frontiers in Psychology

METHODS
published: 05 August 2019
doi: 10.3389/fpsyg.2019.01675

Bayesian Covariance Structure Modeling of Responses and Process Data

Konrad Klotzke and Jean-Paul Fox*

Faculty of BMS, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, Enschede, Netherlands

A novel Bayesian modeling framework for response accuracy (RA), response times (RTs) and other process data is proposed. In a Bayesian covariance structure modeling approach, nested and crossed dependences within test-taker data (e.g., within a testlet, between RAs and RTs for an item) are explicitly modeled. The local dependences are modeled directly through covariance parameters in an additive covariance matrix. The inclusion of random effects (on person or group level) is not necessary, which allows constructing parsimonious models for responses and multiple types of process data. Bayesian Covariance Structure Models (BCSMs) are presented for various well-known dependence structures. Through truncated shifted inverse-gamma priors, closed-form expressions for the conditional posteriors of the covariance parameters are derived. The priors avoid boundary effects at zero, and ensure the positive definiteness of the additive covariance structure at any layer. Dependences of categorical outcome data are modeled through latent continuous variables. In a simulation study, a BCSM for RAs and RTs is compared to van der Linden's hierarchical model (LHM; van der Linden, 2007). Under the BCSM, the dependence structure is extended to allow variations in test-takers' working speed and ability and is estimated with a satisfying performance. Under the LHM, the assumption of local independence is violated, which results in a biased estimate of the variance of the ability distribution. Moreover, the BCSM provides insight in changes in the speed-accuracy trade-off. With an empirical example, the flexibility and relevance of the BCSM for complex dependence structures in a real-world setting are discussed.

Keywords: process data, educational measurement, Bayesian modeling, covariance structure, marginal modeling, cross-classification, response times, latent variable modeling

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Paul De Boeck,
The Ohio State University, United States
Maria Bolsinova,
University of Amsterdam, Netherlands

****Correspondence:***
Konrad Klotzke
k.klotzke@utwente.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *15 September 2018*
Accepted: *02 July 2019*
Published: *05 August 2019*

Citation:
Klotzke K and Fox J-P (2019) Bayesian Covariance Structure Modeling of Responses and Process Data. Front. Psychol. 10:1675. doi: 10.3389/fpsyg.2019.01675

1. INTRODUCTION

Computer-based assessments (CBAs) provide the opportunity to gather responses times (RTs) and other process data in addition to the test-takers' responses. Empirical research has shown that in combination with response patterns, RTs can lend valuable insight into interesting test-taker, item and test characteristics, such as pre-knowledge of items, motivation, time-pressure or differential speededness (Bridgeman and Cline, 2004; Wise and Kong, 2005; Meijer and Sotaridona, 2006; van der Linden et al., 2007; van der Linden and Guo, 2008; Marianti et al., 2014; Qian et al., 2016). New types of process data have been explored lately that carry the potential to lend additional insight into (latent) response processes and to improve inferences about constructs of interest (e.g., Azevedo, 2015; He et al., 2016; Goldhammer and Zehner, 2017; Maddox, 2017). To make valid inferences

from process data, innovative joint models are needed that are capable of utilizing test-taker data beyond RAs and RTs, while accounting for complex relationships in multiple data types.

An important concept is the speed-accuracy trade-off, which states that, on average, a test-taker's ability suffers from an increased working speed (van der Linden, 2009). Test scores depend on test-takers' speed during the test and ignoring this within-subject relationship threatens the validity of inferences about their ability level. In experimental cognitive psychology, the speed-accuracy trade-off can be modeled for individual persons as the relationship between the proportion of correct tasks and the average time spent on the tasks (Luce, 1986). In educational measurement, learning effects can be expected when presenting the same item to a test-taker multiple times (Butler, 2010). Hence, in practical applications, often only a single measurement of RA and RT is obtained for each combination of test-taker and item. Therefore, it is common to assume a certain homogeneity in the speed-accuracy trade-off within a group of test-takers and how they are affected by the condition of interest (Thissen, 1983; Klein Entink et al., 2008; Glas and van der Linden, 2010; Ranger and Kuhn, 2013; Goldhammer and Kroehne, 2014; Goldhammer et al., 2014; Loeys et al., 2014; Molenaar et al., 2015; van der Linden and Fox, 2016). Alternatively, in certain experimental settings, the researcher can control the test-takers' working speed (by imposing time limits) and thereby exclude the person-level working speed variable from the regression equation (Goldhammer and Kroehne, 2014).

More general and flexible approaches to model and test the within-subject dependence structure have been achieved through the generalized linear mixed model (GLMM) (McCulloch, 2003) and mixture models. The within-subject mixture models allow subject-specific changes in the speed-accuracy trade-off across different states. However, in practice the number of states is very limited (Wang and Xu, 2015; Molenaar et al., 2016) to obtain identifiable and stable estimation results. In GLMMs, the measurement model for the RAs or the RTs is extended by including either the person level variable (ability or working speed) or the dependent variable of the respective other measurement model as a covariate in the regression equation. Item-specific person-level and person-specific item-level variables allow the speed-accuracy trade-off to vary between items and allow item parameters to vary across persons, respectively (e.g., Goldhammer et al., 2014, 2015). Furthermore, a non-linear relation between RAs and RTs can be specified (e.g., Molenaar et al., 2015; Bolsinova and Molenaar, 2018).

However, the complexity of a GLMM is drastically increased when including other process data and extending the GLMMs with additional person-level variables. It is therefore questionable whether the GLMM approach can manage the challenges of utilizing new types of process data in complex CBAs. Currently, GLMMs are limited in the amount of process data information that can be utilized to make inferences due to restrictions on the model complexity and the sample size. Furthermore, GLMMs are also limited in how the information is utilized. For instance, correlations between RAs and different types of process data may vary depending on item characteristics or test design. In that case, interaction effects are needed to model item and/or testlet-specific dependences, but this will significantly increase the complexity of the GLMM. To prevent over-parameterization and weak numerical stability, techniques such as principle component analysis, latent class analysis, or various model selection algorithms (e.g., backward elimination, forward selection or all subsets regression) (Thomas, 2002; Efron et al., 2004; Wetzel et al., 2015) have been proposed to reduce the number of covariates in the regression equation. However, this complicates a straightforward modeling approach and can lead to arbitrary assumptions and *ad hoc* decisions. It is well-known that ignoring correlations in test-taker data may cause violations of local independence assumptions and can result in biased inferences about parameters, the reliability of the test, and hinder test equating (e.g., Yen, 1984; Ackerman, 1987; Chen and Thissen, 1997; Bradlow et al., 1999; Baker and Kim, 2004; Jiao et al., 2005, 2012; Wang and Wilson, 2005; Wainer et al., 2007). Therefore, when including new types of process data, care must be taken in modeling the dependence structure to avoid making biased inferences.

The proposed Bayesian Covariance Structure Model (BCSM) can handle different types of nested and cross-classified dependence structures for multiple types of test-taker data. The BCSM extends the marginal model for hierarchically structured item RT data of Klotzke and Fox (2018). In the model of Klotzke and Fox (2018), dependences that follow from nested classifications (e.g., item clusters in a testlet design) are directly modeled as covariances without including random effects. The methodology is extended to classifications across multiple data types. Thus, in addition to modeling nested classifications (within a data type), relationships in data across different types (e.g., RTs and dichotomous responses) are modeled through cross-classifications in the dependence structure. In the same manner as the nested classifications, crossed classifications are modeled explicitly as covariance parameters. Without the inclusion of random effects, the parsimony of the BCSM is preserved, where dependences between each cluster of observations can be modeled with a single covariance parameter. The BCSM assumes a multivariate normal distribution for the data, either directly or through a threshold specification (i.e., for categorical or count data), and allows distinct modeling of the mean and covariance structure. The BCSM parameters can be estimated with an efficient Gibbs-sampling algorithm, even for a reasonably small sample size. Modeling local dependences via covariance parameters instead of modeling dependences through random effects (i.e., the random effect variance defines the covariance between clustered observations) has two advantages: first, covariances can be negative or positive, which allows more flexibility in specifying complex dependence structures than random effect variances. The latter can only model positive dependences. Second, tests for local independence under the BCSM framework do not require testing at the boundary of the parameter space (i.e., the null hypothesis states that the covariance parameter is equal to zero). This stands in contrast to a random effect variance, which is *a-priori* restricted to be positive. In the BCSM, this means that the prior distributions for the covariance parameters are less informative, i.e., they don't assume beforehand that the covariance parameters are

a test-taker's RTs are grouped by the latent factor working speed, and the RAs are grouped by the latent factor ability. Furthermore, observations are grouped across the two data types on a person level, which represents a correlation between a test-taker's ability and working speed. **Table 1** shows the classification matrix and covariance parameters of the BCSM for speed and ability.

2.2. Variable Speed-Accuracy Trade-Off

For the variable speed-accuracy trade-off model, the BCSM for speed and ability is extended with an item-specific cross-covariance between a test-taker's RTs and RAs. This allows to investigate how the speed-accuracy trade-off within a group of test-takers varies between items. Thereby, a certain homogeneity in the relevant response processes is assumed, which leads to test-takers within a group sharing a common speed-accuracy trade-off. The classification diagram for the variable speed-accuracy trade-off model is shown in **Figure 1**. **Table 2** extends **Table 1** with the additional classification rules and covariance parameters implied by a variable speed-accuracy trade-off.

2.3. Blocked Structures of Cross-Covariances

Just as the variable speed-accuracy trade-off model, the blocked structures of cross-covariances model extends the BCSM for speed and ability with a varying cross-covariance between a test-taker's RTs and RAs. However, the cross-covariance is defined to change per blocks of (here: two) items. A possible application for this model is test-taking under varying time-pressure conditions. In such a scenario, it is reasonable to assume local dependence for components (i.e., RTs and RAs) within a block of items that belong to the same time-pressure condition. In the variable speed-accuracy trade-off model on the other hand, the local dependence is defined per individual item. **Table 3** extends **Table 1** with the additional classification rules and covariance parameters of the blocked structures of cross-covariances model.

2.4. Differential Blocked Structures of Cross-Covariances Across Factors

The within-subject dependence structure can also be specified for components within a single data type. In the differential blocked structures of cross-covariances across factors model, the variable speed-accuracy trade-off model is extended with a separate testlet structure for each the RTs and the RAs. The testlet structures are defined independently of each other. **Table 4** extends **Tables 1**, **2**

TABLE 1 | The additive covariance structure of the BCSM for speed and ability is implied by the random effects structure of the LHM with binary factor loadings.

	Classification matrix u											
Covariance	**Response times**						**Response accuracies**					
δ	1	1	1	1	1	1	0	0	0	0	0	0
τ	0	0	0	0	0	0	1	1	1	1	1	1
ϕ	1	1	1	1	1	1	1	1	1	1	1	1

TABLE 2 | The additive covariance structure of the variable speed-accuracy trade-off model is an extension of the BCSM for speed and ability with item-specific cross-covariances between RTs and RAs.

	Classification matrix u											
Covariance	**Response times**						**Response accuracies**					
ν_1	1	0	0	0	0	0	1	0	0	0	0	0
ν_2	0	1	0	0	0	0	0	1	0	0	0	0
ν_3	0	0	1	0	0	0	0	0	1	0	0	0
ν_4	0	0	0	1	0	0	0	0	0	1	0	0
ν_5	0	0	0	0	1	0	0	0	0	0	1	0
ν_6	0	0	0	0	0	1	0	0	0	0	0	1

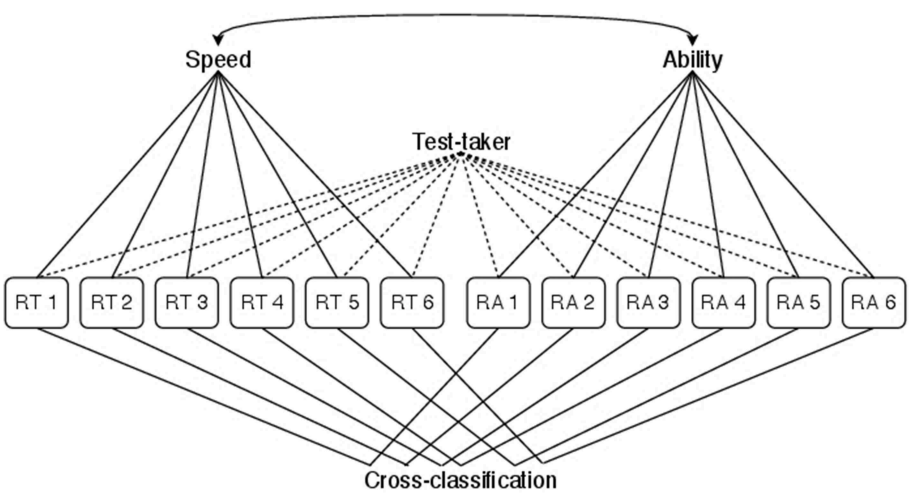

FIGURE 1 | Classification diagram for the variable speed-accuracy trade-off model. The classification implied by the LHM is extended by grouping components item-wise. This allows the group level speed-accuracy trade-off to vary between items.

TABLE 3 | The additive covariance structure of the blocked structures of cross-covariances model is an extension of the BCSM for speed and ability with block-wise cross-covariances between RTs and RAs.

Covariance	Classification matrix u											
	Response times						Response accuracies					
ν_1	1	1	0	0	0	0	1	1	0	0	0	0
ν_2	0	0	1	1	0	0	0	0	1	1	0	0
ν_3	0	0	0	0	1	1	0	0	0	0	1	1

TABLE 4 | The additive covariance structure of the differential blocked structures of cross-covariances across factors model is an extension of the variable speed-accuracy trade-off model with independent testlet structures for separate data types.

Covariance	Classification matrix u											
	Response times						Response accuracies					
Δ_1	1	1	0	0	0	0	0	0	0	0	0	0
Δ_2	0	0	1	1	0	0	0	0	0	0	0	0
Δ_3	0	0	0	0	1	1	0	0	0	0	0	0
Δ_4	0	0	0	0	0	0	1	1	1	0	0	0
Δ_5	0	0	0	0	0	0	0	0	0	1	1	1

with the additional classification rules and covariance parameters of the differential blocked structures of cross-covariances across factors model.

2.5. More Than Two Data Types

A BCSM is not limited to RTs and responses. Additional process data can carry information relevant to the research. In this example, additional process data is available for each combination of test-taker and item. Therefore, $p = 6$ components are added to the model. In the illustrated model, an item-specific cross-covariance between components of all types is assumed. That means for example that RTs and RAs to an item may correlate in a different way than RAs and process data, to the same item. Furthermore, ϕ_1, ϕ_2, and ϕ_3 represent the 3-by-3 covariance of the three latent factors (e.g., ability, working speed, and speed first action) that are related to the three types of data. **Table 5** shows the classification matrix and covariance parameters of the more than two data types model.

2.6. Model Scalability

The models constructed in the BCSM framework are scalable with respect to the length of the test, the number of data types and the specified dependence structure. The number of columns of $\boldsymbol{u}$ corresponds to the number of data components (N_c). If a single observation is available for each combination of test-taker, item and data type, the number of data components is the product of the number of items (p) and the number of data types (N_d), i.e., $N_c = p * N_d$. Consequently, extending the test length with one item increases the number of columns of $\boldsymbol{u}$ by N_d. Similarly, introducing an additional data type increases the number of columns by p.

TABLE 5 | The additive covariance structure for a BCSM that incorporates additional process data, next to RTs and RAs.

Covariance	Classification matrix u																	
	Response times						Response accuracies						Process data					
δ	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
τ	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0
ω	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
ϕ_1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
ϕ_2	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1
ϕ_3	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
ν_1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ν_2	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ν_3	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ν_4	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ν_5	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
ν_6	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
ν_7	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ν_8	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ν_9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ν_{10}	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ν_{11}	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
ν_{12}	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
ν_{13}	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
ν_{14}	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
ν_{15}	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
ν_{16}	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
ν_{17}	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
ν_{18}	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1

The number, if any, of additional rows of $\boldsymbol{u}$ depends on the specified classification structure. For example, under the structure specified in **Table 1**, a change in the number of data components does not affect the number of rows of $\boldsymbol{u}$. Instead, the existing groupings are extended to include the new data components.

In other situations, the number of groupings depends on the number of data components. For example, given the item-specific cross-classifications as defined in **Table 2**, each additional item leads to one additional classification rule (the RA and RT of a test-taker to one item are grouped together) and therefore inserts one row into $\boldsymbol{u}$. Thus, if the variable speed-accuracy trade-off joint-model is applied to a test with $p_2 = 100$ instead of $p_1 = 10$ items, the number of columns increases by $(p_2 - p_1) * N_d = (100 - 10) * 2 = 180$ and the number of rows increases by $p_2 - p_1 = 90$.

3. CATEGORICAL OUTCOME DATA

When recording the test-takers' responses during a test, discrete realizations of latent response variables are observed. The multivariate normally distributed RA data (latent responses) are linked through a threshold specification to their discrete

realizations. However, truncating a multivariate normal distribution in high dimensions is non-trivial (Botev, 2017) and simply truncating independently for each dimension does not lead to the intended multivariate joint-distribution (Horrace, 2005).

The proposed solution is to derive the univariate normal distribution of each latent response component, conditional on all other components. The univariate normal distribution is derived by partitioning the additive covariance structure $\boldsymbol{\Sigma}$, as defined in Equation 1, into four parts. The upper left part, $\boldsymbol{B}_{11}$, gives the variance of the k-th component and the diagonal parts, $\boldsymbol{B}_{12}$ and $\boldsymbol{B}_{21}$, contain the covariance of the k-th component with the remaining components. Finally, $\boldsymbol{B}_{22}$ describes the covariance structure of all components but the k-th:

$$\boldsymbol{\Sigma} = \begin{matrix} \\ Y_{ik} \\ \tilde{\boldsymbol{Y}}_i \end{matrix} \overset{\begin{matrix} Y_{ik} & \tilde{\boldsymbol{Y}}_i \end{matrix}}{\begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix}}, \tag{2}$$

where $\boldsymbol{Y}$ is a $N \times N_c$-dimensional matrix, containing data from all N_c components and N test-takers. A tilde, i.e., a $\sim$, above a vector or matrix indicates that the k-th component is excluded from the data structure. Based on the partitioned covariance matrix, the means and variance of the conditionally univariate normal distribution of the k-th component are derived for each test-taker:

$$\boldsymbol{\mu}_{Y_k|\tilde{Y}} = \boldsymbol{\mu}_{Y_k} + \boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1}(\tilde{\boldsymbol{Y}} - \boldsymbol{\mu}_{\tilde{Y}}), \tag{3}$$

$$\sigma^2_{Y_k|\tilde{Y}} = \boldsymbol{B}_{11} - \boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1}\boldsymbol{B}_{21}. \tag{4}$$

A closed-form expression for $\boldsymbol{B}_{22}^{-1}$ is derived through the Sherman-Morrison formula (e.g., Lange, 2010, p. 261):

$$\boldsymbol{A}_{t+1}^{-1} = (\boldsymbol{A}_t + \lambda\boldsymbol{v}\boldsymbol{v}^T)^{-1} = \boldsymbol{A}_t^{-1} - \frac{\boldsymbol{A}_t^{-1}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{A}_t^{-1}}{1/\lambda + \boldsymbol{v}^T\boldsymbol{A}_t^{-1}\boldsymbol{v}}, \tag{5}$$

where $\boldsymbol{A}_t^{-1} = \tilde{\boldsymbol{\Sigma}}_t^{-1}$ is the inverse of the additive covariance structure for all but the k-th component at the t-th layer, $\lambda = \theta_{t+1}$ is the covariance parameter for the added layer and $\boldsymbol{v} = \tilde{\boldsymbol{u}}_{t+1}$ contains the classification structure for the new layer. Given that the inverse of $\boldsymbol{A}_0^{-1} = \tilde{\boldsymbol{\Sigma}}_0^{-1}$, i.e., the inverse of the diagonal matrix consisting of the measurement error variance parameters for all but the k-th component, is known, the inverse for any additional layer can be derived recursively.

4. BAYESIAN INFERENCE

In line with the approach suggested by Fox et al. (2017) and Klotzke and Fox (2018), closed-form expressions for the conditional posterior distributions of the variance and covariance parameters are derived through truncated shifted inverse-gamma priors. For each of the N_t layers of the additive covariance matrix, a truncation point tr_t is derived by applying the Sherman-Morrison formula (Lange, 2010, p. 260–261). Enforcing the truncation through the indicator function $\mathbb{1}_{tr}$ ensures that the covariance matrix is positive definite at any layer t. This leads to a lower bound for each covariance parameter ($\theta_t > tr_t$) conditional on the classification structure and the inverse of the covariance matrix at the underlying layer ($t-1$):

$$tr_t = -1/\boldsymbol{u}_t^T\boldsymbol{\Sigma}_{t-1}^{-1}\boldsymbol{u}_t. \tag{6}$$

For the measurement error variance parameters (the diagonal terms of $\boldsymbol{\Sigma}_0$) a truncation sets the probability of negative values a-priori to zero.

The reasoning behind the shift parameters is based upon two premises: (1) a draw of θ_t is obtained through sampling $\theta_t + \psi_t$ and subtracting the shift parameter ψ_t iteratively within the Markov chain Monte Carlo (MCMC) (Gilks et al., 1995) algorithm, (2) the probability distribution of $\theta_t + \psi_t$ must incorporate all information that is available in the data about θ_t. It is shown in Equations (7) and (8) that the probability distribution of the person level means across that are grouped together in $\boldsymbol{u}_t$ contains all available information about the covariance parameter θ_t. Note that the person level means are constructed as the mean of (correlated) random normal variables and are therefore univariate normally distributed.

Conditional on the classification structure and the additive covariance matrix at its highest layer ($\boldsymbol{\Sigma}_{N_t}$), the variance of the person level means is derived through the property that the variance of the sum of correlated random variables is the sum of their covariances:

$$\begin{aligned} Var\left(\bar{Y}_{i(k\in\boldsymbol{u}_t)}|\boldsymbol{\Sigma}_{N_t},\boldsymbol{u}\right) &= Var\left(\sum_{k\in\boldsymbol{u}_t} Y_{ik}/\left(\mathbf{1}_{N_c}^T\boldsymbol{u}_t\right)\right) \\ &= \left[\left(\sum_{k=1}^{N_c}\sigma_k^2 u_{tk} + \theta_t\left(\mathbf{1}_{N_c}^T\boldsymbol{u}_t\right)^2 \right.\right. \\ &\quad \left.\left. + \sum_{j\neq t}\theta_j\left(\mathbf{1}_{N_c}^T(\boldsymbol{u}_j\odot\boldsymbol{u}_t)\right)^2\right)/\left(\mathbf{1}_{N_c}^T\boldsymbol{u}_t\right)^2\right] \\ &= \theta_t + \left[\left(\sum_{k=1}^{N_c}\sigma_k^2 u_{tk} + \sum_{j\neq t}\theta_j\left(\mathbf{1}_{N_c}^T(\boldsymbol{u}_j\odot\boldsymbol{u}_t)\right)^2\right)/\left(\mathbf{1}_{N_c}^T\boldsymbol{u}_t\right)^2\right] \\ &= \theta_t + \psi_t, \end{aligned} \tag{7}$$

where $\odot$ denotes the Hadamard product and $\mathbf{1}_{N_c}$ is a N_c-dimensional vector of ones. A sufficient statistic for $Var\left(\bar{Y}_{i(k\in\boldsymbol{u}_t)}|\boldsymbol{\Sigma}_{N_t},\boldsymbol{u}\right) = \theta_t + \psi_t$ is therefore the sum of squares of the deviations of the conditional person level means from the conditional grand mean,

$$SSB_t = \sum_{i=1}^{N}\left(\bar{Y}_{i(k\in\boldsymbol{u}_t)} - \bar{Y}_{.(k\in\boldsymbol{u}_t)}\right)^2. \tag{8}$$

Similarly, the within-component sum of squares is a sufficient statistic for $Var(Y_{ik}) = \sigma_k^2 + \sum_{t=1}^{N_t}\theta_t u_{tk}$, namely

$$SSW_k = \sum_{i=1}^{N}\left(Y_{ik} - \bar{Y}_{.k}\right)^2. \tag{9}$$

From Equations (8) and (9) follow $N_t + N_c$ sufficient statistics for the N_t covariance and N_c variance parameters, out of which the additive covariance structure, as specified in Equation (1), is composed. The model is therefore identified under the condition that the rows of the classification matrix $\boldsymbol{u}$ are mutually distinct.

The truncated shifted inverse-gamma prior extends the default inverse-gamma prior for variance components with a shift and a truncation parameter; the former allowing a covariance parameter to take on negative values, the latter ensuring the positive definiteness of the additive covariance matrix at any layer:

$$IG(x, \alpha_0, \beta_0, \psi_t, tr_t) = \left[\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(x+\psi_t)^{-\alpha_0-1} exp\left(-\frac{\beta_0}{x+\psi_t}\right)\right] \cdot \mathbb{1}_{tr}(x > tr_t), \tag{10}$$

where the truncation point (tr_t) and shift parameter (ψ_t) are computed according to Equations (6) and (7).

Note that conjugacy between the extended inverse-gamma prior and the likelihood function of a normal distribution is preserved, thus leading to truncated shifted inverse-gamma posteriors for the covariance and measurement error variance parameters:

$$\theta_t \sim IG(x, \alpha_0 + N/2, \beta_0 + SSB_t/2, \psi_t, tr_t), \tag{11}$$

$$\sigma_k^2 \sim IG(x, \alpha_0 + N/2, \beta_0 + SSW_k/2, \sum_{t=1}^{N_t} \theta_t u_{tk}, 0). \tag{12}$$

The a-priori restriction of $\sigma_k^2 > 0$ is thus enforced by fixing the truncation point for the measurement error variance parameters to zero.

See Appendix A in **Supplementary Material** for an outline of the MCMC algorithm and the corresponding sampling steps.

5. SIMULATION STUDY

In a simulation study, the within-subject dependence structure under a model for RTs and dichotomous responses is estimated. A comparison is made between a BCSM and the LHM. In the BCSM framework, the dependence structure is directly modeled in an additive covariance matrix. In the LHM framework, the dependence structure is implied by the random effect structure and in particular the random effect variances. Therefore, the focus of this simulation study is the precision and bias of the (co)variance parameter estimates.

In the simulated experiment, across two conditions, $N = 200$ and $N = 1,000$ randomly selected persons are taking a test that consists of $p = 12$ items. Furthermore, the time-pressure on the test-takers systematically changes after every two items. This is assumed to affect the response processes within the group of test-takers over the course of the test. For example, under a perceived high time-pressure, guessing may become more likely. The change in response processes is reflected by the within-subject dependence structure, i.e., the speed-accuracy trade-off may vary between blocks of two items and is common across test-takers.

The length of the test is fixed across the 100 replications for both conditions of the simulation. Within each condition, all test-takers are part of the same group. Within each replication, test-taker data are generated and the BCSM as well as the LHM are fitted with 5000 MCMC iterations and a burn-in phase of 10%. The LHM is fitted using the R-package LNIRT (Fox et al., 2018).

5.1. LHM for Fixed Speed and Ability

On the first level of the hierarchical framework, separate measurement models for the RTs and RAs are specified. The item discrimination parameters are fixed to 1, which gives the following first level models for the RTs (RT) and RAs (RA) of test-taker i and item k:

$$RT_{ik} = \beta_k - \zeta_i + e_{RT_{ik}}, \tag{13}$$

$$RA_{ik} = \theta_i - b_k + e_{RA_{ik}}, \tag{14}$$

where $\zeta_i \sim \mathcal{N}(\mu_\zeta, \delta)$ and $\theta_i \sim \mathcal{N}(\mu_\theta, \tau)$ are random variables on a person level, representing the variation in working speed and ability between test-takers. The time intensity and item difficulty parameters β_k and b_k are item level intercepts and are not given further attention in this simulation study. Finally, $e_{RT_{ik}} \sim \mathcal{N}(0, \sigma_k^2)$ and $e_{RA_{ik}} \sim \mathcal{N}(0, 1)$ are the measurement errors. On the second level, a model for the joint-distribution of the person parameters (working speed and ability) is defined:

$$\boldsymbol{\Sigma}_p = \begin{pmatrix} \delta + \phi & \phi \\ \phi & \tau + \phi \end{pmatrix}. \tag{15}$$

Note that the LHM assumes a constant working speed and ability across the test for a test-taker. From this follows a test-wide cross-covariance between a test-taker's RTs and RAs ϕ.

5.2. BCSM for Variable Speed and Ability

In the BCSM, the within-subject dependence structure is modeled directly in an additive covariance structure with 9 layers. The covariance structure is defined in Equation (1), where $\boldsymbol{\theta} = \{\delta, \tau, \phi, \nu_1, \ldots, \nu_6\}$ are the (cross-)covariance parameters and the classification matrix is specified in **Table 6**. A truncated shifted inverse-gamma prior with $shape = 10^{-3}$ and $scale = 10^3$ is defined for the variance and covariance parameters.

5.3. Data Generation

Data are generated under a generalization of the models specified in Equations (13)–(15) that allows the test-takers' working speed and ability to vary over the course of the test:

$$RT_{ik} = \beta_k - \zeta_{it(k)} + e_{RT_{ik}}, \tag{16}$$

$$RA_{ik} = \theta_{it(k)} - b_k + e_{RA_{ik}}. \tag{17}$$

$$\boldsymbol{\Sigma}_{pk} = \begin{pmatrix} \delta + \phi + \nu_{t(k)} & \phi + \nu_{t(k)} \\ \phi + \nu_{t(k)} & \tau + \phi + \nu_{t(k)} \end{pmatrix}, \tag{18}$$

where $t(k)$ denotes item k in classification group t. The population values of the (co)variance parameters are $\delta = 0.5$, $\tau = 0.5$, $\phi = 0.5$ and $\boldsymbol{\nu} = \{0, -0.05, -0.1, 0.4, 0.2, 0.3\}$. The item level intercepts ($\boldsymbol{\beta}$ and $\boldsymbol{b}$) are set to zero. Finally, the population values of the measurement error variances are generated from a uniform distribution with lower bound 0.5 and upper bound 1.5.

TABLE 6 | The additive covariance structure of the BCSM allows a varying speed-accuracy trade-off between blocks of two items.

Covariance	Classification matrix u																							
	Response times												Response accuracies											
δ	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
τ	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
ϕ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ν_1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
ν_2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
ν_3	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ν_4	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
ν_5	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
ν_6	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1

TABLE 7 | Means and standard deviations of posterior mean estimates across 100 simulated replications of data for 200 and 1,000 test-takers and 12 items.

	Mean (SD) of posterior mean estimates			
	N = 200		N = 1,000	
Cov	BCSM	LHM	BCSM	LHM
$\delta = 0.5$	0.49 (0.03)	0.51 (0.03)	0.50 (0.01)	0.52 (0.01)
$\tau = 0.5$	0.51 (0.07)	0.45 (0.06)	0.48 (0.03)	0.42 (0.03)
$\phi = 0.5$	0.51 (0.03)	0.50 (0.03)	0.50 (0.02)	0.49 (0.01)
$\nu_1 = 0$	0.00 (0.04)		−0.01 (0.02)	
$\nu_2 = -0.05$	−0.05 (0.04)		−0.06 (0.03)	
$\nu_3 = -0.1$	−0.07 (0.03)		−0.11 (0.03)	
$\nu_4 = 0.4$	0.39 (0.06)		0.39 (0.03)	
$\nu_5 = 0.2$	0.18 (0.05)		0.19 (0.02)	
$\nu_6 = 0.3$	0.28 (0.06)		0.30 (0.02)	

A comparison is made between a BCSM and the LHM. In the BCSM framework, the full within-subject dependence structure is modeled.

5.4. Results

Under the LHM, the test-wide cross-covariance and the variance of the test-taker working speed distribution are successfully estimated. The variance of the ability distribution (τ) is underestimated for both sample size conditions under the LHM, which can be attributed to ignoring the block-wise deviations from the test-wide cross-covariance. Under the BCSM, the full within-subject dependence structure is successfully estimated. Cross-covariances near zero (ν_1, ν_2, and ν_3) are estimated without bias regardless of sample size, which can be attributed to the non-informative truncated shifted inverse-gamma priors. The standard deviations of the posterior mean estimates are comparable for both models. Increasing the sample size leads to smaller standard deviations of the posterior mean estimates for both models. Under the BCSM, an average correlation of 0.99 (SD: 0.01) is observed under both conditions between the simulated measurement error variance parameters and their posterior mean estimates. The results of the simulation study are summarized in **Table 7**.

6. EMPIRICAL EXAMPLE: PIAAC 2012

The Programme for the International Assessment of Adult Competencies (PIAAC) study deploys a computer-based large scale assessment to gain insight into adult competencies across the domains of numeracy, literacy and problem solving (OECD, 2013). The computer-based nature of the assessment allows recording behavioral process data, in addition to the scored responses. It is assumed that the process data correlate with the scored responses and therefore contain information about the latent competencies of interest. Describing these correlations requires paying attention to local dependences within the data. Local dependences follow from shared item characteristics (e.g., response mode), the test design (e.g., testlets), the manner the process data is obtained (e.g., a single measurement per type, test-taker and item, multiple measurements or aggregated data) and the latent factor structure (e.g., data components load on test-takers' ability and working speed). Furthermore, test-taker characteristics such as computer experience or gender may affect the associations of data components (e.g., the correlation of RTs and RAs of an item may differ between test-takers with and without computer experience). It will be shown that a BCSM can be constructed that (a) takes the complex dependence structure within test-taker data into account, (b) allows correcting for between-subject differences in the dependence structure by including test-taker background variables, and (c) can be estimated given a reasonable sample size.

6.1. Data Set

The data set consists of responses and process data for $N = 745$ Canadian test-takers and $p = 15$ items. For each combination of item and test-taker, three data points are available: the scored dichotomous response, the total (log) RT it took the test-taker to complete the item and the (log) time it took the test-taker until they took their first action on that item. Nine of the items measure numeracy competencies, the remaining six items measure literacy competencies. Furthermore, the items differ in their response mode. See **Table 8** for an overview of the included items and their characteristics. Moreover, the test-takers' gender (0: male, 1: female), computer experience (0: no, 1: yes), whether or not they are a native speaker (0: no, 1: yes)

TABLE 8 | Id, name, domain, and response mode of the 15 PIAAC items included in the data analysis of the empirical example.

Item id.	Name	Domain	Response mode
1	Wine 1	Numeracy	Number match
2	Wine 2	Numeracy	Stimulus clicking
3	Gas gauge	Numeracy	Number match
4	Photo 1	Numeracy	Number match
5	Photo 2	Numeracy	Stimulus clicking
6	Photo 3	Numeracy	Exact match
7	Urban population	Numeracy	Number match
8	Tiles	Numeracy	Exact match
9	Package	Numeracy	Stimulus clicking
10	Baltic stock market 1	Literacy	Stimulus clicking
11	Baltic stock market 2	Literacy	Stimulus highlighting
12	Baltic stock market 3	Literacy	Stimulus clicking
13	Baltic stock market 4	Literacy	Stimulus clicking
14	TMN antitheft 1	Literacy	Stimulus highlighting
15	TMN antitheft 2	Literacy	Stimulus highlighting

and their educational level (1: low, 2: medium, 3: high) were recorded. Further information on test-taker demographics and item characteristics can be found in Statistics Canada (2013).

6.2. Dependence Structure

Data that are naturally grouped may be stronger correlated than (conditionally) unrelated data. In the data set at hand, items are grouped through their domain (numeracy or literacy) and their response mode (number match, exact match, stimulus clicking or stimulus highlighting). For each grouping, three layers are defined: one for each pair of data types. This allows to explore how the dependences between, respectively, RAs and RTs, RAs and times to first action (TAs), and RTs and TAs vary across item domains and response modes, while controlling for the rest of the dependence structure. Furthermore, data components that load on a common latent factor may be correlated. Latent factors are the test-taker's ability, working speed and speed first action. The correlation between the latent factors is modeled in separate layers. **Figure 2** illustrates the classifications that follow from the groupings. Data within each classification group may be locally dependent. The corresponding classification matrix for the $N_c = 45$ data components and $N_t = 24$ classification groups is shown in Appendix B (**Supplementary Material**).

6.3. Statistical Model

Under the BCSM framework, a model for response and process data is constructed. In the mean structure of the joint-model, test-taker background data are modeled as predictor variables. The dependence structure is modeled through an additive covariance matrix that defines the relationship of the multivariate normally distributed error terms:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{B} + \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}_{N_c}, \boldsymbol{\Sigma}), \tag{19}$$

where $\boldsymbol{Y} = \{\boldsymbol{RA}, \boldsymbol{RT}, \boldsymbol{TA}\}$ is a $N \times N_c$-dimensional matrix containing the RAs that underlie the scored dichotomous responses (RA), the total RTs per item for each test-taker (RT), and the time passed until the test-taker's first action per item (TA). The $N \times 5$-dimensional matrix $\boldsymbol{X}$ contains the grand-mean centered test-taker background variables (gender, computer experience, native speaker and education level) and a vector of ones as first column. $\boldsymbol{B}$ is a $5 \times N_c$ matrix containing the regression weights for each of the four covariates on the N_c data components, and the intercepts. The first column of $\boldsymbol{B}$ contains the item-specific intercepts, which can be interpreted as item difficulty, time intensity and average time to first action parameters. The weights and intercepts are thus modeled for each data component and are equal across test-takers, therefore representing fixed effects. Note that no random variance components are associated with fixed effects, whereby they don't enter the modeled dependence structure. The $N_c \times N_c$-dimensional additive covariance matrix $\boldsymbol{\Sigma}$ consists of $N_t = 24$ layers that correspond to the specified dependence structure:

$$\boldsymbol{\Sigma} = diag(\boldsymbol{\sigma}) + \sum_{t=1}^{N_t} \theta_t \boldsymbol{u}_t \boldsymbol{u}_t^T. \tag{20}$$

For the RA components, the measurement error variance parameters are fixed to one. Furthermore the scale of the IRT model is set by fixing the mean of the item-specific intercepts (i.e., the mean of the item difficulty parameters) to zero. The classification matrix $\boldsymbol{u}$ is shown in Appendix B (**Supplementary Material**). A truncated shifted inverse-gamma prior with $shape = 10^{-3}$ and $scale = 10^3$ is defined for the variance and covariance parameters. No a-priori information about the regression weights is used: the prior guesses for the scale matrix and the mean matrix of $\boldsymbol{B}$ equal the identity matrix and a matrix of zeros, respectively.

6.4. Results

The model parameters are estimated with a single MCMC chain of 55,000 iterations from which the first 15,000 iterations are discarded as burn-in period. Visual inspection of traceplots and applying the Heidelberger and Welch' criterion (Heidelberger and Welch, 1983) using the R-package coda (Plummer et al., 2016) indicate a satisfying exploration of the parameter space and do not provide evidence against convergence of the MCMC algorithm. The posterior means and standard deviations of the twenty-four covariance parameters in the additive covariance structure are summarized in **Table 9**. **Figure 3** shows the corresponding 95%-Highest Posterior Density (HPD) intervals.

Given the observed data, it can be concluded that the probability of local dependence in the ability and speed first action latent factor classification groups is at least 95%. Furthermore, a positive interdependence in the higher order relationship between RTs and TAs is found. This implies that on average, test-takers who work overall faster also lose less time before making the first move in the item solving process. The results indicate that it is necessary to model the implied covariance structure of the correlated person effects on each type of test-taker data (RAs, RTs, and TAs). The variation in the data explained on a person level that is captured by the latent factors (ability, working speed, and speed first action) and

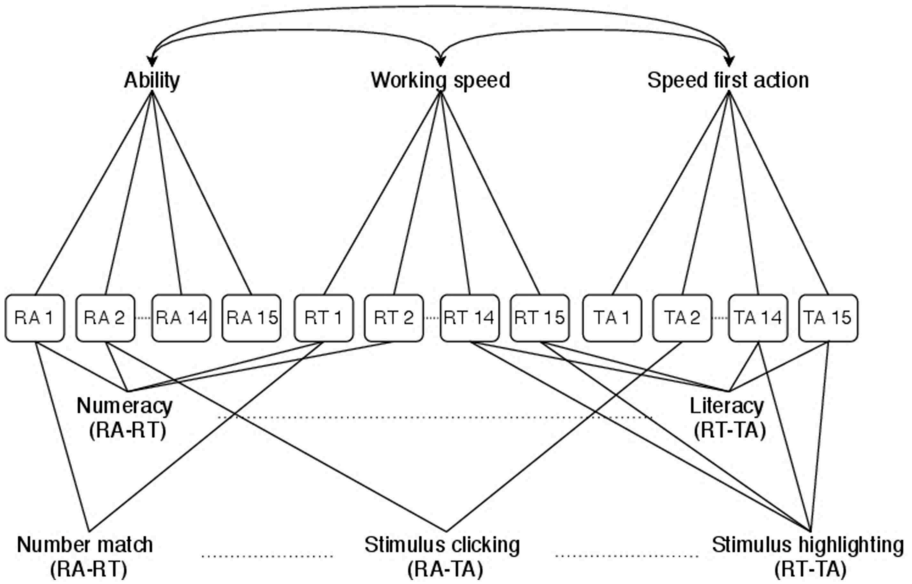

FIGURE 2 | Classification diagram for the PIAAC 2012 BCSM. The classification structure specifies dependences between scored responses and behavioral process data for varying item characteristics (domain and response mode) and a correlated latent factors structure. RA, RAs that underlie the scored dichotomous responses; RT, total RTs per item; TA, times to first action per item.

their correlation is estimated through the corresponding layers in the additive covariance structure: modeling the person effects themselves is not required.

Neither variations in the item domains, nor in the response modes caused local dependence in the data. For each domain and response mode, three sources of local dependence are independently evaluated: the relationships between, respectively, (1) RAs and RTs, (2) RAs and TAs, and (3) RTs and TAs. Modeling the 3-by-3 covariances for each specified subset of items shows that the interdependences across data types and the investigated item characteristics are sufficiently captured by the covariance layers through which the dependences of the latent factors structure are specified. It can therefore be concluded that the items' domain and response mode do not explain a noticeable amount of variance in the test-taker data when controlling for the rest of the dependence structure.

The occurrence of a vast number of covariance parameter estimates close to, or approximately equal to, zero highlights the importance of the truncated shifted inverse-gamma prior specification that avoids boundary effects by moving the edge of the parameter space away from zero. For instance, a default inverse-gamma prior would presume that $\theta_{20} > 0$ and would therefore be informative with regard to the probability of local dependence caused by the cross-relationship of RAs and TAs that belong to items with the stimulus clicking response mode: it decreases the estimated probability of local independence, i.e., the (estimated) probability that the true value of θ_{20} is zero, and can thereby provoke false conclusions about the underlying response processes. Finally, measurement error variance parameters are estimated for the fifteen RT components (mean: 0.62, SD: 0.47) and the fifteen TA components (mean: 0.31, SD: 0.19).

7. DISCUSSION

A novel Bayesian framework to model local dependences in test-taker data is proposed. The BCSM allows specifying dependences across different types of data (RAs, RTs and other process data) and multiple levels (e.g., within a testlet, clustered data per item and test-taker). The local dependences are specified through a cross-classification structure and are explicitly modeled as covariance parameters. In an additive covariance structure, nested and/or cross-classified data structures are modeled through covariance parameters.

Recording test-taker data during CBAs is not limited to scored responses and RTs. For researchers and assessors these additional process data are of utility: they can increase the precision of test-taker ability estimates and lend new insights into underlying response processes. However, using process data to draw inferences is problematic in the GLMM framework: each additional type of data requires the inclusion of new person-level variables. If interaction effects occur, the model's complexity further increases drastically. A highly complex model is prone to over-parameterization and weak numerical stability, which may strongly limit its utility in practical applications.

The BCSM framework allows the construction of parsimonious models without requiring random effects (on a person or group level) to model data dependences. Contrary to common marginal modeling approaches such as GEE, the dependence structure is however fully modeled in an additive covariance structure. This allow testing for interaction effects and to estimate the random effects *post-hoc* from the residuals of the model. By estimating random effects *post-hoc*, inferences

TABLE 9 | Posterior means and standard deviations of the $N_t = 24$ covariance parameters in the additive covariance structure.

Layer	Classification	Level	Posterior distribution	
			Mean	SD
1	Ability	Latent factor	0.47	0.16
2	Working speed	Latent factor	0.01	0.03
3	Speed first action	Latent factor	0.05	0.02
4	Ability-Working speed	Latent factor	0.01	0.01
5	Ability-Speed first action	Latent factor	−0.03	0.02
6	Working speed-Speed first action	Latent factor	0.12	0.02
7	Numeracy: RA-RT	Item domain	0.01	0.03
8	Numeracy: RA-TA	Item domain	0.05	0.03
9	Numeracy: RT-TA	Item domain	0.04	0.02
10	Literacy: RA-RT	Item domain	0.00	0.03
11	Literacy: RA-TA	Item domain	0.01	0.03
12	Literacy: RT-TA	Item domain	0.03	0.02
13	Exact match: RA-RT	Response mode	−0.01	0.17
14	Exact match: RA-TA	Response mode	0.11	0.11
15	Exact match: RT-TA	Response mode	0.07	0.10
16	Number match: RA-RT	Response mode	−0.02	0.05
17	Number match: RA-TA	Response mode	0.07	0.07
18	Number match: RT-TA	Response mode	0.04	0.04
19	Stimulus clicking: RA-RT	Response mode	0.01	0.04
20	Stimulus clicking: RA-TA	Response mode	0.00	0.03
21	Stimulus clicking: RT-TA	Response mode	0.00	0.02
22	Stimulus highlighting: RA-RT	Response mode	−0.02	0.04
23	Stimulus highlighting: RA-TA	Response mode	0.01	0.03
24	Stimulus highlighting: RT-TA	Response mode	0.02	0.02

Each layer of the covariance structure corresponds to one classification. Classifications are made across three data types (RA, response accuracies that underlie the scored dichotomous responses; RTs, response times; TAs, times to first action taken) based on (correlated) latent factors, item domains, and item response modes.

about test-taker characteristics can be made conditional on a complex within-subject dependence structure that follows from combining various auxiliary process data types into a coherent model. There is no theoretical limitation to the number of data types to combine, or in the number of components within each type (e.g., test length).

Modeling local dependences through covariance parameters instead of random effect variance parameters results in an extended parameter space. This allows more flexibility in specifying complex dependence structures (covariances can be negative, zero or positive). Compared to default inverse-gamma priors for variance parameters, truncated shifted inverse-gamma priors for the covariance parameters are less informative and allow more objective inferences about the dependence structure. The truncation is furthermore used to ensure the positive definiteness of the additive covariance structure, and can be utilized for inequality hypothesis testing (e.g., $\theta_1 < \theta_2 < \theta_3$). Through conjugacy of the proposed priors, BCSMs can be fit with an efficient Gibbs-sampling algorithm.

In a simulation study, a complex within-subject dependence structure was successfully estimated under a BCSM for responses and RTs. The model used for data generation allowed the test-takers' working speed and ability to vary over the course of a test. The LHM was not capable to capture this variation and showed bias in the variance estimate of the ability distribution. Under the BCSM, variation in test-takers' working speed and ability did not violate the condition of local independence: the dependence structure was extended to account for the variation. Furthermore, by estimating the extended dependence structure, insight into the development of the speed-accuracy trade-off on group level across the test was obtained.

The empirical example based on the PIAAC study showed a complex real-world dependence structure in response and process data. Covariance, measurement error variance and item parameters were estimated conditional on a dependence structure that took into account the classifications across three data types (scored dichotomous responses, RTs, TAs), item characteristics (domain, response mode), and the latent factor structure (data components load on the correlated factors ability, working speed and speed first action). Furthermore, test-taker background variables were included as covariates to correct for between-subject differences in the dependence structure. Through additive layers in a single covariance matrix, 3-by-3 covariance structures were modeled for each specified subset of items. This allowed to evaluate the cross-dependence between all pairs of data types individually for each of the item domains and response modes. The results indicated, that the interdependences across data types and the investigated item characteristics were sufficiently captured by the covariance layers through which the dependences of the latent factors structure were specified. The empirical example illustrates how, in the BCSM framework, the modeled dependence structure can be flexibly adapted to the design and the underlying theoretical constructs of an assessment. Furthermore, the vague nature of the truncated shifted inverse-gamma prior specification promotes unbiased inferences about the dependence structure. In the empirical example, this was in particular important due to the vast number of covariance parameter estimates close to, or approximately equal to zero. In this situation, a prior specification that does not take boundary effects into account artificially increases the estimated probability of local independence and hence provokes false conclusions about the dependence structure and the underlying response processes.

In addition to integrating multiple types of test-taker data, dependences can follow from the test design, item properties, the (sub-)population of test-takers, test-taking modes, test-taking conditions, and from an interaction of these characteristics. Examples are testlet structures, in which data within a testlet is often more alike then data across testlets (e.g., Wainer and Kiely, 1987; Yen, 1993; Wainer et al., 2007), or the interaction of culturally loaded concepts in items and diverse (sub-)populations of test-takers (e.g., with and without migration background) (e.g., Steele and Aronson, 1995; Paniagua, 2000; Good et al., 2003; Robinson, 2010). The resulting dependences in test-taker data form a threat for the flawless psychometric equivalence of an assessment, if not accounted for Helms (1992).

In educational measurement, factor loadings, or slope parameters, are utilized to assess differential item functioning (DIF) across groups, test-taking modes and over time (Millsap, 2010), allow multidimensional item response theory (MIRT)

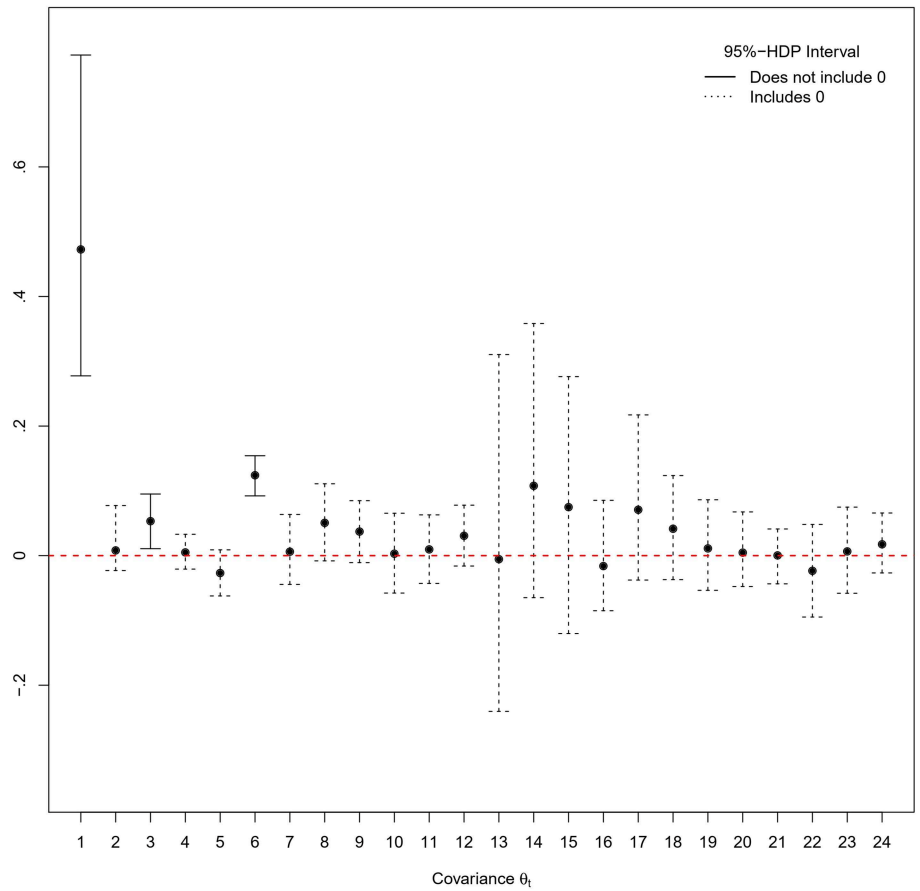

FIGURE 3 | 95%-Highest Posterior Density (HPD) intervals for the $N_t = 24$ covariance parameters in the additive covariance structure. Black dots correspond to posterior mean estimates.

(Reckase, 2009), and are used to represent the quality of an item to discriminate between distributions of test-takers with a different level of ability or speed (van der Linden, 2007; Klein Entink et al., 2008). As discussed by Klotzke and Fox (2018), factor loadings integrate seamlessly into the proposed modeling framework. In fact, the inclusion of factor loadings solely removes the restriction of values being either zero or one in the classification matrix, hence keeping the modeling structure and the therein derived equations intact. However, while this allows to include pre-calibrated factor loadings into the model, no estimation procedure has been described so far. In a conditional-BCSM hybrid model, the factor loadings can also be modeled in the mean structure instead of in the covariance structure. For example, a 2PL-IRT model with item-discrimination parameters can be specified in the mean structure and the dependences implied by a testlet structure can be explicitly modeled in the multivariate distribution of the error terms. This approach is straightforward and suited for practical applications. A downside is, that a trade-off is been made between the parsimony of the model and the number of person level variables included in the mean structure. In the empirical PIAAC data example showcased, the factor loadings were predefined given the test design and item characteristics. Freeing the factor loadings will further increase the flexibility in the modeled dependence structure and thereby the utility of BCSM for practical applications in educational measurement.

It has been shown that modeling a non-linear relationship between RAs and RTs can be beneficial (e.g., Molenaar et al., 2015; Bolsinova and Molenaar, 2018). Through the additive covariance structure in BCSM, the conditional dependence between RTs and RAs is not limited to vary solely based on item membership (i.e., data points that belong to the same item are conditionally more alike), but is allowed to change based on item characteristics (e.g., domain and response mode), test form (e.g., computer based vs. paper-and-pencil) and test design (e.g., a testlet structure). Individual test-taker characteristics that may cause between-subject differences in the dependences of RTs and RAs are controlled for through modeling test-taker background variables as covariates in the mean structure (e.g., the relationship between RTs and RAs may vary based on the test-takers' age or a pre-test speed categorization). This differs from methods that model a non-linear relationship between RTs and RAs through a predefined function that involves person-specific random components

and/or item parameters (Molenaar et al., 2015; Bolsinova et al., 2017; Bolsinova and Molenaar, 2018): in BCSM, test-taker characteristics that may affect the relationship between data types are controlled for in the mean structure, and the person-specific random effects are not modeled. Item characteristics are modeled in the mean structure (e.g., item difficulty parameters) and through additive layers in the covariance structure (e.g., item response mode). It is an interesting future prospect to see in how far the BCSM framework can be extended for covariance structures that follow from curvilinear functions for the relationship between data types. Furthermore, the BCSM approach must be distinguished from methods that model a person-specific covariance matrix (e.g., Meng et al., 2015): by their nature, models that explicitly specify a covariance matrix for each test-taker heavily increase in complexity with growing sample size and thus must impose strong restrictions on the modeled dependence structure to achieve model identification. In contrast, BCSM aims at designing parsimonious models that are easily identified when complex dependence structures are modeled.

The BCSM framework is not limited to RTs and dichotomous responses. Dependences between dichotomous responses and RTs were modeled through latent continuous variables. Expressions for the mean and variance of the conditional normal distribution of a latent variable were obtained by partitioning the additive covariance matrix and analytically deriving its inverse. Information from the observed responses (whether or not a test-taker responded correctly to an item) was utilized by truncating the respective distribution. Modeling dependences through latent continuous variables can be extended to data with more than two ordered or unordered response categories (e.g., Castro et al., 2012). This extends the range of process data that can be integrated into a BCSM. For example, sequential action patterns can be operationalized as count variables through N-grams (He et al., 2016). It is interesting to see under which conditions a BCSM allows to draw inferences about the interdependence between responses, RTs and action patterns and which new insights into latent response processes can be obtained. Further future prospects of BCSMs are the application to additional real-world empirical settings, extensions to unbalanced data and nested classifications on a person level (e.g., a test-taker is part of a school and classroom), and evaluating the utility of estimating test-taker effects *post-hoc*. Finally, it is of interest to compare the plausibility of different dependence structures in a Bayesian model selection framework (e.g., Kass and Raftery, 1995).

AUTHOR CONTRIBUTIONS

KK wrote the manuscript, developed the software and performed the analysis. J-PF and KK developed the modeling framework and the MCMC algorithm. J-PF contributed to the structure of the manuscript, and gave suggestions for the analysis and writing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01675/full#supplementary-material

REFERENCES

Ackerman, T. A. (1987). *The Robustness of LOGIST and BILOG IRT Estimation Programs to Violations of Local Independence (Research Report No. 87-14)*. Iowa City: The American College Testing Program.

Azevedo, R. (2015). Defining and measuring engagement and learning in science: conceptual, theoretical, methodological, and analytical issues. *Educ. Psychol.* 50, 84–94. doi: 10.1080/00461520.2015.1004069

Baker, F. B., and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques, 2nd Edn*. New York, NY: CRC Press.

Bolsinova, M., and Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Front. Psychol.* 9:1525. doi: 10.3389/fpsyg.2018.01525

Bolsinova, M., Tijmstra, J., and Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *Brit. J. Math. Stat. Psychol.* 70, 257–279. doi: 10.1111/bmsp.12076

Botev, Z. I. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *J. R. Stat. Soc. Ser. B* 79, 125–148. doi: 10.1111/rssb.12162

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika* 64, 153–168. doi: 10.1007/BF02294533

Bridgeman, B., and Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *J. Educ. Meas.* 41, 137–148. doi: 10.1111/j.1745-3984.2004.tb01111.x

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 1118–1133. doi: 10.1037/a0019902

Castro, M., Paleti, R., and Bhat, C. R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: application to pre- dicting crash frequency at intersections. *Transport. Res. B Methodol.* 46, 253–272. doi: 10.1016/j.trb.2011.09.007

Chen, W.-H., and Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *J. Educ. Behav. Stat.* 22, 265–289. doi: 10.3102/10769986022003265

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2013). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.* 32, 407–499. doi: 10.1214/009053604000000067

Fox, J.-P., Klotzke, K., and Klein Entink, R. H. (2018). *LNIRT: LogNormal Response Time Item Response Theory Models*. Available online at: https://CRAN.R-project.org/package=LNIRT

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

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. Boca Raton, FL: CRC Press.

Glas, C. A., and van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *Brit. J. Math. Stat. Psychol.* 63(Pt 3), 603–626. doi: 10.1348/000711009X481360

Goldhammer, F., and Kroehne, U. (2014). Controlling individuals' time spent on task in speeded performance measures: experimental time limits, posterior time limits, and response time modeling. *Appl. Psychol. Meas.* 38, 255–267. doi: 10.1177/0146621613517164

Goldhammer, F., Naumann, J., and Greiff, S. (2015). More is not always better: the relation between item response and item response time in Raven's matrices. *J. Intell.* 3, 21–40. doi: 10.3390/jintelligence3010021

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated

by task diffi- culty and skill: insights from a computer-based large-scale assessment. *J. Educ. Psychol.* 106, 608–626. doi: 10.1037/a0034716

Goldhammer, F., and Zehner, F. (2017). What to make Of and how to interpret process data. *Measurement* 15, 128–132. doi: 10.1080/15366367.2017.1411651

Good, C., Aronson, J., and Inzlicht, M. (2003). Improving adolescents' standardized test performance: an intervention to reduce the effects of stereotype threat. *J. Appl. Dev. Psychol.* 24, 645–662. doi: 10.1016/j.appdev.2003.09.002

Gupta, A. K., and Nagar, D. K. (1999). *Matrix Variate Distributions*. New York, NY: CRC Press.

He, Q., and Von Davier, M. (2016). "Analyzing process data from problem-solving items with n-grams: insights from a computer-based large-scale assessment," in *Handbook of Research on Technology Tools for Real-World Skill Development*, eds Y. Rosen, S. Ferrara, and M. Mosharraf (New York, NY: Chapman and Hall/CRC Press), 749–776. doi: 10.4018/978-1-4666-9441-5.ch029

Heidelberger, P., and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operat. Res.* 31, 1109–1144. doi: 10.1287/opre.31.6.1109

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *Am. Psychol.* 47, 1083–1101. doi: 10.1037//0003-066X.47.9.1083

Horrace, W. C. (2005). Some results on the multivariate truncated normal distribution. *J. Multivar. Anal.* 94, 209–221. doi: 10.1016/j.jmva.2004.10.007

Jiao, H., Kamata, A., Wang, S., and Jin, Y. (2012). A multilevel testlet model for dual local dependence. *J. Educ. Meas.* 49, 82–100. doi: 10.1111/j.1745-3984.2011.00161.x

Jiao, H., Wang, S., and Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *J. Appl. Meas.* 6, 311–321.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795. doi: 10.1080/01621459.1995.10476572

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2008). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74:21. doi: 10.1007/s11336-008-9075-y

Klotzke, K., and Fox, J.-P. (2018). *Response Times in a Bayesian Marginal Modeling Framework*. LSAC Research Report Series. Newport, PA: Law School Admission Council (USA).

Lange, K. (2010). *Numerical Analysis for Statisticians, 2nd Edn*. New York, NY: Springer Publishing Company, Incorporated.

Lee, Y., and Neider, J. A. (2004). Conditional and marginal models: another view. *Stat. Sci.* 19, 219–228. doi: 10.1214/088342304000000305

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22. doi: 10.1093/biomet/73.1.13

Loeys, T., Legrand, C., Schettino, A., and Pourtois, G. (2014). Semi-parametric proportional hazards models with crossed random effects for psychometric response times. *Brit. J. Math. Stat. Psychol.* 67, 304–327. doi: 10.1111/bmsp.12020

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press.

Maddox, B. (2017). Talk and gesture as process data. *Measurement* 15, 113–127. doi: 10.1080/15366367.2017.1392821

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., and Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *J. Educ. Behav. Stat.* 39, 426–451. doi: 10.3102/1076998614559412

McCulloch, C. E. (2003). "Generalized linear mixed models," in *NSF-CBMS Regional Conference Series in Probability and Statistics* (Beachwood, OH), Vol. 7, i-84.

Meijer, R. R., and Sotaridona, L. S. (2006). *Detection of Advance Item Knowledge Using Response Times in Computer Adaptive Testing (info:eu-repo/semantics/report No. CT 03-03)*. Newton, PA: Law School Admission Council.

Meng, X.-B., Tao, J., and Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *J. Educ. Meas.* 52, 1–27. doi: 10.1111/jedm.12060

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: an introduction. *Child Dev. Perspect.* 4, 5–9. doi: 10.1111/j.1750-8606.2009.00109.x

Molenaar, D., Oberski, D., Vermunt, J., and De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivar. Behav. Res.* 51, 606–626. doi: 10.1080/00273171.2016.1192983

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivar. Behav. Res.* 50, 56–74. doi: 10.1080/00273171.2014.962684

OECD (2013). *Technical Report of the Survey of Adult Skills (PIAAC) (Tech. Rep.)*. Paris: OECD Publishing.

Paniagua, F. A. (2000). *Handbook of Multicultural Mental Health: Assessment and Treatment of Diverse Populations*. San Diego, CA: Academic Press.

Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar, D., Bates, D., et al. (2016). *CODA: Output Analysis and Diagnostics for MCMC*. Available online at: https://CRAN.R-project.org/package=coda

Qian, H., Staniewska, D., Reckase, M., and Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educ. Meas.* 35, 38–47. doi: 10.1111/emip.12102

Ranger, J., and Kuhn, J.-T. (2013). Analyzing response times in tests with rank correlation approaches. *J. Educ. Behav. Stat.* 38, 61–80. doi: 10.3102/1076998611431086

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer-Verlag.

Robinson, J. P. (2010). The effects of test translation on young english learners' mathematics performance. *Educ. Res.* 39, 582–590. doi: 10.3102/0013189X10389811

Statistics Canada (2013). *Skills in Canada: First Results From the Programme for the International Assessment of Adult Competencies (PIAAC) (Monograph)*. Ottawa, ON: Statistics Canada.

Steele, C. M., and Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *J. Pers. Soc. Psychol.* 69, 797–811. doi: 10.1037/0022-3514.69.5.797

Thissen, D. (1983). "9 - Timed testing: an approach using item response theory," in *New Horizons in Testing*, ed D. J. Weiss (San Diego, CA: Academic Press), 179-203.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika* 67, 33–48. doi: 10.1007/BF02294708

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

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *J. Educ. Meas.* 46, 247–272. doi: 10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J., Breithaupt, K., Chuah, S. C., and Zhang, Y. (2007). Detecting differential speededness in multistage testing. *J. Educ. Meas.* 44, 117–130. doi: 10.1111/j.1745-3984.2007.00030.x

van der Linden, W. J., and Fox, J.-P. (2016). "Joint hierarchical modeling of responses and response times," in *Handbook of Item Response Theory, Volume One, Models*, ed W. J. van der Linden (New York, NY: Chapman and Hall/CRC Press), 481–500.

van der Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika* 73, 365–384. doi: 10.1007/s11336-007-9046-8

Wainer, H., Bradlow, E. T., and Wang, X. (2007). *Testlet Response Theory and Its Applications*. New York, NY: Cambridge University Press.

Wainer, H., and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: a case for testlets. *J. Educ. Meas.* 24, 185–201. doi: 10.1111/j.1745-3984.1987.tb00274.x

Wang, C., and Wilson, M. (2005). The rasch testlet model. *Appl. Psychol. Meas.* 29, 126–149. doi: 10.1177/0146621604271053

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Brit. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wetzel, E., Xu, X., and von Davier, M. (2015). An alternative way to model population ability distributions in large-scale educational surveys. *Educ. Psychol. Meas.* 75, 739–763. doi: 10.1177/00131644 14558843

Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl. Psychol. Meas.* 8, 125–145. doi: 10.1177/014662168 400800201

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *J. Educ. Meas.* 30, 187–213. doi: 10.1111/j.1745-3984.1993.tb00423.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.