_g = .2$, $\mu_{\zeta_g} = \mu_{jg} = 0$, $\lambda_{gk} \sim N(0, 1)$ and $\sigma_{gk}^2 = \mathbf{1}_p$. Data are generated under the restrictions of the respective models: (a) the a priori assumption about whether or not the testlet (co)variance parameters may be negative is taken into account when simulating response times; and (b) to identify the random-effects model, the variance of the random speed effects is fixed, that is, $\zeta_{ig} \sim N(0, .2)$. Consequently, the variance is also fixed when generating data for the random-effects models. Note that all parameters in the BCSM covariance structure are free, i.e., δ_g is not fixed for the BCSM.

Both the BCSM and the random-effects model are fitted in a Bayesian framework. The Gibbs-sampling algorithm of the BCSM framework is implemented in R (R Core Team, 2017), and the random-effects model is fitted with the R package R2jags (Su, 2015). Estimates refer to the mean of the respective posterior distributions and the credible intervals are equally-tailed. The coverage rate of the credible intervals and the distribution of the posterior mean estimates is based on 1000 replications. Each replication consists of 10,000 MCMC iterations, and a burn-in phase of 10% is applied. For the BCSM, a truncated shifted inverse-gamma prior with shape $= 10^{-8}$ and scale $= 10^8$ is defined for the testlet (co)variance parameters. For the random-effects model, a default non-informative inverse-gamma prior on the variance parameters is likely to cause the MCMC sample chains of very small variance parameters to often get stuck at zero (Browne, Steele, Golalizadeh, & Green, 2009; Lesaffre & Lawson, 2012). In practice, the resulting autocorrelation renders obtaining information about the posteriors given a reasonable number of MCMC iterations futile. As a remedy, parameter expansion is implemented through a half-Cauchy prior with mode $= 0$ and scale $= 25$ on the testlet standard deviation parameters as proposed by Gelman (2006).

A visual inspection of the model parameters' trace plots showed no evidence against convergence of the MCMC algorithms. The results of the parameter estimation and the coverage rates are shown in Table 2. Due to the skewness of the respective distributions, the posterior mean estimates of the two smallest testlet variance parameters are positively biased for the random-effects model. Under the BCSM, no bias is observed. The standard deviation of the posterior mean estimates is smaller for the random-effects model. However, as described above, fewer restrictions were applied when generating data for the BCSM. Under the BCSM, the empirical coverage rates correspond to the theoretical coverage of the credible intervals. Under the random-effects model,

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

TABLE 2.
Upper part: mean and standard deviation of posterior mean estimates of testlet (co)variance parameters. Lower part: empirical coverage of corresponding 95%-credible intervals.

	Trunc. shifted IG	Half-Cauchy
Δ	*Empirical mean (SD) of posterior mean estimates*	
0	.000 (.034)	.021 (.015)
.01	.011 (.033)	.026 (.019)
.05	.051 (.037)	.054 (.030)
Δ	*Empirical coverage of 95%-credible intervals*	
0	94.3	0
.01	95.1	84.5
.05	95.5	75.4

Results based on 1000 simulated replications with $N = 300$ persons, $p = 30$ items and $N_t = 3$ testlets.

a true value of zero is not included in any of the computed 95%-credible intervals. For true values close to zero, a significant undercoverage is observed.

7.2. Model Selection

Given a testlet structure, the assumption of local independence states that a person's response times within and between testlets are independent when controlling for the person's speed. In the random-effects model specified in Eq. (30) and the proposed BCSM specified in Eq. (32), the assumption of local independence is violated if the testlet (co)variance is not equal to zero. In this simulation, the plausibility of two versions of the model described in Eq. (32) is evaluated with a Bayes factor. In the null model M_0, the covariance parameter Δ_1 is restricted to zero. In the alternative model M_a, all covariance parameters are unrestricted. In other words, according to the null model local independence holds for the items within the first testlet, and the alternative hypothesis indicates local dependence. The model evidence is compared between M_0 and M_a for a set of 7 population values of Δ_1, namely $\{-.2, 0, .2, .4, .6, .8, 1\}$. For each value of Δ_1, 50 samples are drawn from the respective population and for each sample the log-Bayes factor is computed. This is done twice, first for three groups of test-takers of size $N_1 = N_2 = N_3 = 100$ and second for a group size of $N_1 = N_2 = N_3 = 150$. In both cases, data for $p = 18$ items are simulated and 3000 MCMC iterations are run. A truncated shifted inverse-gamma prior with shape $= 10^{-3}$ and scale $= 10^3$ is defined for the testlet covariance parameters. The means of each replication are summarized in Fig. 4.

The Bayes factor behaves as expected: a larger discrepancy between the population value of Δ_1 and zero makes the alternative model more plausible. Figure 5 shows the empirical density of the log-Bayes factor at $\Delta_1 = .4$ for group sizes of $N = 100$ and $N = 150$. Both figures illustrate that more data lead to a greater statistical power.

8. Empirical Example

Realistic Mathematics Education (RME) is an approach to teaching and learning theory that is based on the idea of providing students with problems that are perceived as useful and relevant. This aims at making mathematical education accessible to a wider range of students and therefore giving numeracy a stronger focus in society. Based on an empirical dataset by Buschers (2016), we investigated whether or not presenting contextual numeracy items in different formats has an

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

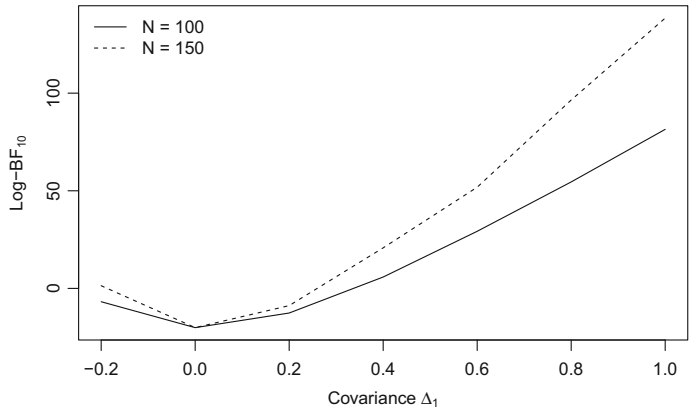


FIGURE 4.
Average log-Bayes factor across 50 replications quantifying the evidence for $H_a : \Delta_1 \neq 0$ against the evidence for $H_0 : \Delta_1 = 0$. A positive value indicates that H_a is more plausible. The comparison is made for three groups of size $N_1 = N_2 = N_3 = 100$, respectively $N_1 = N_2 = N_3 = 150$.

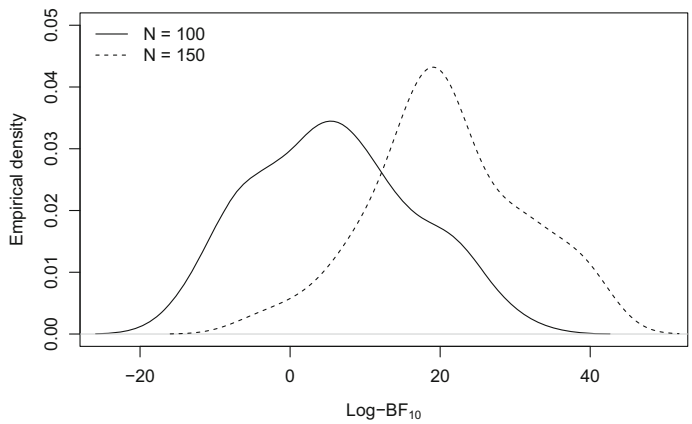


FIGURE 5.
Empirical density of the log-Bayes factor across 50 replications quantifying the evidence for $H_a : \Delta_1 \neq 0$ against the evidence for $H_0 : \Delta_1 = 0$. A positive value indicates that H_a is more plausible. Samples are drawn from a population with $\Delta_1 = .4$. The comparison is made for three groups of size $N_1 = N_2 = N_3 = 100$, respectively $N_1 = N_2 = N_3 = 150$.

effect on students' response processes. To gain insight into the latent response processes, response times are utilized (Molenaar, Tuerlinckx, & Maas, 2014). Three different presentation formats were considered: text only, image only and text and image (see Table 3 for an example). In general, a numeracy problem in a test should not overload or distract the student with redundant information and should furthermore ensure that the available information is easily accessible to the student. For the given empirical example, the focus is on the translation of the contextual problem to a mathematical problem. In this context, two cognitive theories are considered. First of all, the cognitive load theory states that the cognitive capacity of a student, and in particular his or her short term memory, is limited and can thus be overloaded. Hence, in the majority of cases it seems better to not repeat the information included in an image in the accompanying text. Second, according to the dual channel principle students have separate channels to process verbal and pictorial information. Therefore, it is not only the amount of information that plays a role, but also how the information is presented and thus processed. For example, it can be argued that

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

TABLE 3.

Note. Reprinted from Words, pictures or both?: the influence of the presentation of contextual numeracy problems on student performance in (pre) vocational education, by Buschers (2016), *unpublished Master's thesis*, p. 7.

Text variant	**Image variant**	**Text + Image Variant**
Two market stalls are offering strawberries. The first stall sells 500 grams of strawberries for EUR 1.95. The second stall sells 150 grams of strawberries for EUR 0.75. What is the difference in price between stall 1 and stall 2 for a kilo of strawberries?	Two market stalls are offering strawberries.  What is the difference in price between stall 1 and stall 2 for a kilo of strawberries?	Two market stalls are offering strawberries. The first stall sells 500 grams of strawberries for EUR 1.95. The second stall sells 150 grams of strawberries for EUR 0.75.  What is the difference in price between stall 1 and stall 2 for a kilo of strawberries?

presenting information in text and image allows the student to focus on the format of presentation, or combination thereof, that works best for him or her.

In total, data from 301 respondents were recorded in various Dutch schools. The randomly assigned groups are of size $N_1 = 99$, $N_2 = 96$ and $N_3 = 94$ when including respondents for whom data for at least five items were available. The respondents are students of the three levels of prevocational education and the third level of vocational education. A partially counterbalanced Latin square design with three blocks is employed. Each block corresponds to one of the three presentation formats. Thus, within the three randomly assigned groups, each student is presented all $p = 35$ items, but with different formats. The order of the presentation format varies per group in accordance with the Latin square design. Within each group, item order effects are possible. These effects are accounted for when the three counterbalanced groups are merged for the purpose of statistical inference (i.e., parameter estimation and hypothesis testing).

The goal of this research is to investigate whether or not the students' performance differs between the three item presentation formats. This is operationalized as differences in response times between the three variants. More specifically, it is of interest whether or not the response times within one variant are more alike than the response times across all variants. Furthermore, if it is plausible that the response times within two variants are more alike than the response times across all variants, then comparing the covariance within the two variants is of interest. The plausibility of the hypotheses is evaluated with the approximated Bayes factor specified in Eq. (29).

8.1. The Statistical Model

Each item presentation format variant corresponds to a separate testlet j in a testlet structure. The item–testlet combination (i.e., which item belongs to which testlet) varies across the three groups of students. An appropriate mixed-effects model is the following:

$$T_{igk} = \lambda_{gk} - (\zeta_{ig} + \theta_{igj(k)}) + \varepsilon_{igk},\ \varepsilon_{igk} \sim N(0, \sigma^2_{gk}), \tag{33}$$

where the random speed and testlet effects are normally distributed with $\zeta_{ig} \sim N(\mu_{\zeta_g}, \delta_g)$ and $\theta_{igj(k)} \sim N(\mu_{\theta_{gj}}, \Delta_{gj})$. The goal of the present research is to make inferences about the effect of

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

the testlet structure (i.e., the different item presentation formats) on the interdependence between a person's response times. More specifically, it is of interest to evaluate whether or not response times are more alike within a testlet than across testlets. This translates into a statement about the local dependence of a person's response times within each testlet. For the given application, the focus thus lies on the population-average of the (co)variance parameters Δ_{gj} and not on the individual random effects, i.e., $\theta_{igj(k)}$.

In the BCSM framework, the mixed-effects model in Eq. (33) is marginalized and the interdependency between a person's response times is modeled in an additive covariance matrix with four layers:

$$\boldsymbol{\Sigma}_{T_g} = \left[\text{diag}(\boldsymbol{\sigma}_g^2) + \delta_g \boldsymbol{J}_p\right] + \left[\Delta_{g1}\boldsymbol{u}_{g1}\boldsymbol{u}_{g1}^T\right] + \left[\Delta_{g2}\boldsymbol{u}_{g2}\boldsymbol{u}_{g2}^T\right] + \left[\Delta_{g3}\boldsymbol{u}_{g3}\boldsymbol{u}_{g3}^T\right], \tag{34}$$

where δ_g describes the covariance across all items and Δ_{g1}, Δ_{g2} and Δ_{g3} describe the additional covariance within the variants "Text", "Image" and "Text and Image". The fact that the testlet effect $\theta_{igj(k)}$ has three categories is represented in the BCSM by the design vectors $\boldsymbol{u}_{g1}$, $\boldsymbol{u}_{g2}$ and $\boldsymbol{u}_{g3}$, which thus specify the order of the presentation formats in group g. Two rules are defined to identify the model. First, the group speed mean is set to zero in all groups ($\mu_{\zeta_1} = \mu_{\zeta_2} = \mu_{\zeta_3} = 0$). As a result, the time intensity parameters are on the same scale across groups, which allows the extraction of the presentation variant effects. Second, the measurement error variance parameter of the last item (σ_p^2) is set to be equal in all groups. This ensures that the covariance parameters are on the same scale across groups (e.g., $\Delta_{11} = \Delta_{21} = \Delta_{31}$). A truncated shifted inverse-gamma prior with shape $= 10^{-3}$ and scale $= 10^3$ is defined for the variance and covariance parameters. For the fixed item effects (λ_{gk}) a locally uniform prior is approximated with $N(0, 10^{10})$. Finally, data are assumed to be missing at random (MAR).

8.2. Results

The model parameters are estimated with one MCMC chain of 50,000 iterations. A burn-in phase of 10% is applied. A visual inspection of the model parameters' trace plots showed no evidence against convergence of the MCMC algorithm. The posterior mean estimate of the covariance parameter across all variants is .417 (SD .135). The additional covariances in the "Text", "Image" and "Text and Image" variants are estimated as .100 (SD .130), .036 (SD .071) and .043 (SD .082), respectively. The plausibility of the hypothesis stating that response times are more alike within a variant than across variants is evaluated with an approximated Bayes factor by comparing the evidence for said hypothesis to the evidence in favor of the respective complementary null hypotheses: $H_{01} : \Delta_1 = 0$, $H_{02} : \Delta_2 = 0$ and $H_{03} : \Delta_3 = 0$.

The results of the hypothesis testing are summarized in Table 4. Following the guidelines of Kass and Raftery (1995) to interpret the results, very strong evidence is found against the three alternative hypotheses. This means that, given the data at hand, it is highly implausible

TABLE 4.
Approximated log-Bayes factor quantifying the plausibility of the alternative hypothesis, i.e., response times within a presentation format variant are more alike than response times across variants, against the null hypothesis; i.e., the response times are not more alike.

	$H_{a1} : \Delta_1 \neq 0$	$H_{a2} : \Delta_2 \neq 0$	$H_{a3} : \Delta_3 \neq 0$
Log-BF_{a0}	− 19	− 66	− 23

A positive value indicates evidence in favor of the alternative hypothesis.

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

that response times are more alike if they are collected under the same item presentation format. In other words, variation in presentation format does not cause local dependence within the corresponding blocks of items. This result is in line with the very small average effects of the presentation variants on the response times. The effects are extracted from the residuals of the model and indicate that the log-response times are, on average, the lowest in the "Text" variant ("Text" − "Image" = − .025; "Text" − "Text and Image" = − .070) and the highest in the "Text and Image" variant ("Text and Image" − "Image" = .045).

9. Discussion

In a novel Bayesian modeling framework, a multivariate generalization of the log-normal response time model has been proposed. The BCSM framework allows the specification of models based on, but not limited to, an integrated likelihood approach. Under the integrated likelihood approach, the random effects are integrated out, and their implied dependencies between observations are directly modeled in a covariance structure in which the random-effect variance parameters serve as covariance parameters. The complexity of the BCSMs is easily controlled, since each random-effect structure is modeled in a separate layer of an additive covariance structure. This is much more difficult in a conditional modeling approach, where each random effect introduces many model parameters and the exact number of parameters depends stronger on the fit of the model.

In the conditional random-effects models, inferences about variance parameters are also problematic. For instance, a random-effect variance of zero can be of specific interest, but the value zero is the lower bound of the corresponding parameter space. The prior specification of a positive variance component can lead to biased parameter estimates and can complicate testing the support for a random effect. In the BCSM, these so-called boundary effects can be avoided, or at least weakened, by extending the parameter space to include negative values. Therefore, shifted inverse-gamma priors are proposed for the variance components, which include a restriction on the parameter space to ensure that the covariance matrix is positive definite but allow negative parameter values, thereby accounting for boundary effects and creating a more exhaustive hypothesis space. Contrary to other priors for variance components such as the half-t or half-Cauchy priors, conjugacy is preserved with the proposed truncated shifted inverse-gamma priors. This greatly increases the efficiency of the MCMC sampling algorithm. The proposed priors furthermore lead to less skewness in the posterior distribution if a covariance is close to zero, when compared to priors for variance components. As a result, bias of posterior mean estimates and undercoverage of credible intervals is avoided in a situation where the true value of the (co)variance parameter is located near zero.

The sample size requirements for the BCSM for response times to obtain stable estimates are minimal: for each random-effect structure only two items are needed; this means that the latent speed effect or multidimensional effects (e.g., testlet structure) can be measured with two or more items. In general terms, it is sufficient to have observations from two items to measure an additional dependency, which is modeled as a separate layer in the additive covariance matrix. Furthermore, explicitly modeling each layer of the covariance structure allows testing model assumptions within and across layers. For example, Lee and Neider (2004) point out that in common marginal models it is impossible to test for treatment–random-effect interaction as the marginal models are inferentially identical regardless of whether or not the interaction is present in the corresponding conditional model. In the proposed BCSM framework, these interactions are explicitly modeled and can be tested for, as demonstrated with the testlet structure in the second simulation study and in the context of the empirical example.

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

The estimates of the BCSM for response times may not be directly comparable to those of a conditional model. This can be caused by different constraints on the parameters space; that is, covariance parameters may take on negative values in the BCSM, while variance parameters in the conditional model have a lower bound at zero. For the above-mentioned reasons, we argue that the (co)variance estimates of the proposed BCSM are more accurate representations of their true values. A related point of caution is the recovery of random effects in the BCSM. As demonstrated in the context of the empirical example, it is possible to recover the random-effect information from the model's residuals post hoc. Due to the different constraints on the parameter space, the random-effect estimates that are made by the BCSM for response times can be seen as originating from a qualitatively different model, when compared to estimates from a conditional model.

An interesting future prospect of the BCSM for response times is a to combine it with the marginal IRT model by Fox et al. (2017) into a joint-model where the interdependence between response accuracy and speed is explicitly modeled as item-specific cross-covariance parameters. This may lend insight into the effect of, for example different item forms, testlet structures, or time pressure conditions on the speed-accuracy trade-off within a group of persons. Existing approaches to joint-models either assume a constant correlation between response accuracy and speed across persons and items (Glas & van der Linden, 2010; Klein Entink et al., 2008; Loeys, Legrand, Schettino, & Pourtois, 2014; Ranger & Kuhn, 2013; Thissen, 1983; van der Linden & Fox, 2016); do not allow the explicit modeling of the item-specific cross-covariance parameters (Goldhammer & Kroehne, 2014; Goldhammer et al., 2014; Molenaar, Tuerlinckx, & Maas, 2015); or strictly limit the number of states in the speed-accuracy relationship that can be modeled given a reasonable sample size in educational measurement research (Molenaar, Oberski, Vermunt, & Boeck, 2016; Wang & Xu, 2015).

The proposed framework can furthermore be extended with link functions, which translate a latent multivariate normally distributed variable (e.g., response accuracy) into observations that follow a different distribution (e.g., dichotomous item responses). Finally, it is possible to sample directly from the posterior predictive distribution of the data. In the empirical example, this is utilized for proper imputation. However, it also makes the creation of posterior predictive checks (PPC) straightforward. PPCs have been shown to be useful in checking assumptions of IRT models such as multidimensionality or conditional independence (Levy, Mislevy, & Sinharay, 2009) and have been extended to joint-models that incorporate speed and accuracy (Bolsinova & Tijmstra, 2016).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix A

The sampling steps described below correspond to the sampling scheme as outlined in Algorithm 1. $\boldsymbol{T}^* = \{\boldsymbol{T}, \boldsymbol{\omega}\}$ is the imputed dataset.

9.1. Drawing Samples from the Full Joint Posterior

The missing data, item, group, covariance and measurement error variance parameters are iteratively sampled from their respective conditional posterior distribution.

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

9.1.1. Sample Missing Data Parameters The missing data parameters $\boldsymbol{\omega}$ are sampled from the distribution of the replicated data (i.e., the posterior predictive distribution of the data):

$$p(\boldsymbol{\omega}|\boldsymbol{T}, \boldsymbol{R}) = \int p(\boldsymbol{\omega}|\boldsymbol{T}, \boldsymbol{R}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{T}, \boldsymbol{\omega}, \boldsymbol{R}) \mathrm{d}\boldsymbol{\theta},$$

where $\boldsymbol{R}$ are the missing data indicators, $\boldsymbol{\theta}$ is the vector of model parameters, and $p(\boldsymbol{\theta}|\boldsymbol{T}, \boldsymbol{\omega}, \boldsymbol{R})$ is the posterior distribution of $\boldsymbol{\theta}$.

9.1.2. Sample Item Parameters The conditional posterior distribution of the item time intensity parameters is univariate normal with mean $E\left(\lambda_{gk}|\cdot\right)$ and variance $\mathrm{Var}\left(\lambda_{gk}|\cdot\right)$:

$$E\left(\lambda_{gk}|\boldsymbol{T}^*_{gk}, \mu_{\zeta_g}, \sigma^2_{gk}, \delta_g, \mu_{\lambda 0}, \sigma^2_{\lambda 0}\right) = \mathrm{Var}\left(\lambda_{gk}|\cdot\right)\left(\frac{n_g(\bar{T}^*_{\cdot gk} + \mu_{\zeta_g})}{\sigma^2_{gk} + \delta_g} + \frac{\mu_{\lambda 0}}{\sigma^2_{\lambda 0}}\right),$$

$$\mathrm{Var}\left(\lambda_{gk}|\sigma^2_{gk}, \delta_g, \sigma^2_{\lambda 0}\right) = \left(\frac{n_g}{\sigma^2_{gk} + \delta_g} + \frac{1}{\sigma^2_{\lambda 0}}\right)^{-1}.$$

9.1.3. Sample Group Parameters The conditional posterior distribution of each speed group mean parameter is univariate normal with mean $E\left(\mu_{\zeta_g}|\cdot\right)$ and variance $\mathrm{Var}\left(\mu_{\zeta_g}|\cdot\right)$:

$$E\left(\mu_{\zeta_g}|\boldsymbol{T}^*_g, \boldsymbol{\lambda}_g, \boldsymbol{\sigma}^2_g, \delta_g, \mu_{\zeta 0}, \sigma^2_{\zeta 0}\right) = \mathrm{Var}\left(\mu_{\zeta_g}|\cdot\right)\left(\frac{n_g(\bar{\lambda}_g - \bar{T}^*_{\cdot g \cdot})}{\left(\sum_{k=1}^{p}(\sigma^2_{gk}/p + \delta_g)\right)/p} + \frac{\mu_{\zeta 0}}{\sigma^2_{\zeta 0}}\right),$$

$$\mathrm{Var}\left(\mu_{\zeta_g}|\boldsymbol{T}^*_g, \boldsymbol{\sigma}^2_g, \delta_g, \sigma^2_{\zeta 0}\right) = \left(\frac{n_g}{\left(\sum_{k=1}^{p}(\sigma^2_{gk}/p + \delta_g)\right)/p} + \frac{1}{\sigma^2_{\zeta 0}}\right)^{-1}.$$

9.1.4. Sample Covariance Parameters The covariance parameter δ_g is sampled from its conditional inverse-gamma posterior distribution as defined in Eq. (19). Covariance parameters from additional layers are sampled directly from their conditional posterior that follows from Eq. (22).

9.1.5. Sample Measurement Error Variance Parameters The measurement error variance parameters of the response times $\boldsymbol{\sigma}^2_g$ are sampled from their respective conditional inverse-gamma posterior distribution as defined in Eq. (20). The shift parameter and the truncation point are adjusted if layers are added to the covariance matrix.

Appendix B

Let $\boldsymbol{U}$ have a mean vector equal to $\mathbf{0}_p$ (e.g., when $\boldsymbol{U}$ are realizations of the multivariate normally distributed error variances of the model, and thus the residuals). The trace of the diagonalizable covariance matrix, $\mathrm{tr}\left(\boldsymbol{\Sigma}_U\right)$, is equal to the sum of its eigenvalues (Axler, 2014, p. 302). In other words, if the basis of the corresponding vector space changes as a result of decorrelating the data (i.e., by multiplying the data matrix with the inverse of the eigenvectors), then the sum

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

of the within sum of squares across all items remains unchanged. To apply this theorem to the problem at hand, first the trace of the CS covariance structure defined in Eq. (16) is derived:

$$\text{tr}\,(\boldsymbol{\Sigma}_U) = (\sigma_{g1}^2 + \delta_g) + \cdots + (\sigma_{gp}^2 + \delta_g) = p(\bar{\sigma}_g^2 + \delta_g).$$

Subsequently, let $\boldsymbol{X}$ be the decorrelated data. Furthermore, let the vector $\boldsymbol{Z}_{n_g}$ contain the person means across all items of the transformed data: $\boldsymbol{Z}_{n_g} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_p$. From this follows

$$Z \sim N(0,\, p(\bar{\sigma}_g^2 + \delta_g)),$$

whereby the variance thus equals the trace of the covariance matrix as shown in Eq. (5). A truncated inverse-gamma prior that ensures that $\bar{\sigma}_g^2 > 0$ then leads to a posterior truncated shifted inverse-gamma distribution with location parameter δ_g:

$$\bar{\sigma}_g^2 \sim IG(\alpha_0 + n_g/2,\, \beta_0 + \text{SSW}/(2p),\, \delta_g,\, 0),$$

where $\text{SSW} = \sum_{i=1}^{n_g} \sum_{k=1}^{p} \left(U_{igk} - \bar{U}_{.gk}\right)^2$.

References

Axler, S. (2014). *Linear algebra done right*. Berlin: Springer.

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. London: Macmillan Education.

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models, pp. 1122–1129. In *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section*. Alexandria, VA: American Statistical Association.

Bolsinova, M., & Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics*, *41*(2), 123–145. https://doi.org/10.3102/1076998616631746.

Browne, W. J., Steele, F., Golalizadeh, M., & Green, M. J. (2009). The use of simple reparameterizations to improve the efficiency of Markov chain Monte Carlo estimation for multilevel models with applications to discrete time survival models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(3), 579–598. https://doi.org/10.1111/j.1467-985X.2009.00586.x.

Buschers, L. (2016). *Words, pictures or both?: The influence of the presentation of contextual numeracy problems on student performance in (pre) vocational education*. Unpublished Masters thesis, University of Twente, Enschede. Retrieved December 5, 2017 from http://essay.utwente.nl/70306/1/Buschers_MA_BMS.pdf.

de Bruijn, N. G. (1970). *Asymptotic methods in analysis*. Mineola: Courier Corporation.

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. (2013). *Analysis of longitudinal data*. Oxford: OUP.

El Leithy, H. A., Abdel Wahed, Z. A., & Abdallah, M. S. (2016). On non-negative estimation of variance components in mixed linear models. *Journal of Advanced Research*, *7*(1), 59–68. https://doi.org/10.1016/j.jare.2015.02.001.

Faes, C., Molenberghs, G., Aerts, M., Verbeke, G., & Kenward, M. G. (2009). The effective sample size and an alternative small-sample degrees-of-freedom method. *The American Statistician*, *63*(4), 389–399. https://doi.org/10.1198/tast.2009.08196.

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

Fox, J.-P., Mulder, J., & Sinharay, S. (2017). Bayes factor covariance testing in item response models. *Psychometrika*, *82*(4), 979–1006. https://doi.org/10.1007/s11336-017-9577-6.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409. https://doi.org/10.2307/2289776.

Gelfand, A. E., Smith, A. F. M., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*(418), 523–532. https://doi.org/10.2307/2290286.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*. https://doi.org/10.1214/06-BA117A.

Glas, C. A. W., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *The British Journal of Mathematical and Statistical Psychology*, *63*(Pt 3), 603–626. https://doi.org/10.1348/000711009X481360.

Goldhammer, F., & Krohne, U. (2014). Controlling individuals time spent on task in speeded performance measures: Experimental time limits, posterior time limits, and response time modeling. *Applied Psychological Measurement*, *38*(4), 255–267. https://doi.org/10.1177/0146621613517164.

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626. https://doi.org/10.1037/a0034716.

Gustafson, P., Hossain, S., & MacNab, Y. C. (2006). Conservative prior distributions for variance parameters in hierarchical models. *The Canadian Journal of Statistics*, *34*(3), 377–390.

Hecht, M., Siegle, T., & Weirich, S. (2017). A model for the estimation of testlet response time to optimize test assembly in paper-and-pencil large-scale assessments. *Journal for Educational Research Online*, *9*(1), 32–51.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2008). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*(1), 54–75.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/10.1037/1082-989X.10.4.477.

Lange, K. (2010). *Numerical analysis for statisticians* (2nd ed.). Berlin: Springer.

Lee, Y., & Neider, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, *19*(2), 219–228.

Lesaffre, E., & Lawson, A. B. (2012). *Bayesian biostatistics*. New York: Wiley.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*(7), 519–537. https://doi.org/10.1177/0146621608329504.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.2307/2336267.

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*(2), 304–327. https://doi.org/10.1111/bmsp.12020.

Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, *78*(3), 384–408. https://doi.org/10.1177/0013164417693666.

Molenaar, D., Oberski, D., Vermunt, J., & Boeck, P. D. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*(5), 606–626. https://doi.org/10.1080/00273171.2016.1192983.

Molenaar, D., Tuerlinckx, F., & Maas, H. (2014). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12042.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). 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.

Muth, C., Bales, K. L., Hinde, K., Maninger, N., Mendoza, S. P., & Ferrer, E. (2016). Alternative models for small samples in psychological research: Applying linear mixed effects models and generalized estimating equations to repeated measures data. *Educational and Psychological Measurement*, *76*(1), 64–87.

Overholser, R., & Xu, R. (2014). Effective degrees of freedom and its application to conditional AIC for linear mixed-effects models with correlated error structures. *Journal of multivariate analysis*, *132*, 160–170. https://doi.org/10.1016/j.jmva.2014.08.004.

Perrakis, K., Ntzoufras, I., & Tsionas, E. G. (2014). On the use of marginal posteriors in marginal likelihood estimation via importance sampling. *Computational Statistics & Data Analysis*, *77*, 54–69. https://doi.org/10.1016/j.csda.2014.03.004.

Pryseley, A., Tchonlafi, C., Verbeke, G., & Molenberghs, G. (2011). Estimating negative variance components from Gaussian and non-Gaussian data: A mixed models approach. *Computational Statistics and Data Analysis*, *55*(2), 1071–1085. https://doi.org/10.1016/j.csda.2010.09.002.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Ranger, J., & Kuhn, J.-T. (2013). Analyzing response times in tests with rank correlation approaches. *Journal of Educational and Behavioral Statistics*, *38*(1), 61–80. https://doi.org/10.3102/1076998611431086.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136.

Su, Y.-S. (2015). R2jags. Retrieved February 21, 2018 from https://CRAN.R-project.org/package=R2jags.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 179–203). San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50019-6.

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

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

van der Linden, W. J., & Fox, G. J. A. (2016). Joint hierarchical modeling of responses and response times. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models*, pp. 481–500. Chapman and Hall/CRC Press.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001.

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x.
Wang, C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149.
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. https://doi.org/10.1111/bmsp.12054.
Wood, S. N. (2013). A simple test for random effects in regression models. *Biometrika*, *100*(4), 1005–1010. https://doi.org/10.1093/biomet/ast038.

Manuscript Received: 22 MAR 2018
Published Online Date: 16 MAY 2019

https://doi.org/10.1007/s11336-019-09671-8 Published online by Cambridge University Press