



Longitudinal joint modeling for assessing parallel interactive development of latent ability and processing speed using responses and response times

Peida Zhan^{1,2,3} · Qipeng Chen¹ · Shiyu Wang⁴ · Xiao Zhang⁵

Accepted: 21 March 2023 / Published online: 14 April 2023
© The Psychonomic Society, Inc. 2023

Abstract

To measure the parallel interactive development of latent ability and processing speed using longitudinal item response accuracy (RA) and longitudinal response time (RT) data, we proposed three longitudinal joint modeling approaches from the structural equation modeling perspective, namely unstructured-covariance-matrix-based longitudinal joint modeling, latent growth curve-based longitudinal joint modeling, and autoregressive cross-lagged longitudinal joint modeling. The proposed modeling approaches can not only provide the developmental trajectories of latent ability and processing speed individually, but also exploit the relationship between the change in latent ability and processing speed through the across-time relationships of these two constructs. The results of two empirical studies indicate that (1) all three models are practically applicable and have highly consistent conclusions in terms of the changes in ability and speed in the analysis of the same data set, and (2) additional analysis of the RT data and acquisition of individual processing speed measurements can reveal the parallel interactive development phenomena that are difficult to detect using RA data alone. Furthermore, the results of our simulation study demonstrate that the proposed Bayesian Markov chain Monte Carlo estimation algorithm can ensure accurate model parameter recovery for all three proposed longitudinal joint models. Finally, the implications of our findings are discussed from the research and practice perspectives.

Keywords Longitudinal joint modeling · Response times · Latent growth model · Autoregressive cross-lagged model · Item response theory · Longitudinal data analysis

In psychological and behavioral science, researchers are often interested in studying the developmental changes of a group or of multiple groups of individuals, such as changes in their cognitive levels and behavioral patterns over time. Longitudinal studies are often conducted to investigate these problems, and results

from these studies can yield convincing arguments pertaining to the relationships between variables (e.g., directionality of causality) by constructing a theoretical or temporal back-and-forth logic between said variables (Ferrer & McArdle, 2010; Leszczensky & Wolbring, 2022; Toh & Hernán, 2008). The measurement of developmental changes relies on longitudinal data collected using multiple measures of constructs over time. Typically, these changes can be captured with longitudinal latent variable models falling into two main categories (McArdle, 2009; Muthén & Muthén, 2000): (1) longitudinal models that focus on changes in categorical latent variables, such as hidden Markov models (or latent transition models) (e.g., Bartolucci et al., 2013; Collins et al., 1997; Wang, Yang, et al., 2018a), and (2) longitudinal models that focus on metrical changes in continuous latent variables, such as the latent growth curve models (e.g., Bollen & Curran, 2006; Duncan et al., 2006) and longitudinal item response theory (IRT) models (e.g., Andersen, 1985; von Davier et al., 2011; Wang & Nydick, 2020). Most of these longitudinal models use item response accuracy (RA) data, such as binary responses indicating whether an answer to a multiple-choice question is right or wrong or ordinal responses

✉ Peida Zhan
pdzhan@gmail.com

¹ School of Psychology, Zhejiang Normal University, Jinhua, China

² Intelligent Laboratory of Child and Adolescent Mental Health and Crisis Intervention of Zhejiang Province, Zhejiang Normal University, Jinhua 321004, China

³ Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua 321004, China

⁴ Department of Educational Psychology, University Georgia, Athens, GA, USA

⁵ Faculty of Education, The University of Hong Kong, Hong Kong, China

to Likert-type items, as observed indicators of the measured latent construct in psychological studies.

Thanks to advancements in computer- and web-based learning, assessment, and experimental systems, many types of multimodal data, such as reaction or response times (RTs), mouse clicks, action sequences, eye tracking, and brain activation, are accessible in addition to RA data (Gorin, 2006; Jeon et al., 2021; Jiao & Lissitz, 2018; van der Maas & Jansen, 2003; Zhan et al., 2022). Among these types of data, participants' RTs to test item, which is the amount of time spent by an individual to consider and solve each task or item, have been used to explore behavioral patterns in many psychological and behavioral case studies (e.g., Meijering & van Rijn, 2009; Siegler, 1989; van der Maas & Jansen, 2003) and to address the measurement issues encountered in many methodological studies (e.g., Bolsinova & Tjijmstra, 2018; Man & Harring, 2021; van der Linden & Guo, 2008; Zhan et al., 2018). For an assessment at a given time point, RA and RT data are collected simultaneously to provide parallel information pertaining to participants' cognitive processes or behavioral patterns for the same task (e.g., it takes 60 s for a participant to respond to an item correctly). As a complement to RA data, RT data may offer additional information related to the contents of the cognitive processes that underlie problem-solving on tasks that are difficult to analyze using RA data alone (De Boeck & Jeon, 2019). A joint analysis of RA and RT data can not only directly reveal the relationship between latent ability and processing speed, but also improve the accuracy of parameter estimation with the measurement models (Bolsinova & Tjijmstra, 2018; van der Linden, 2007; Zhan et al., 2018).

Several recent computer-based longitudinal studies, such as dynamic studies focusing on learning development (e.g., adaptive learning [Wang, Yang, et al., 2018a] and intelligent tutoring [Woolf, 2009]) and longitudinal behavioral experimental studies focusing on child development (e.g., Ouyang et al., 2022) have facilitated the simultaneous collection of RA and RT data across multiple time points, and we refer to these data as longitudinal RA and longitudinal RT. Moreover, these studies have demonstrated the promise of using both longitudinal RA and longitudinal RT data to study changes in cognitive levels and behavioral patterns. Human behavior or cognitive process is a product of multiple constructs that are systematically related to one another. Hence, individual constructs are likely to undergo developmental change not independently but on the basis of the influences of their respective developmental changes on each other. Such change in constructs that develop separately but influence each other is called "*parallel interactive development*" in the present study. Our focus is on the parallel interactive development of two latent constructs, namely latent ability and processing speed, which are measured using RA and RT, respectively, with the aim of advancing our holistic understanding of the developmental changes in individuals.

With the availability of longitudinal RA and longitudinal RT data, a statistical method must be used to capture the parallel interactive development of latent ability and processing speed from the data. Joint-hierarchical latent variable modeling is one of the most popular approaches for simultaneously analyzing RA and RT data (van der Linden, 2007). However, this modeling approach is suitable for cross-sectional data, and it assumes a constant continuous ability (reflected by RA) and a constant continuous latent speed (reflected by RT) through a test at a given point in time. Many variants of this model have been proposed to generalize its assumptions to diverse test scenarios and applications, which are nonetheless limited to cross-sectional data (e.g., Fox & Mariani, 2016; Klein Entink et al., 2009; Man et al., 2019; Man & Harring, 2021; Molenaar et al., 2015, 2016; Zhan et al., 2021). Although a few studies have extended this joint modeling framework to track changes in speed and ability, they have defined latent ability as a categorical latent variable and used a cognitive diagnostic model (CDM) (e.g., Junker & Sijtsma, 2001) for measuring RA data (Wang et al., 2020; Wang, Zhang, et al., 2018b). By contrast, our study focuses on detecting the parallel interactive development of two continuous latent variables, namely latent ability and latent processing speed. Continuous latent variables and categorical latent variables represent two different quantitative perspectives on latent constructs, and they are used for different purposes. Many latent constructs, such as intelligence, are considered continuous, and in such a case, we seek a scale on which individuals can be located. In other words, continuous variables portray individual development more finely than categorical variables, especially when the magnitude of change is small (e.g., only quantitative but not qualitative change) (e.g., Zhan, 2021).

The objective of our study is to develop novel statistical modeling approaches for assessing the parallel interactive development of continuous latent ability and processing speed using both longitudinal RA and longitudinal RT data. The proposed longitudinal joint modeling approaches are within the domain of structural equation modeling (SEM) (McArdle & Nesselroade, 2014), and we adapt three widely used modeling approaches in our setup: unstructured-covariance-matrix-based modeling, latent (parallel process) growth curve modeling, and autoregressive cross-lagged modeling. The first modeling approach is often used in longitudinal IRT models, and the latter two are frequently used for analyzing longitudinal data in the field of developmental and applied psychology. In addition, the proposed modeling approaches can not only provide developmental trajectories in terms of changes in latent ability and processing speed over time, but also provide additional information about the across-time relationships between these two constructs. In this sense, these approaches can be viewed as extensions of joint-hierarchical latent variable modeling (van der Linden, 2007) in a longitudinal setup from the SEM perspective.

In the remaining sections, we first describe the proposed longitudinal joint models, including the model formulas, model estimation procedure, and differences from a few related models. Then, two empirical examples are presented to illustrate the applicability of the proposed models, followed by a simulation study to explore the psychometric performance of the proposed models under different simulated test conditions. Finally, we present our findings and provide an outline for future research on this topic.

Longitudinal joint-hierarchical latent variable modeling

Consider N students participating in a longitudinal assessment that integrates learning components across P time points. Let I_p be the index of the number of items administered at time point p ($p = 1, \dots, P$) and the entire assessment consists of $I = \sum_{p=1}^P I_p$ items. Two types of longitudinal data can be collected from these items at each time point: longitudinal RA and longitudinal RT. Let Y_{nip} and T_{nip} be the n ($n = 1, \dots, N$)-th student's RA and RT for item i ($i = 1, \dots, I_p$) at time point p , respectively. Following the SEM framework, the proposed longitudinal joint models consist of two components: a *measurement model* that describes the associations between the observed variables and latent variables at a given time point and a *structural model* that describes the changes in and relationships among the latent variables over time. Our focus is on presenting three types of structural models that can be used to describe the mechanisms of change in latent ability and processing speed over time. The following sections describe the measurement model component and the proposed structural models.

Measurement model

At a given time point p , we assume that RA data follow the two-parameter logistic IRT model (Birnbaum, 1968), which can be expressed as follows:

$$\text{Prob}(Y_{nip} = 1 | \theta_{np}, a_{ip}, b_{ip}) = \frac{\exp(a_{ip}\theta_{np} - b_{ip})}{1 + \exp(a_{ip}\theta_{np} - b_{ip})}, \quad (1)$$

where θ_{np} denotes the latent ability of student n at time point p , and a_{ip} and b_{ip} denote the discrimination and difficulty of item i at time point p , respectively. Equation (1) represents a two-parameter extension of Andersen's (1985) longitudinal Rasch model with an additional item-discrimination parameter (see also von Davier et al., 2011).

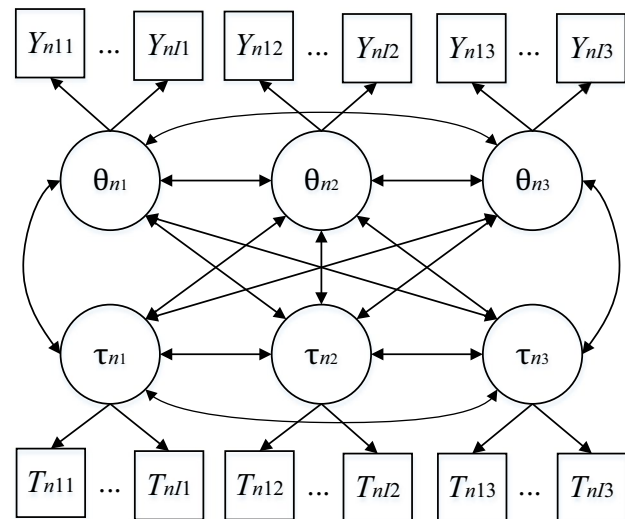


Fig. 1 Graphical representation of unstructured-covariance-matrix-based longitudinal joint models (three time points)

In addition, at a given time point p , we assume that RT data follow the lognormal RT model (van der Linden, 2006) with an additional time-discrimination parameter (Klein Entink et al., 2009), which can be expressed as follows:

$$\log T_{nip} = \xi_{ip} - \phi_{ip}\tau_{np} + \varepsilon_{nip}, \varepsilon_{nip} \sim N(0, \omega_{ip}^{-2}), \quad (2)$$

where τ_{np} denotes the latent processing speed of student n at time point p , and ξ_{ip} , ϕ_{ip} , and ω_{ip} denote the time-intensity, time-discrimination, and time-precision of item i at time point p , respectively.

Note that the dichotomous RA data are considered only as an example to illustrate the conceptualization of the proposed modeling approaches. Different IRT models can be applied to ordinal or nominal response data with more than two categories (Bock, 1972; Samejima, 1969). A few studies have explored the joint analysis of ordinal response and RT data in cross-sectional personality tests based on rating scales (e.g., Ranger, 2013).

Structural model

Unstructured-covariance-matrix-based structural model

To describe the relationship between θ_{np} and τ_{np} across P time points, one of the most straightforward methods is to construct an unstructured covariance matrix (e.g., von Davier et al., 2011; Zhan et al., 2019), as illustrated in Fig. 1. The structural model component of the proposed joint model based on an unstructured covariance matrix, denoted by COV, is as follows:

$$\begin{pmatrix} \theta_n \\ \tau_n \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \Sigma\right), \Sigma = \begin{pmatrix} \sigma_{\theta_1}^2 & \cdots & \sigma_{\theta_1\theta_p} & \sigma_{\theta_1\tau_1} & \cdots & \sigma_{\theta_1\tau_p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{\theta_p\theta_1} & \cdots & \sigma_{\theta_p}^2 & \sigma_{\theta_p\tau_1} & \cdots & \sigma_{\theta_p\tau_p} \\ \sigma_{\tau_1\theta_1} & \cdots & \sigma_{\tau_1\theta_p} & \sigma_{\tau_1}^2 & \cdots & \sigma_{\tau_1\tau_p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{\tau_p\theta_1} & \cdots & \sigma_{\tau_p\theta_p} & \sigma_{\tau_p\tau_1} & \cdots & \sigma_{\tau_p}^2 \end{pmatrix}, \tag{3}$$

where $\theta_n = (\theta_{n1}, \dots, \theta_{nP})'$ denotes the latent ability vector consisting of P elements, $\tau_n = (\tau_{n1}, \dots, \tau_{nP})'$ denotes the latent processing speed vector consisting of P elements. The vectors $\mu_\theta = (\mu_{\theta_1}, \dots, \mu_{\theta P})'$ and $\mu_\tau = (\mu_{\tau_1}, \dots, \mu_{\tau P})'$ are the population mean vector of latent abilities and population mean vector of latent processing speeds, respectively. Σ is a variance-covariance matrix, in which $\sigma_{\theta_p}^2$ is the variance of θ_p , $\sigma_{\tau_p}^2$ is the variance of τ_p , $\sigma_{\theta_p\theta_p'}$ is the covariance of θ_p and θ_p , $\sigma_{\tau_p\tau_p'}$ is the covariance of τ_p and τ_p , and $\sigma_{\theta_p\tau_p}$ is the covariance of θ_p and τ_p .

This structural model directly outputs estimates of the latent ability and processing speed at each time point. Therefore, $\hat{\theta}_n$ and $\hat{\tau}_n$, respectively, can be used to directly describe the estimated developmental trajectories of the latent ability and processing speeds of individuals. In other words, $\hat{\theta}_{n(p+1)} - \hat{\theta}_{np}$ and $\hat{\tau}_{n(p+1)} - \hat{\tau}_{np}$ can be used to describe the estimated degrees of changes in individuals' latent abilities and processing speeds. The estimate of the population-level average changes in ability and speed at two adjacent time points can then be denoted by $\hat{\mu}_{\theta(p+1)} - \hat{\mu}_{\theta p}$ and $\hat{\mu}_{\tau(p+1)} - \hat{\mu}_{\tau p}$, respectively. Furthermore, the estimate of the population-level scale changes in ability and speed at two adjacent time points can be described by $\hat{\sigma}_{\theta(p+1)}/\hat{\sigma}_{\theta p}$ and $\hat{\sigma}_{\tau(p+1)}/\hat{\sigma}_{\tau p}$, respectively (Paek et al., 2014).

The unstructured covariance matrix in COV makes it possible to consider various relationships between the latent constructs across all time points. However, when the number of time points is large, the computational cost increases dramatically, and the nonconvergent estimation issue may be encountered.

Latent growth curve-based structural model

The second structural model is based on the latent growth curve (e.g., Curtis, 2010; Wang & Nydick, 2020), and we denote the joint model based on it as the latent growth curve longitudinal joint model (LGC). This model is illustrated in Fig. 2 and is expressed as follows:

$$\theta_{np} = \pi_{n0} + \pi_{n1}(p - 1) + \varepsilon_{\theta_{np}}, \varepsilon_{\theta_{np}} \sim N(0, \sigma_{\theta_p}^2), \tag{4}$$

$$\tau_{np} = \delta_{n0} + \delta_{n1}(p - 1) + \varepsilon_{\tau_{np}}, \varepsilon_{\tau_{np}} \sim N(0, \sigma_{\tau_p}^2), \tag{5}$$

$$\begin{pmatrix} \pi_{n0} \\ \pi_{n1} \\ \delta_{n0} \\ \delta_{n1} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_{\pi_0} \\ \mu_{\pi_1} \\ \mu_{\delta_0} \\ \mu_{\delta_1} \end{pmatrix}, \begin{pmatrix} \sigma_{\pi_0}^2 & \cdots & \sigma_{\pi_0\delta_1} \\ \vdots & \ddots & \vdots \\ \sigma_{\delta_1\pi_0} & \cdots & \sigma_{\delta_1}^2 \end{pmatrix}\right), \tag{6}$$

where π_{n0} and π_{n1} are the individual growth intercept and growth slope parameters for latent ability, respectively, and δ_{n0} and δ_{n1} are the individual growth intercept and growth slope parameters for processing speed, respectively. In addition, μ_{π_0} and μ_{π_1} denote the population mean and population average developmental change in latent ability, respectively, and μ_{δ_0} and μ_{δ_1} denote the population mean and population average developmental change in processing speed, respectively. Furthermore, $\varepsilon_{\theta_{np}}$ and $\varepsilon_{\tau_{np}}$ are the residual terms of latent ability and processing speed, respectively.¹

Unlike the COV, which directly estimates the values of θ_{np} and τ_{np} at each time point, this model estimates the growth factors (i.e., π_{n0} , π_{n1} , δ_{n0} , and δ_{n1}) of each individual. Accordingly, $\hat{\pi}_{n1}$ and $\hat{\delta}_{n1}$ can be used to describe the estimated amounts of change in latent ability and processing speed, respectively, of each individual between adjacent time points, and $\hat{\mu}_{\pi_1}$ and $\hat{\mu}_{\delta_1}$ can be used to describe the estimated amounts of change in latent ability and processing speed, respectively, of the corresponding population means between adjacent time points.

As expressed in Eq. (6), growth factors are assumed to follow a multivariate normal distribution, indicating that the starting values (i.e., π_{n0} and δ_{n0}) of and the amounts of developmental change (i.e., π_{n1} and δ_{n1}) in the latent constructs mutually influence each other. In such cases, for example, $\rho_{\pi_0\delta_0} = \sigma_{\delta_0\pi_0} / \sigma_{\delta_0}\sigma_{\pi_0}$ can be used to describe the correlation between latent ability and processing speed at the starting point, and $\rho_{\pi_0\pi_1} = \sigma_{\pi_0\pi_1} / \sigma_{\pi_0}\sigma_{\pi_1}$ and $\rho_{\delta_0\delta_1} = \sigma_{\delta_0\delta_1} / \sigma_{\delta_0}\sigma_{\delta_1}$ can be used to describe the correlation between the starting values of and amounts of developmental change in two latent constructs, respectively. If these values are significantly greater

¹ Notably, a few studies do not include the residual term in the latent growth curve model or fix $\sigma_{\theta_p}^2 = \sigma_{\tau_p}^2 = 0$ (e.g., Curtis, 2010; Fox & Mariani, 2016), resulting in perfectly linear growth of the individuals' latent constructs over time.

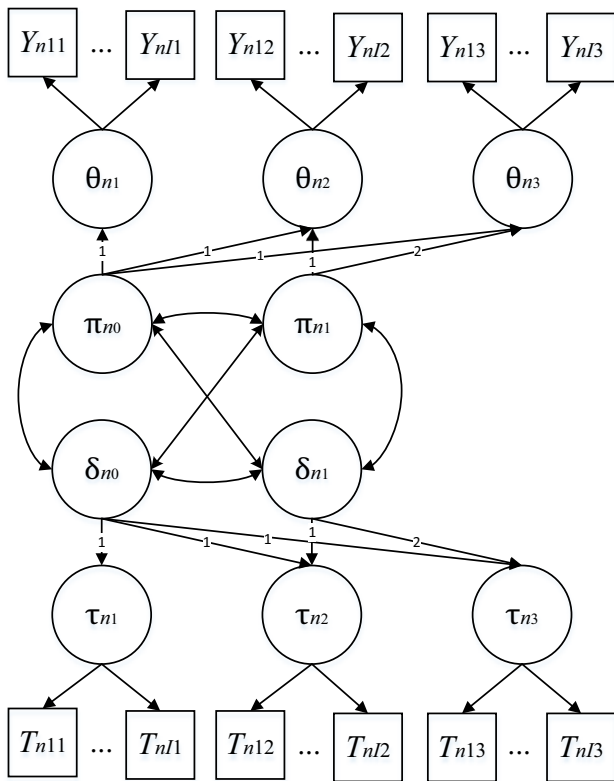


Fig. 2 Graphical representation of latent growth curve longitudinal joint models (three time points)

than 0, it follows that the higher the starting level, the faster is the individual’s development, and vice versa.

The proposed LGC is a linear growth model. A more complex nonlinear growth process can be considered by adding quadratic terms (e.g., $\pi_{n2}(p - 1)^2$ and $\delta_{n2}(p - 1)^2$, where π_{n2} and δ_{n2} are the quadratic growth parameters of latent ability and processing speed, respectively) to Eqs. (4) and (5), respectively. This nonlinear LGC was also fitted to two data sets presented in the empirical examples section. However, its relative model–data fit, including the model complexity penalty, was worse than that of the linear LGC in empirical example 1 and marginally better than that of the linear LGC in empirical example 2 (Tables S8 and S9 in the online appendix present the relative and absolute model–data fits, respectively). For this reason, we choose to not present this model in this study.

Autoregressive cross-lagged structural model

The last approach we consider for describing the developmental changes in θ_{np} and τ_{np} over time is autoregressive

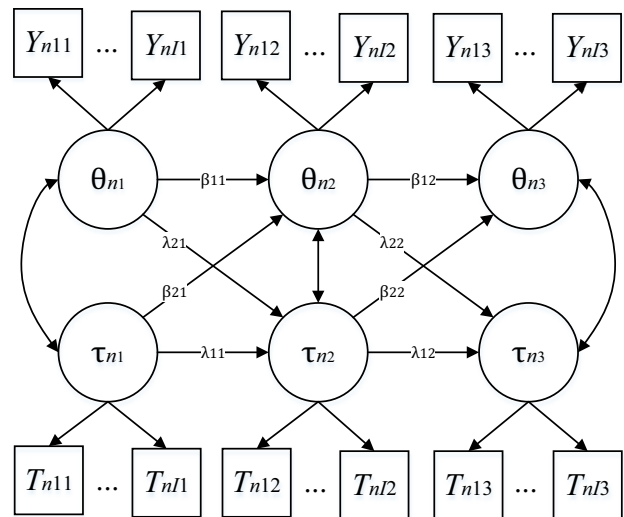


Fig. 3 Graphical representation of autoregressive cross-lagged longitudinal joint model (three time points)

cross-lagged modeling (e.g., Bentler, 1980; Mayer, 1986; McArdle, 2009), as depicted in Fig. 3. We express the structural model of the autoregressive cross-lagged longitudinal joint model (denoted as ACL) as follows:

$$\theta_{np} = \beta_{0p} + \beta_{1p}\theta_{n(p-1)} + \beta_{2p}\tau_{n(p-1)} + \varepsilon_{\theta_{np}}, \tag{7}$$

$$\tau_{np} = \lambda_{0p} + \lambda_{1p}\tau_{n(p-1)} + \lambda_{2p}\theta_{n(p-1)} + \varepsilon_{\tau_{np}}, \tag{8}$$

$$\begin{pmatrix} \varepsilon_{\theta_{np}} \\ \varepsilon_{\tau_{np}} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\varepsilon_{\theta_p}}^2 & \sigma_{\varepsilon_{\theta_p}\varepsilon_{\tau_p}} \\ \sigma_{\varepsilon_{\tau_p}\varepsilon_{\theta_p}} & \sigma_{\varepsilon_{\tau_p}}^2 \end{pmatrix}\right), \tag{9}$$

where β_{0p} and λ_{0p} are the intercepts of latent ability and processing speed at time point p , respectively. Moreover, β_{1p} and λ_{1p} denote the autoregressive effects of latent ability and processing speed at time point p , respectively, and they describe the stability of the latent constructs from one time point to the next. β_{2p} and λ_{2p} denote the cross-lagged effects of latent ability and processing speed at time point p , respectively, and they describe the effect of one construct on another measured at a later time point. In addition, $\varepsilon_{\theta_{np}}$ and $\varepsilon_{\tau_{np}}$ are the residual terms of latent ability and processing speed, respectively. In contrast to the residual terms in the LGC, the residual terms in the ACL are assumed to follow a bivariate normal distribution for describing the correlation between the unexplained parts of latent ability and processing speed.

Compared to the COV (Fig. 1), the ACL strictly limits the directionality of the influence between variables (i.e.,

Table 1 Comparison of individual level and population mean level among the three models at three time points

Model	Construct	Individual			Population mean		
		$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
COV	Ability	θ_{n1}	θ_{n2}	θ_{n3}	$\mu_{\theta 1}$	$\mu_{\theta 2}$	$\mu_{\theta 3}$
	Speed	τ_{n1}	τ_{n2}	τ_{n3}	$\mu_{\tau 1}$	$\mu_{\tau 2}$	$\mu_{\tau 3}$
LGC	Ability	$\pi_{0n} + \varepsilon_{\theta np}$	$\pi_{0n} + \pi_{1n} + \varepsilon_{\theta np}$	$\pi_{0n} + 2\pi_{1n} + \varepsilon_{\theta np}$	$\mu_{\pi 0}$	$\mu_{\pi 0} + \mu_{\pi 1}$	$\mu_{\pi 0} + 2\mu_{\pi 1}$
	Speed	$\delta_{0n} + \varepsilon_{\tau np}$	$\delta_{0n} + \delta_{1n} + \varepsilon_{\tau np}$	$\delta_{0n} + 2\delta_{1n} + \varepsilon_{\tau np}$	$\mu_{\delta 0}$	$\mu_{\delta 0} + \mu_{\delta 1}$	$\mu_{\delta 0} + 2\mu_{\delta 1}$
ACL	Ability	$\theta_{n1} = \beta_{01} + \varepsilon_{\theta np}$	$\theta_{n2} = \beta_{02} + \beta_{12}\theta_{n1} + \beta_{22}\tau_{n1} + \varepsilon_{\theta np}$	$\beta_{03} + \beta_{13}\theta_{n2} + \beta_{23}\tau_{n2} + \varepsilon_{\theta np}$	$\bar{\theta}_1 = \beta_{01}$	$\bar{\theta}_2 = \beta_{02} + \beta_{12}\beta_{01} + \beta_{22}\lambda_{01}$	$\beta_{03} + \beta_{13}\bar{\theta}_2 + \beta_{23}\bar{\tau}_2$
	Speed	$\tau_{n1} = \lambda_{01} + \varepsilon_{\tau np}$	$\tau_{n2} = \lambda_{02} + \lambda_{12}\tau_{n1} + \lambda_{22}\theta_{n1} + \varepsilon_{\tau np}$	$\lambda_{03} + \lambda_{13}\tau_{n2} + \lambda_{23}\theta_{n2} + \varepsilon_{\tau np}$	$\bar{\tau}_1 = \lambda_{01}$	$\bar{\tau}_2 = \lambda_{02} + \lambda_{12}\lambda_{01} + \lambda_{22}\beta_{01}$	$\lambda_{03} + \lambda_{13}\bar{\tau}_2 + \lambda_{23}\bar{\theta}_2$

COV unstructured covariance matrix-based longitudinal joint model, LGC latent growth curve longitudinal joint model, ACL autoregressive cross-lagged longitudinal joint model, p time point, θ latent ability, τ latent speed

prediction). Under strict experimental design, the ACL can be used to reveal causal relationships between variables if irrelevant variables are controlled (Bentler, 1980), while the COV cannot. In addition, similar to the COV, $\hat{\theta}_{n(p+1)} - \hat{\theta}_{np}$ and $\hat{\tau}_{n(p+1)} - \hat{\tau}_{np}$ can be used to estimate the degrees of change in individuals’ latent ability and processing speed, respectively. However, the calculations of the degrees of change in the population means of latent ability and processing speed at adjacent time points are complex. For ease of understanding, an example is summarized in Table 1.

Summary and comparison between three longitudinal joint models

In summary, we presented three types of joint models for longitudinal RA and RT data with differences in the formulations of the structural models, denoted by COV, LGC, and ACL. Table 1 presents a comparison between individual level and population mean level among the three structural models at three time points. First, at the starting point (i.e., $p = 1$), the three models are equivalent, though different notations for parameters related to ability and speed are used given their own model formulations. Second, the three models have different assumptions for describing the developmental changes in ability and speed. Specifically, the COV has the most lenient assumption, and it directly estimates the latent ability and processing speed at each time point. The LGC assumes a linear growth and estimates the coefficients of the growth curves (i.e., growth factors) of ability and speed over time. The ACL further models the influence of constructs on themselves and the relationships between constructs at two adjacent time points. Third, both the LGC and the ACL can be treated as special cases of the COV by reparameterizing the mean and the variance of the distributions of the latent variables in the COV. For example, in the LGC,

$$\theta_{np} \sim N\left(\mu_{\pi_0} + \mu_{\pi_1}(p - 1), \sigma_{\pi_0}^2 + \sigma_{\pi_1}^2(p - 1)^2 + 2(p - 1)Cov(\pi_0, \pi_1) + \sigma_{\theta_p}^2\right)$$

addition, to our understanding, ACL and LGC develop independently, and there is no theoretical nested relationship between them. Fourth, when $P \geq 3$, all three models can be used. When $P = 2$, the COV and the ACL are recommended because two time points do not satisfy the identifiability requirement of LGC (Bollen & Curran, 2006).² When $P = 1$, all three models are reduced to the joint-hierarchical item response model for cross-sectional data (Klein Entink et al., 2009; van der Linden, 2007).

Related models

A limited number of models have been proposed for analyzing longitudinal RA and RT data simultaneously. One area of research is within the ambit of the dynamic CDM framework (Chen & Culpepper, 2020; Wang et al., 2020; Wang, Zhang, et al., 2018b). The aforementioned models use the CDM model and the lognormal RT model for measuring RA and RTs, respectively, at a given time point and then model the transition of a discrete latent ability variable based on different assumptions of the latent ability variable and latent speed. The proposed models differ from these joint models mainly in terms of modeling the change in a continuous latent ability instead of the discrete latent ability. In addition, our proposed models allow different possibilities of changes in both abilities and speed, which are more realistic in terms of the application in diverse longitudinal assessment scenarios.

Another branch of research that is closely related to this study is joint modeling of RA and RT data in cross-sectional scenarios. For example, Zhan et al. (2021) proposed

² The LGC can be identified at two time points by imposing some constraints. For example, for latent ability, by constraining $Cov(\pi_0, \pi_1) = 0$ and $\sigma_{\theta_1}^2 = \sigma_{\theta_2}^2$, we get $\mu_{\pi_0} = E(Y_1)$, $\mu_{\pi_1} = E(Y_2) - E(Y_1)$, $\sigma_{\pi_0}^2 = Cov(Y_1, Y_2)$, and $\sigma_{\pi_1}^2 = Var(Y_2) - Var(Y_1)$.

a multidimensional joint model for RA and RT data, which can be used to analyze the multidimensionality of ability and speed in cross-sectional assessments. In fact, the COV proposed in this study can be considered an application of the multidimensional joint model to the analysis of longitudinal bimodal data (RA and RTs), similar to the application of multidimensional IRT models to the analysis of longitudinal RA data (von Davier et al., 2011). In addition, to address the within-person relationship between ability and speed in cross-sectional assessments, Fox and Marianti (2016) proposed a joint model of RA and RT and modeled the relationship between latent ability and speed using a latent growth curve model. In contrast to their assumption that an individual has constant ability but variable speed across items, our LGC assumes that both ability and speed change over time but remain constant across items at a specific time point.

Model estimation

Model identification

To ensure the comparability of latent constructs across time points, an anchor-item design or a repeated measurement design (i.e., all items are anchor items) is often used. Kolen and Brennan (2004) recommended that assessments use at least 20% of the items in question to anchor the parameters to the common scale. If an adequate number of items is linked across time and no item parameter drift is assumed, the estimation of model parameters requires the imposition of a few model identifiability constraints. We impose constraints on the mean and variance of ability and the speed to identify the scales of ability and speed. In the proposed models, the mean of latent ability at the first time point is set to zero to identify the mean of the ability scale. Similarly, the mean of processing speed at the first time point is set to zero to identify the mean of the speed scale. Specifically, in the COV, we set $\mu_{01} = \mu_{\tau 1} = 0$; in the LGC, we set $\mu_{\pi_0} = \mu_{\delta_0} = 0$; and in the ACL, we set $\beta_{01} = \lambda_{01} = 0$. Next, in all the proposed models, the product of item discrimination parameters is restricted to 1 to identify the variance of the ability scale, and the product of item time-discrimination parameters is restricted to 1 to identify the variance of the speed scale. Once the ability and speed scales are identified, all higher-level model parameters can be identified.

Bayesian parameter estimation

The parameters of the proposed models can be estimated using the Markov chain Monte Carlo (MCMC) method within the Bayesian estimation framework. In this study, the PyMC3 package (version 3.11.2) (Salvatier et al., 2016; <https://docs.pymc.io>) in the Python software environment

is used to implement the MCMC method. Moreover, the No-U-Turn Sampler (NUTS) (Hoffman & Gelman, 2014) in PyMC3 is used as the MCMC sampling algorithm because it uses gradient information from the likelihood to converge considerably faster than traditional sampling algorithms (e.g., Gibbs sampling), especially in the case of complexity models.

Herein, medium-informative hyper-priors are used to increase the generalizability of our code. The priors used in this study are listed in Section S1.1 of the online appendix. In addition, a robustness analysis of the model parameter estimation for low-, medium-, and high-informative hyper-priors is presented in Section S1.2 of the online appendix. The results indicate that the three proposed models are highly robust against prior distributions containing different amounts of information.

In longitudinal studies, missing data are encountered commonly owing to factors such as sample attrition and test design (i.e., missing by design). A common practice in Bayesian estimation is to have the algorithm automatically fill in missing values on the basis of the sampled model parameter values (e.g., Pan & Zhan, 2020). However, when the proportion of missing values is high, the computational cost increases considerably. In addition, when the model does not fit the data well, the automatically filled data may have large biases. By contrast, if the data are missing completely at random (MCAR) or missing at random (MAR), we suggest using only the observed data to fit the proposed models.³ For the MCAR or MAR, Rubin (1976) postulated that respondents were similar to nonrespondents within a subcategory; methods such as full information maximum likelihood estimation, which ignores missing values and calculates the likelihood function using observed data, will work reasonably well in such cases (Pokropek, 2011). To delete the missing data that are devoid of information at the response level (i.e., missing data of specific persons for specific items), we adjust the data format from the traditional item response matrix to the item response vector with tags that include time points, persons, and items. Further details pertaining to the aforementioned data format conversion can be found in Section S2 in the online appendix.

³ In practice, testing the missing data mechanism is not an easy task. Fortunately, the mechanisms of missing data due to some specific factors are easier to determine. For example, participants in longitudinal studies are frequently requested to complete only a portion of a test in order to lessen their cognitive burden; the resulting missing data by design are usually identified as MCAR (Peugh & Enders, 2004; Pokropek, 2011). In addition, Little's test of MCAR (Little, 1988) is often used to test for missing data mechanism.

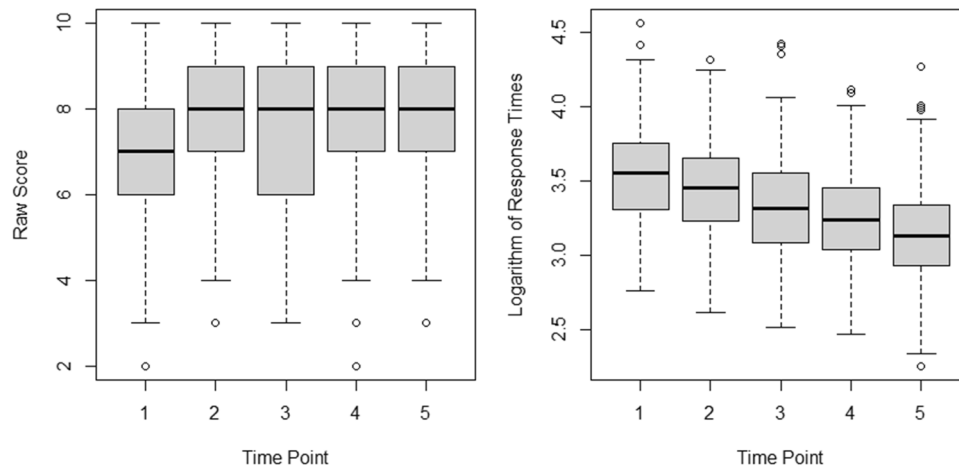


Fig. 4 Box plot of students' total raw scores and average logarithm of response times at each time point in empirical example 1

Applications

The proposed models were applied for analyzing two sets of data. The first set was collected from a computer-based learning platform that aims to improve students' spatial rotation skills. The experiment was conducted over 1 to 2 hours with short intervals between time points. The second data set was collected from a computer-based behavioral experiment in the field of developmental psychology, and it was characterized by long intervals between time points. In both experiments, the participants' longitudinal RA and RT data were recorded. The three proposed joint models were fitted to both data sets. In addition, to evaluate the benefit of using a joint modeling framework, we also separately fitted the longitudinal RA and longitudinal RT data by their corresponding measurement model (denoted as sep-COV, sep-LGC, and sep-ACL, respectively). That is, ignoring the structural relationship between latent ability and processing speed.

Empirical example 1: Computer-based spatial rotation learning experiment

Data description

The spatial rotation data set employed herein has been used in a few previous studies to track learning trajectories in dynamic CDM frameworks (e.g., Wang et al., 2020; Wang, Yang, et al., 2018a; Wang, Zhang, et al., 2018b). In the experiment, a total of 350 participants answered 50 questions from five testing blocks sequentially and received a learning intervention between any two adjacent testing blocks. Each testing block contains 10 items and represents a time point. To balance the item positions and avoid the

empirical identifiability problem, a total of five versions of the test were developed following the Latin square design, and they were randomly assigned to the participants to guarantee that different test blocks have the same chance of being the first block among all participants. Detailed descriptions of the test questions, learning intervention, and the entire experimental process can be found in Wang, Yang, et al. (2018a).

Essentially, this data set represents a longitudinal assessment with the repeated measure design. To facilitate the proposed model fitting process, we reorganized the data set using 350 individuals' responses to 250 items at five time points (50 items per time point). At a given time point, a participant only answered 10 items, and responses to the other 40 items remained missing because of the design. These missing data due to test design were considered as MCAR and were ignored in the data analysis. Figure 4 depicts box plots of the total raw score and the average logarithm of the RTs (logRTs) of all students on the 50 items at each time point (excluding the missing values). A clear decreasing trend of the average logRTs can be observed; as for the total raw score, it shows an increase between time points 1 and 2, and the changes between the other time points are negligible.

Analysis

In this application, latent ability represents a student's spatial rotation ability, while latent processing speed denote a student's latent speed when completing a rotation question. To each model, two Markov chains were applied. In each chain, 10,000 iterations were performed, with the first 5000 iterations as burn-in, and the remaining 5000 iterations (10,000 per chain in total) performed for model parameter inference. PyMC3 provides different initialization schemes

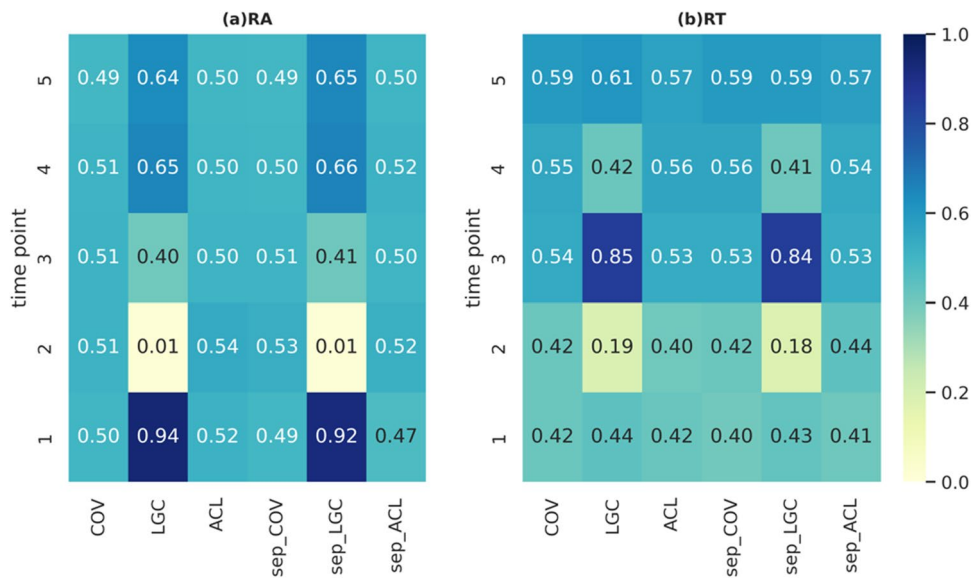


Fig. 5 Posterior predictive probability values of six models to response accuracy (RA) and response time (RT) data at each time point in empirical example 1. Note. COV unstructured covariance

matrix-based longitudinal joint model, LGC latent growth curve longitudinal joint model, ACL autoregressive cross-lagged longitudinal joint model; sep- separate modeling

for MCMC chains, as well as a set of tools for automatically diagnosing convergence after sampling.⁴ In this study, the *advi+adapt_diag* initialization scheme, which runs automatic differentiation variational inference and, subsequently, adapts the resulting diagonal mass matrix on the basis of the variance of the tuning samples, was used. This scheme initializes the chain at the test value, which depends on the prior distribution and is usually the mean or mode of the prior distribution. We checked for convergence using the criterion that the potential scale reduction factor (PSRF) should be less than 1.1 (Brooks & Gelman, 1998) or 1.2 (de la Torre & Douglas, 2004).

Posterior predictive model checking (PPMC; Gelman et al., 2014) was used to evaluate the absolute model–data fit. A posterior predictive probability (*ppp*) value near 0.5 indicates no systematic differences between the realized and predictive values and, thus, an adequate model fit. By contrast, a *ppp* value smaller than 0.05 or larger than 0.95 was considered the indication of an inadequate model fit. In this study, the *ppp* value was computed separately for each time point. Only the differences between the observed and predicted data were compared, and these differences were used to compute the *ppp* values (Levy & Mislevy, 2016). Specifically, for both RA and logRT on each time point, the differences between the observed data, *X*, and posterior predicted

data, $X^{postpred}$, were compared in computing the PPMC, as $ppp = \sum_{e=1}^E (Sum(X^{postpred(e)} \geq Sum(X)) / E$, where *E* is the total number of iterations in MCMC sampling; $X^{postpred(e)}$ were the posterior predicted data in *e*-th iteration, which were generated from the item response function (Eq. (1) for RA and Eq. (2) for logRT) based on the samplings of the model parameters from the posterior distributions. The deviance information criterion (DIC) and the widely available information criterion (WAIC) (Gelman et al., 2014) were computed for model selection. Smaller DIC and WAIC values indicate a better model–data fit.

Results

The PSRFs of all the parameters in each model were less than 1.1, indicating good convergence under the specified setting. Figure 5 depicts the *ppp* values of the six models for the RA and RT data of each time point. The results of absolute fitting of the data with the joint model and its corresponding separate model are generally consistent. Both the COV and the ACL were able to fit these data well at all five time points. As for the LGC, its RA model was not able to fit the RA data at time point 2, mainly because the change in total raw score was not linear with respect to time (as shown in Fig. 4), but its RT model was able to fit the RT data.

Table 2 summarizes the relative model–data fit indices of the six models and their computation times. According to the $-2LL$, among the six models, the COV model provided the best fit for both RA and RT data, regardless

⁴ NUTS convergence and sampling speed are highly dependent on the choice of mass matrix. Different methods for choosing or adapting the mass matrix can be found in <https://docs.pymc.io/en/latest/api/generated/pymc.initnuts.html>.

Table 2 Relative model–data fits and computation times (in seconds) of six models in empirical example 1

Analysis model	Response accuracy model			Response times model			Computation time
	–2LL	DIC	WAIC	–2LL	DIC	WAIC	
COV	18,859.25	19,611.46	19,245.21	39,104.56	40,170.01	39,655.39	3362.525
LGC	18,897.69	19,497.79	19,204.12	39,147.48	40,013.14	39,592.88	4924.093
ACL	18,892.19	19,484.25	19,194.67	39,144.22	40,028.14	39,599.17	6441.862
sep-COV	18,867.10	19,574.90	19,230.10	39,120.69	40,108.27	39,630.25	3069.254
sep-LGC	18,894.14	19,516.55	19,212.36	39,144.47	40,029.14	39,599.98	4659.859
sep-ACL	18,893.79	19,502.56	19,204.93	39,145.14	40,036.33	39,603.91	5940.281

COV unstructured covariance matrix-based longitudinal joint model, LGC latent growth curve longitudinal joint model, ACL autoregressive cross-lagged longitudinal joint model, sep- separate modeling, –2LL –2 log likelihood, DIC = deviance information criterion, WAIC widely available information criterion. The numbers in bold type face indicate the smallest values

of model complexity, because it was the most generalized model. When the model complexity penalty was considered, in terms of the DIC and WAIC, the ACL model provided the best fit for the RA data, and the LGC model provided the best fit for the RT data. In addition, the joint model generally fitted the data better than its corresponding separate model,

except the COV. A possible reason is that the COV incorporates more covariance coefficients between latent ability and processing speed, resulting in larger complexity than the sep-COV, and hence higher DIC and WAIC. Furthermore, because the separate models are more parsimonious than the joint models, the computation times of the former

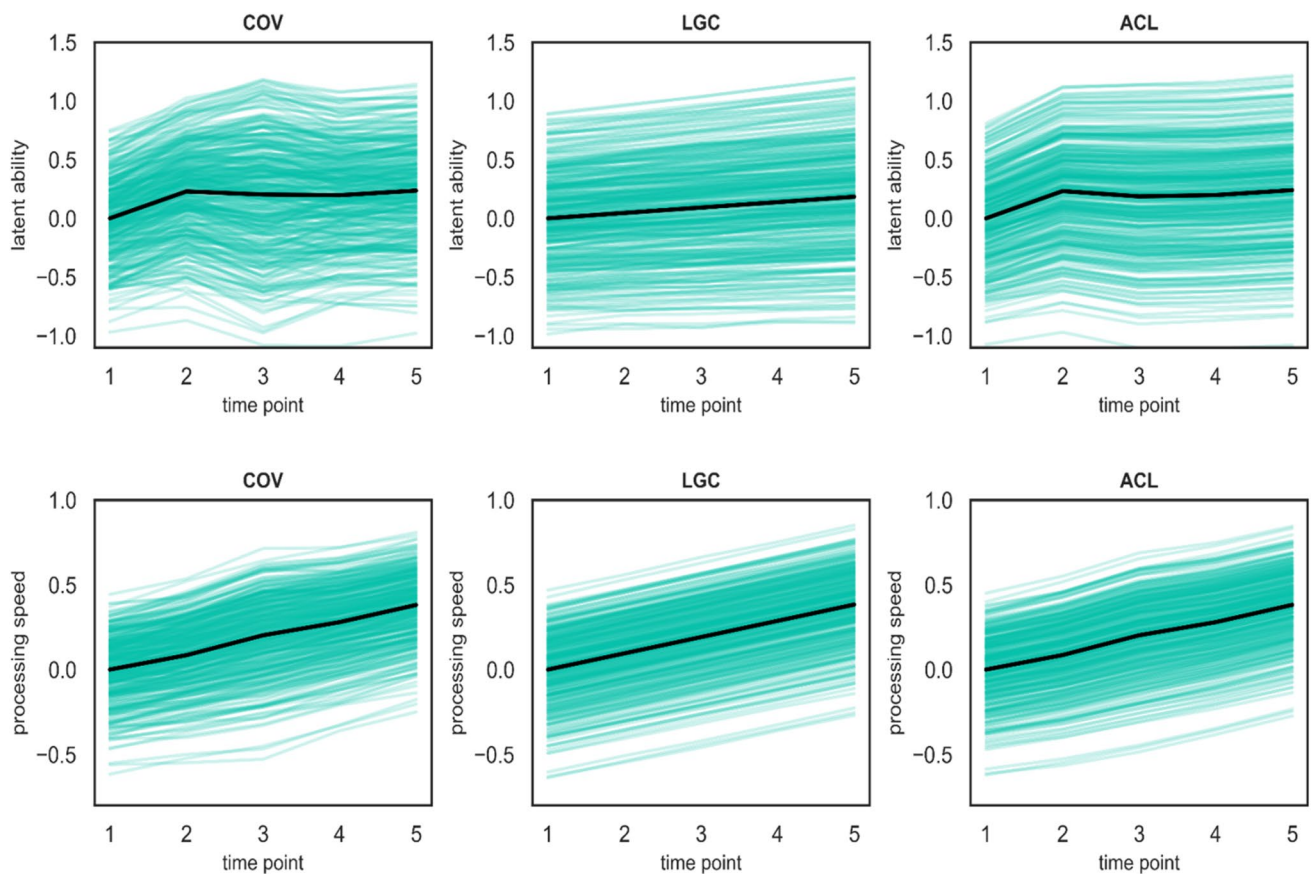


Fig. 6 Estimated developmental trajectories of latent ability and processing speed in empirical example 1. Note. COV unstructured covariance matrix-based longitudinal joint model, LGC latent growth curve

longitudinal joint model, ACL autoregressive cross-lagged longitudinal joint model. The thick solid line represents the developmental trajectory of the population mean

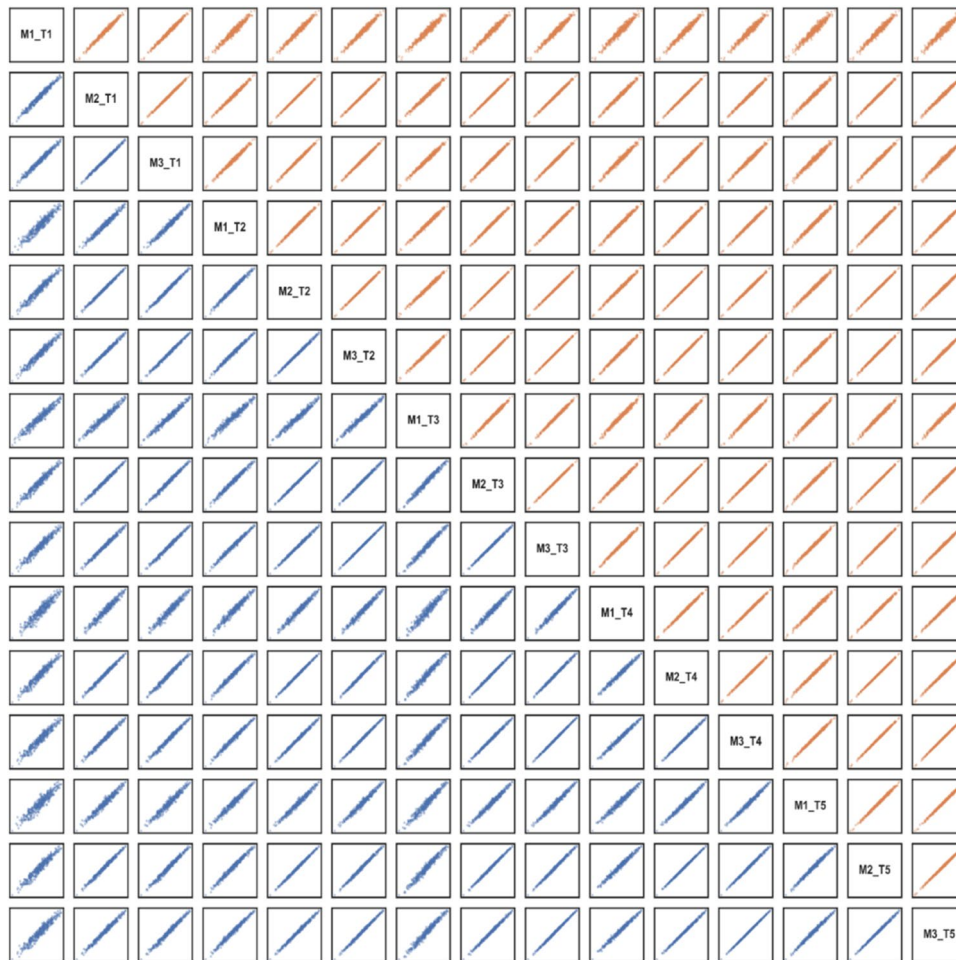


Fig. 7 Scatterplot between estimates of latent ability and processing speed in empirical example 1. *Note.* *M1* unstructured-covariance-matrix-based longitudinal joint model, *M2* latent growth curve longitudinal joint model, *M3* autoregressive cross-lagged longitudinal joint model, *T_x* *x*th time point. The upper and lower triangular matrices

represent the estimates of processing speed and that of latent ability, respectively. The position of each dot on the horizontal and vertical axes in each scatter plot represents the estimates of the two models corresponding to the row and column in which the plot is located

were shorter than those of the latter. The following section discusses the results of the joint models.

Figure 6 displays the developmental trajectories of latent ability and processing speed estimated using the three joint models. First, there was a high degree of consistency among the developmental trajectories estimated using the three models, especially between the results from the COV and ACL models. Second, the latent ability increased clearly between time point 1 and time point 2, but there was almost no change between time point 2 and time point 5. Meanwhile, the processing speed exhibited a steady increasing trend from time point 1 to time point 5 for all three models. These results are consistent with the findings of previous studies (Wang et al., 2020; Wang and Chen, 2020), and they explain the change trends of the observed variables in Fig. 4. Moreover, these results

reflect the benefit of additional RT data analysis. That is, the benefit of using the designed intervention is not only in improving the latent ability but also in increasing processing speed., which cannot be detected if using RA data alone.

Furthermore, in the case of the COV model, the estimated scale changes of latent ability and processing speed over time (i.e., $\hat{\sigma}_{\theta(p+1)}/\hat{\sigma}_{\theta p}$ and $\hat{\sigma}_{\tau(p+1)}/\hat{\sigma}_{\tau p}$, respectively) were (1.093, 1.287, 0.849, 1.020)' and (1.038, 1.120, 0.902, 0.970)', respectively. In the case of the ACL model, the $\hat{\sigma}_{\varepsilon_{\theta(p+1)}}/\hat{\sigma}_{\varepsilon_{\theta p}}$ and $\hat{\sigma}_{\varepsilon_{\tau(p+1)}}/\hat{\sigma}_{\varepsilon_{\tau p}}$ over time were (1.110, 1.073, 0.995, 1.008)' and (1.038, 1.051, 0.964, 0.988)', respectively. In the case of the LGC model, neither $\hat{\rho}_{\pi_0\pi_1} = 0.296$ (95% highest posterior density [HDP] of [-0.375, 0.931]) nor $\hat{\rho}_{\delta_0\delta_1} = 0.039$ (95% HDP of [-0.664, 0.746]) was significantly different from zero. These results indicate the absence

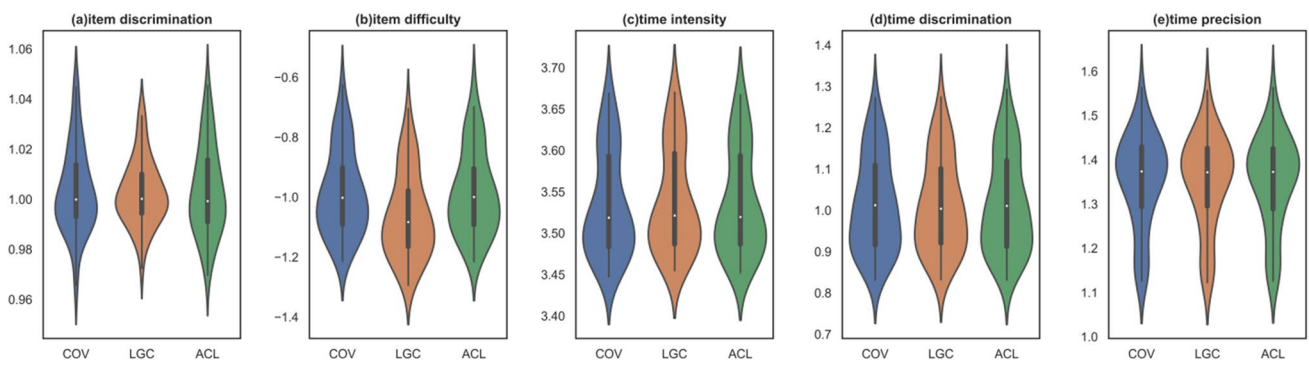


Fig. 8 Violin plot of estimates of item parameters in empirical example 1. *Note.* *COV* unstructured-covariance-matrix-based longitudinal joint model, *LGC* latent growth curve longitudinal joint model, *ACL* autoregressive cross-lagged longitudinal joint model

Table 3 Estimates of structural model parameters (i.e., correlation coefficients) in the unstructured-covariance-matrix-based longitudinal joint model in empirical example 1

	θ_1	θ_2	θ_3	θ_4	θ_5
τ_1	0.420	0.319	0.426	0.299	0.295
τ_2	0.296	0.279	0.307	0.192 [#]	0.235
τ_3	0.314	0.308	0.362	0.224	0.287
τ_4	0.290	0.218 [#]	0.286	0.205 [#]	0.221
τ_5	0.269	0.230	0.304	0.185 [#]	0.259

[#] means that the 95% highest posterior density (HPD) includes zero, that is, the correlation coefficient is not significantly different from zero. 95% HPD of each estimate was presented in Table S10 in the online appendix

of a significant Matthew effect⁵ for the changes in latent ability and processing speed.

Figure 7 presents the scatter plot between the estimates of latent ability and processing speed across all time points for all models. First, for each model, both latent ability and processing speed exhibit a high degree of correlation among the estimates at five time points, indicating a high degree of consistency in latent ability or processing speed at different time points (mainly because of the short interval between the time points in this test). Second, the estimates of both latent ability and processing speed obtained using different models at a given point exhibit a high degree of correlation, indicating a high degree of consistency between the estimates yielded by different models.

Figure 8 displays the estimates of the item parameters of all the models considered in this empirical study. The item parameter estimation was highly consistent across the three models, especially the COV and ACL models. Most of the item difficulty parameter estimates were negative, indicating that the overall difficulty of the test was low, and the

⁵ The Matthew effect means that participants starting out at a higher level of proficiency gain more on average than participants starting at a lower proficiency level.

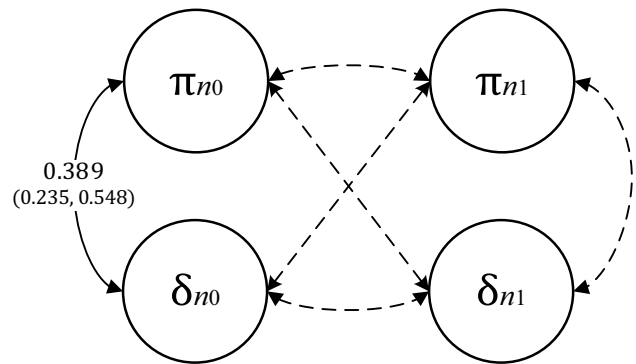


Fig. 9 Estimates of structural model parameters (i.e., correlation coefficients) of the latent growth curve longitudinal joint model in empirical example 1. *Note.* 95% highest posterior density (HPD) in parentheses. The dotted line indicates that the correlation coefficient is not significantly different from zero (i.e., 95% HPD includes zero). 95% HPD of each estimate was presented in Fig. S3 in the online appendix

resulting ceiling effect may explain the observation that the average total raw score hardly changed after time point 2.

In addition to the comparative results of the three above-mentioned models, a few characteristics of the data itself can be identified using the parameter estimates of the different models. For example, Table 3 presents the estimates of correlations among latent ability and latent speed across all time points in the COV model. A low to medium degree of correlation was found between latent ability and latent speed at each of the time points. Figure 9 shows the estimates of the structural model parameters (i.e., correlation coefficients) of the LGC. In these data, there exists a correlation only between the starting level of latent ability and the starting level of processing speed. Moreover, there is no interaction between the development of these two constructs over time. Figure 10 presents the estimates of the structural model parameters (i.e., autoregressive coefficients and cross-lagged coefficients) of the ACL model. In these

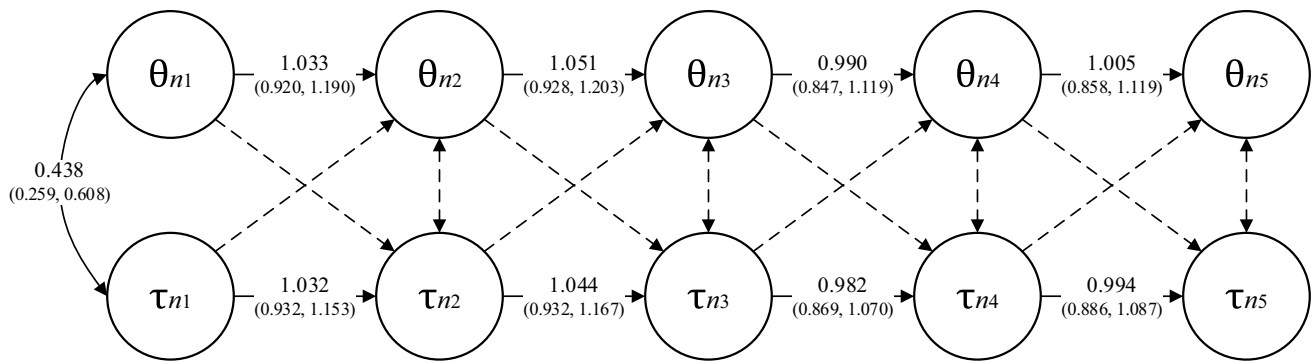


Fig. 10 Estimates of structural model parameters (i.e., path coefficients) of the autoregressive cross-lagged longitudinal joint model in empirical example 1. *Note.* 95% highest posterior density (HPD) in parentheses. The dotted line indicates that the path coefficient is not

significantly different from zero (i.e., 95% HPD includes zero). 95% HPD of each estimate was presented in Figure S4 in the online appendix

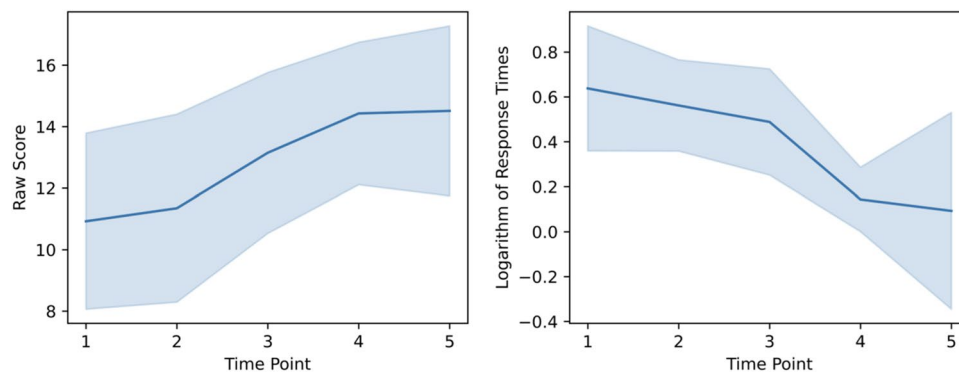


Fig. 11 Change trends in total raw score and average response times over time (with standard deviation range)

data, a parallel relationship exists between developmental change in latent ability and developmental change in processing speed, and they do not affect each other over time.

Empirical example 2: Subitizing task in developmental psychology

Data description and analysis

This subitizing task was adapted from LeFevre et al. (2010). In the task, the participants were presented with a series of black dots (the number of dots ranged from 1 to 4) on a computer screen and asked to indicate verbally the number of dots they saw as fast and accurately as possible. In each trial, the picture of the dots was displayed after a 200-ms mask. The testers recorded the RA and RT data of each participant by pressing a key. The test consists of 18 items, and each series of dots (containing 1–4 dots) was presented thrice, albeit at different locations on the screen. A total of 204 Cantonese-speaking children participated in the task. They

were followed for three years across five test time points.⁶ At each of these time points, 190, 189, 183, 180, and 186 children completed the same task containing 18 items. The data set used herein was deployed in a previous study related to development psychology (Ouyang et al., 2022).

As depicted in Fig. 11, the average raw score increased over time, while the average RT decreased over time, possibly indicating that the participants' latent ability and latent processing speed tended to increase and decrease, respectively, over time. Little's test of MCAR was performed (across the five time points, approximately 8.85% of the data was missing), and the results indicated that the null hypothesis "the missing data are MCAR" cannot be rejected (for RA: $\chi^2 = 534.863$, $df = 522$, $p = 0.339$; for RT: $\chi^2 = 528.768$, $df = 522$, $p = 0.409$). The analysis process was identical to that followed for empirical example 1.

⁶ The original data contained a total of seven time points, and owing to the small number of children at the first two time points, only the data of the last five time points were used in this study (new children were added starting from the third time point).

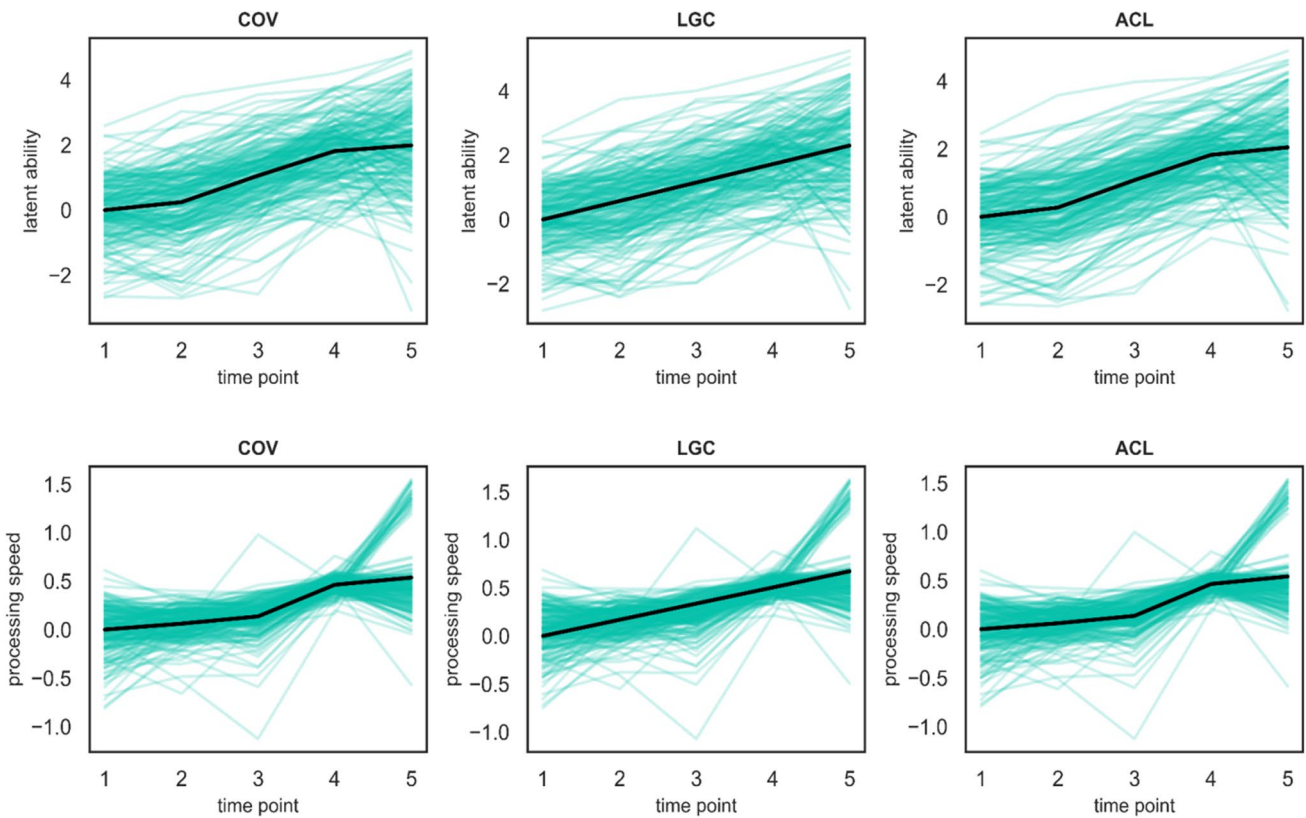


Fig. 12 Estimated developmental trajectories of latent ability and processing speed in empirical example 2. *Note.* COV = unstructured-covariance-matrix-based longitudinal joint model, LGC = latent

growth curve longitudinal joint model, ACL = autoregressive cross-lagged longitudinal joint model. The thick solid line represents the developmental trajectory of the population mean

Results

For simplicity, only a few main results are discussed here, and details about the results obtained for this data set can be found in Section S3.3 in the online appendix. The PSRFs of all the parameters of the six models were less than 1.1. Figure S5 in the online appendix displays the *ppp* values of the six models applied to the RA and RT data at each time point. The results of absolute fitting the data with the joint

model and its corresponding separate model are generally consistent. Overall, the COV and ACL models were able to fit the data at all time points (nearly unfitted to the RTs at

Table 4 Estimated structural model parameters (i.e., correlation coefficients) of the unstructured-covariance-matrix-based longitudinal joint model in empirical example 2

	θ_1	θ_2	θ_3	θ_4	θ_5
τ_1	0.498	0.198	0.218	0.292	0.250
τ_2	0.406	0.340	0.395	0.161 [#]	0.416
τ_3	0.405	0.300	0.430	0.208	0.240
τ_4	0.182 [#]	0.050 [#]	0.266	0.264	0.259
τ_5	0.128 [#]	0.230	0.060 [#]	0.125 [#]	0.148 [#]

[#] indicates that the 95% highest posterior density (HPD) includes zero, that is, the correlation coefficient is not significantly different from zero. 95% HPD of each estimate was presented in Table S12 in the online appendix

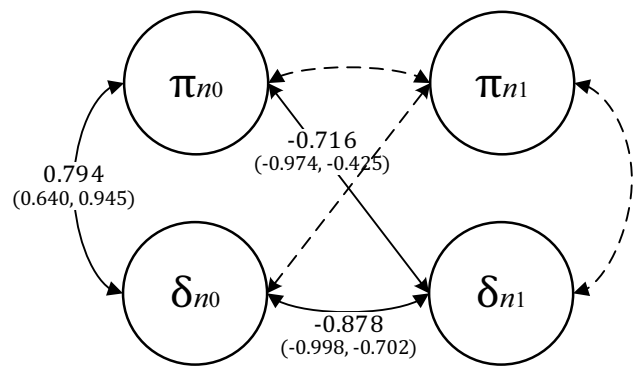


Fig. 13 Estimated structural model parameters (i.e., correlation coefficients) of the latent growth curve longitudinal joint model in empirical example 1. *Note.* 95% highest posterior density (HPD) in parentheses. The dotted line indicates that the correlation coefficient is not significantly different from zero (i.e., 95% HPD includes zero). 95% HPD of each estimate was presented in Figure S8 in the online appendix

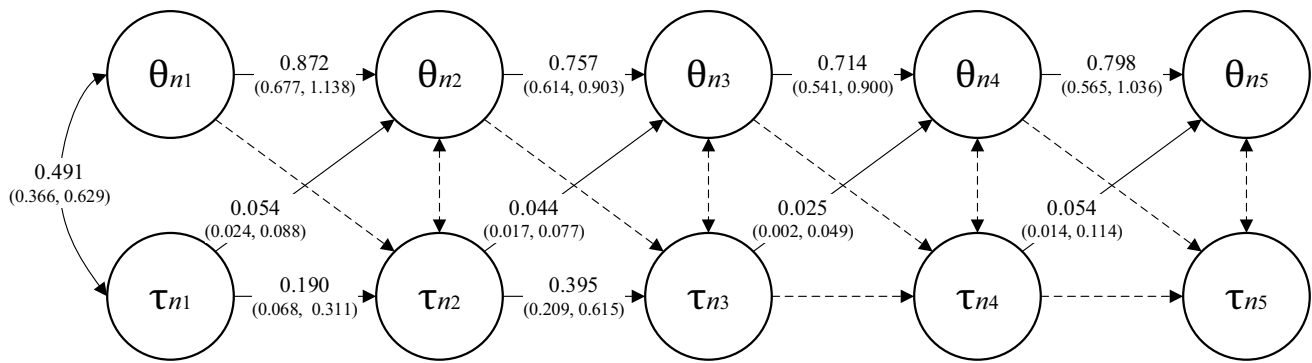


Fig. 14 Estimated structural model parameters (i.e., path coefficients) of the autoregressive cross-lagged longitudinal joint model in empirical example 1. *Note.* 95% highest posterior density (HPD) in paren-

theses. The dotted line indicates that the path coefficient is not significantly different from zero (i.e., 95% HPD includes zero). 95% HPD of each estimate was presented in Figure S9 in the online appendix

time point 5), while the LGC model was nearly unfitted to the RA data at time points 1 and 2, and it did not fit the RT data at time points 3–5. Table S11 in the online appendix summarizes the relative model–data fit indices and the computation times of the six models. The ACL model was found to be preferable on the basis of the DIC and WAIC values for both RA and RT data. Since the joint model generally fitted the data better than its corresponding separate model, the following section discusses the results of the joint models.

Figure 12 displays the estimated developmental trajectories of latent ability and processing speed of the three proposed models. There was a high degree of consistency in the estimated developmental trajectories from the three joint models. The estimated mean population latent ability and latent processing speed both increased over time. In addition, there was a significant upward shift in the processing speed of several children at time point 5, indicating the existence of heterogeneity among the children at that time point for some reason (e.g., rapid guessing owing to poor motivation).

Figure S6 in the online appendix shows a scatter plot between the estimates of latent ability and processing speed across all time points for all models. The results indicate a high degree of consistency between the estimates obtained using different models, as well as a moderate correlation between latent abilities at different time points and a weak correlation between processing speeds at different time points. A possible reason for this phenomenon is the long interval between time points. Figure S7 displays the estimated item parameters of all models in the empirical study. There was a high level of agreement between the estimates of each of the item parameters of the three models.

Table 4 presents the estimated correlations among latent abilities and latent speed across all time points in the COV model. A low to moderate degree of correlation was found between latent ability and latent speed at each time point.

Figure 13 shows the estimates of structural model parameters of the LGC model.⁷ The results indicate that (a) a high degree of correlation existed between the starting levels of latent ability and processing speed, and (b) the increase in processing speed was slower for children with high starting levels of latent ability and processing speed. Figure 14 displays the estimates of the structural model parameters of the ACL model. The results mainly indicated that the latent ability at the current time point was positively influenced by both latent ability and latent processing speed at the previous time point. These parallel developmental phenomena with cross-time relationships between latent ability and processing speed cannot be observed from RA data alone.

Simulation study

A simulation study was conducted to further explore the psychometric performance of the proposed models in different test scenarios. Note that we did not intend to compare the performance across the three modeling approaches, because they have their own modeling assumptions, and therefore cannot be compared fairly using the same data-generation mechanism.

Design, data generation, and analysis

For each model, we considered three factors, namely, sample size $N = 250$ and 500 , test length at each time point $I_p = 15$ and 30 , and the number of test time points, $P = 3$ and 5 . We

⁷ Notably, the LGC model was not able to fit these data, and the results were primarily used to demonstrate the practical use of the LGC model for reference purposes only.

assume the same items were used at each time point, as in the empirical studies. The true item parameters were simulated by referring to those in the literature (e.g., Bolsinova & Tijmstra, 2018; Fox & Mariani, 2016; Man et al., 2019; Wang, Zhang, et al., 2018b; Zhan et al., 2018). Specifically, $b_{it} = b_i \sim N(\mu_b, \sigma_b^2) = N(0, 1)$, $a_{it} = a_i \sim N(\mu_a, \sigma_a^2) = N(1, 0.05)$, $\xi_{ip} = \xi_i \sim N(\mu_\xi, \sigma_\xi^2) = N(4, 0.25)$, $\phi_{ip} = \phi_i \sim N(\mu_\phi, \sigma_\phi^2) = N(1, 0.05)$, and $\omega_{ip} = \omega_i \sim N(\mu_\omega, \sigma_\omega^2) = N(2, 0.05)$.

The person parameters of the COV model were generated using Eq. (3), where $\mu_0 = \mu_\tau = (0, 0.25, 0.5, 0.75, 1)'$. The generated unstructured covariance matrix Σ is presented in Table S13 in the online appendix, and it was generated mimicking the $\hat{\Sigma}$ in empirical example 1. The person parameters of the LGC model were generated using Eqs. (4) and (5). Four growth factors were generated using Eq. (6), where $\mu_{\pi_0} = \mu_{\delta_0} = 0$, $\mu_{\pi_1} = \mu_{\delta_1} = 0.25$, $\sigma_{\pi_0}^2 = \sigma_{\delta_0}^2 = 1$, $\sigma_{\pi_1}^2 = \sigma_{\delta_1}^2 = 0.02$, $\rho_{\delta_0\pi_0} = 0.5$, and the correlations between all the other growth factors were set to 0.2. The residual terms were generated from $\varepsilon_{\theta_{np}} \sim N(0, \sigma_{\theta_p}^2) = N(0, 0.001)$ and $\varepsilon_{\tau_{np}} \sim N(0, \sigma_{\tau_p}^2) = N(0, 0.001)$. The person parameters of the ACL model were generated using Eqs. (7) and (8), where $\beta_{01} = \lambda_{01} = 0$, $\beta_{0p \geq 2} = \lambda_{0p \geq 2} = 0.25$, $\beta_{1p} = \lambda_{1p} = 0.95$, and $\beta_{2p} = \lambda_{2p} = 0.1$. In Eq. (9), $\begin{pmatrix} \varepsilon_{\theta_{n1}} \\ \varepsilon_{\tau_{n1}} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$ and $\begin{pmatrix} \varepsilon_{\theta_{np \geq 2}} \\ \varepsilon_{\tau_{np \geq 2}} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.001 & 0 \\ 0 & 0.001 \end{pmatrix}\right)$. Finally, for each of the three joint models, 50 data sets were generated for each simulated test situation. In each data set, the model parameters were regenerated from the same distribution mentioned above. The three joint models and their corresponding separate models were used to fit the data.

To evaluate the estimation accuracy, we computed bias and root mean square error (RMSE) as $\text{bias}(\hat{x}) = \frac{\sum_{r=1}^{50} \hat{x}_r - x_r}{50}$ and $\text{RMSE}(\hat{x}) = \sqrt{\frac{\sum_{r=1}^{50} (\hat{x}_r - x_r)^2}{50}}$, respectively, where x_r is the true value of the model parameter generated in replicate r ($r = 1, \dots, 50$), and \hat{x}_r is the corresponding parameter estimate in replicate r . In addition, the correlations between the generated and estimated values (denoted as Cor) of the model parameters were computed. Other analysis processes were identical to those in the empirical examples, except that the length of each Markov chain was shortened to 4000 iterations, of which the first 1000 iterations were earmarked for burn-in. The PSRFs of all model parameters were less than 1.1, suggesting good convergence under the specified setting.

Results

Table 5 presents the recovery of the item parameters of the three joint models (Table S14 in online appendix presents the

recovery of the item parameters of three separate models). Except for the time precision parameter (ω), which is largely unaffected, the recoveries of the other item parameters of the joint model were better than those of the corresponding separate model. For three joint models, it was possible to recover all the item parameters across all simulated conditions, where all the biases were close to zero, RMSEs were less than 0.15 for the RA model parameters and less than 0.1 for the RT model parameters, and all Cor values were close to 1. A larger sample size and a longer test length improved the recovery of the item parameters. However, the test length had a relatively weaker effect than the sample size. In addition, the number of time points had little effect on the recovery of the item parameters.

Table 6 presents the RMSE and Cor values of the person parameter estimates obtained using the three models (Table S15 in online appendix presents the RMSE and Cor values of the person parameter estimates of three separate models). In addition, bias values of the six models are presented in Tables S16 and S17, respectively, in the online appendix. The recoveries of two person parameters of the joint model were better than those of the corresponding separate model. For three joint models, the recovery of latent processing speed was better than that of latent ability. A longer test length improved the recovery of the person parameters. The increase in the number of time points was beneficial for recovery in the ACL model, but it had no effect on recovery for the COV and LGC models. The sample size had little effect on the recovery of the person parameters for all models.

Furthermore, Table S18 in online appendix presents the average computation time for all models under all conditions. Generally speaking, the computation time of the joint model was longer than that of its corresponding separate model, and the computation time increased as the sample size, number of time points, and number of items increased.

Conclusion and discussion

In this study, we proposed three longitudinal joint modeling approaches to simultaneously analyze longitudinal RA and RT data for assessing the parallel interactive development of latent ability and processing speed: the COV, LGC, and ACL models. Analysis results from these models can not only inform the developmental trajectories of latent ability and processing speed individually, but also show the relationship between changes in latent ability and processing speed through the across-time relations of these two constructs. The results of the empirical studies indicate that (1) all three joint models are practically applicable, and their conclusions were highly consistent in terms of the changes in ability and speed in the analysis of the same data set, and

Table 5 Recovery of item parameters of joint models in simulation study

Model	P	N	I_p	bias					RMSE					Cor				
				a	b	ϕ	ξ	ω	a	b	ϕ	ξ	ω	a	b	ϕ	ξ	ω
COV	3	250	15	0.001	-0.003	-0.006	0.000	0.001	0.108	0.116	0.069	0.018	0.057	0.999	0.999	0.999	0.999	0.999
			30	0.001	0.001	0.005	0.000	-0.003	0.104	0.114	0.066	0.018	0.053	0.999	0.999	0.999	0.999	0.999
		500	15	0.000	0.005	0.009	0.000	-0.002	0.080	0.083	0.041	0.013	0.040	0.999	0.999	0.999	0.999	0.999
			30	-0.001	-0.002	-0.010	0.000	0.000	0.076	0.078	0.043	0.012	0.039	0.999	0.999	0.999	0.999	0.999
	5	250	15	-0.001	0.018	0.020	0.000	-0.002	0.079	0.116	0.064	0.013	0.043	0.999	0.999	0.999	0.999	0.999
			30	0.000	-0.009	0.001	0.000	0.000	0.081	0.096	0.056	0.013	0.040	0.999	0.999	0.999	0.999	0.999
		500	15	0.000	0.010	0.011	0.000	0.001	0.060	0.077	0.051	0.009	0.030	0.999	0.999	0.999	0.999	0.999
			30	0.001	0.004	0.005	0.000	0.001	0.060	0.070	0.056	0.009	0.030	0.999	0.999	0.999	0.999	0.999
LGC	3	250	15	0.001	-0.016	-0.005	0.000	-0.003	0.096	0.122	0.078	0.018	0.055	0.998	0.998	0.998	0.998	0.998
			30	0.000	-0.010	-0.019	0.000	-0.001	0.096	0.117	0.070	0.018	0.053	0.999	0.999	0.999	0.999	0.999
		500	15	0.000	-0.004	0.009	0.000	-0.002	0.076	0.087	0.053	0.012	0.040	0.999	0.999	0.999	0.999	0.999
			30	0.000	-0.013	-0.005	0.000	0.000	0.070	0.091	0.054	0.013	0.038	0.999	0.999	0.999	0.999	0.999
	5	250	15	0.000	0.020	0.006	0.000	-0.001	0.078	0.105	0.068	0.013	0.043	0.999	0.999	0.999	0.999	0.999
			30	-0.001	0.022	0.022	0.000	0.001	0.076	0.104	0.075	0.013	0.042	0.999	0.999	0.999	0.999	0.999
		500	15	0.000	0.002	-0.003	0.000	-0.002	0.056	0.069	0.058	0.009	0.030	0.999	0.999	0.999	0.999	0.999
			30	0.000	0.012	0.003	0.000	0.000	0.056	0.074	0.054	0.009	0.028	0.999	0.999	0.999	0.999	0.999
ACL	3	250	15	0.001	0.013	-0.001	0.000	0.000	0.097	0.113	0.070	0.018	0.054	0.999	0.999	0.999	0.999	0.999
			30	0.000	0.000	-0.007	0.000	0.002	0.098	0.117	0.082	0.018	0.052	0.999	0.999	0.999	0.999	0.999
		500	15	0.001	-0.004	-0.004	0.000	0.000	0.072	0.083	0.039	0.012	0.038	0.999	0.999	0.999	0.999	0.999
			30	0.001	0.001	-0.002	0.000	0.000	0.072	0.089	0.047	0.013	0.037	0.999	0.999	0.999	0.999	0.999
	5	250	15	-0.001	0.006	0.006	0.000	-0.003	0.076	0.108	0.072	0.012	0.042	0.999	0.999	0.999	0.999	0.999
			30	0.002	0.000	-0.003	0.000	-0.001	0.073	0.100	0.065	0.013	0.042	0.999	0.999	0.999	0.999	0.999
		500	15	0.000	-0.013	0.000	0.000	0.000	0.052	0.074	0.041	0.009	0.029	1.000	1.000	1.000	1.000	1.000
			30	0.001	-0.004	0.007	0.000	-0.001	0.054	0.076	0.047	0.009	0.029	1.000	1.000	1.000	1.000	1.000

P number of time points, N sample size, I_p test length at each time point, *COV* unstructured-covariance-matrix-based longitudinal joint model, *LGC* latent growth curve longitudinal joint model, *ACL* autoregressive cross-lagged longitudinal joint model, a item discrimination, b item difficulty, ϕ time discrimination, ξ time-intensity, ω time precision, *RMSE* root mean square error, *Cor* correlation between the generated and estimated values

(2) additional analysis of RT data and acquisition of individual processing speed measurements can reveal the parallel interactive development phenomena that are difficult to detect when using RA data alone. In other words, without the measures of processing speed, the interaction between latent ability and processing speed cannot be detected during the development process. In addition, the results of the simulation study demonstrate that the proposed Bayesian MCMC estimation algorithm can ensure accurate model parameter recovery for all three proposed longitudinal joint models. In addition, for both empirical examples and simulation studies, we also compared the fit of joint models and of the separate models on one type of data at a time. Results of these comparisons indicate that the joint analysis of bimodal data by considering the relationships among different latent variables can improve model–data fit and model parameter recovery to some extent compared with the separate models for one type of data at a time. On one hand, we do note that this conclusion may depend on the

correlation between latent variables. For example, when the correlation between latent ability and processing speed is low, the joint model does not necessarily fit the data better than its corresponding separate model, especially if one uses the relative model–data fit indices which account for the model complexity penalty. On the other hand, the result from the joint models allows us to capture the exact magnitude of the correlation between different latent variables, which is more informative than not knowing or considering the correlation between them.

The choice of the modeling approach depends on the specific context of a longitudinal analysis. The COV model can be used if the primary objective is to determine the overall growth at each time point. The LGC model is useful for obtaining the developmental trajectories of latent ability and processing speed, such as the consistency of or differences in changes between the two constructs. The ACL model is suitable when the objective is to analyze the self-influences of latent ability and processing speed and

determine the relationships between the two constructs at adjacent time points. In addition, the LGC model did not fit well to the two empirical data sets in this study. This is mainly due to its strictest assumptions related to the developmental trajectories of the latent constructs (i.e., linear developmental trajectory) compared with the other two models, and these assumptions are difficult to satisfy in practice. By contrast, the COV and ACL models have more lenient assumptions related to the developmental trajectory of the latent constructs, and therefore they are more inclusive of practice data than the LGC model.

The results of this study shed light on the use of both RA and RT data to solve problems in various subdomains of psychology, including clinical psychology and developmental psychology, which rely heavily on the use of RA data alone. For example, the results of common psychiatric scales in clinical psychology tend to be interpreted using RA data alone. Specifically, decisions regarding whether an individual has a certain mental illness or abnormal tendency are made by comparing the individual's scale score with a certain cutoff point. However, the outward manifestations of an individual's mental illness may not be fully expressed using their choices of Likert-type items. For example, usually, people with depression or depressive tendencies take more time to complete the depression scale compared to people who are not affected by depression. In this sense, the joint analysis of RA and RT data might provide superior diagnoses of mental illnesses. In addition, as has been demonstrated in our empirical data analysis, the joint analysis of longitudinal RA and RT data may reveal a significant change in speed when the change in ability is not significant (example 1). This might provide additional information for assessing the effectiveness of the interventions related to mental health when a repeated measure design is adopted.

Next, we discuss the limitations of our study. First, we have only considered the single-group situation in this study. That is, all individuals within the population are assumed to have homogeneous average developmental trajectory for a specific construct. To address the heterogeneity of developmental trajectories among individuals, we can extend the current models through the multigroup modeling (e.g., von Davier et al., 2011) and mixture modeling (e.g., Muthén & Shedden, 1999; Zhang & Wang, 2019) frameworks in the future. Second, herein, we assume the lack of item parameter drift throughout all analyses. This means that all item parameters (e.g., item difficulty and time-intensity) of the same item are assumed to be invariant over time. However, item parameter drift might be expected in repeated measure design, and in that case, more work would be required to address this issue when generating estimates using the proposed models. Third, although this study provides insights into the measurement of individual growth from a holistic perspective,

only two data modalities, namely RA and RT, and the constructs measured using these modalities, are considered. In recent years, with the increasing popularity of technology-enhanced assessments (Jiao & Lissitz, 2018), the acquisition of multimodal data beyond RA and RTs, such as action sequences (Han et al., 2021), eye-tracking (Man et al., 2022; Zhan et al., 2022), and brain activation (Jeon et al., 2021), has become possible. Technology-enhanced longitudinal assessments could be considered in the future to assess the parallel interactive development of multiple constructs (e.g., latent ability, processing speed, visual engagement, and brain activation) through the analysis of longitudinal multimodal data. Fourth, the proposed modeling approaches are longitudinal extensions of the cross-sectional joint-hierarchical modeling approach (van der Linden, 2007), in which separate measurement models are implemented for RA and RT data under conditional independence assumptions. A few recent studies have proposed the joint cross-loading modeling approach (e.g., Bolsinova & Tijmstra, 2018; Molenaar et al., 2015), which attempts to relax the conditional independence assumptions and improve the precision of latent ability estimation through direct extraction of information from the RT data. We will investigate longitudinal extension of the joint cross-loading modeling approach in the future. Fifth, given the focus of the present study on extending the applicability of the joint-hierarchical latent variable modeling approach (van der Linden, 2007), the proposed models draw on the three most commonly used longitudinal modeling approaches in SEM, but several new extensions of these approaches are not considered in the present study (e.g., Bianconcini & Bollen, 2018; Bishop et al., 2015; Hamaker et al., 2015; Kohli & Harring, 2013).

Finally, we discuss several other directions of study that can be explored in the future. First, in terms of the model assumptions related to dimensionality of the latent construct, one route to increase the capability of the proposed longitudinal joint models in terms of describing the interactions of students and items is to hypothesize that persons vary on a wide range of latent constructs. Cognitive science (Frederiksen et al., 1990) shows that subsets of those constructs are important for correct response to specific items. Hence, we could consider modeling the parallel interactive development of multidimensional latent ability (Reckase, 2009) and multidimensional processing speed (Zhan et al., 2021) in multidimensional longitudinal assessments. Another straightforward extension of the proposed models is to incorporate observed covariates (e.g., background variables and number of interventions) to explain the variations in individual growth. For example, for the COV model, the constructs at a specific time point may be regressed upon the observed covariates. For the LGC model, either the growth intercept or the growth slope or both can be regressed upon

the observed covariates to explain the variations in initial statuses or growth rates. For the ACL model, the observed covariates can be directly added to the regression model. Finally, dimensionality-reduction methods (e.g., Cai, 2010; Gibbons & Hedeker, 1992) can be investigated and incorporated into the model estimation procedure. This is especially necessary when there are large numbers of latent constructs and time points, which may lead to the high-dimensionality problem in the proposed joint models.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13428-023-02113-5>.

Code availability All code used in simulation study can be downloaded and used at this website: https://osf.io/fjdp8/?view_only=daf59a3da44145299bc7add54874dfb

Authors' contributions Peida Zhan contributed to the conception, design, and drafting and revision of the paper. Qipeng Chen contributed to the data analysis. Shiyu Wang and Xiao Zhang both contributed to manuscript revision and empirical data sharing.

Funding The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China (Grant number 31900795).

Data availability For simulation study, the data sets generated during and analyzed during the current study are available in the OSF repository, https://osf.io/fjdp8/?view_only=daf59a3da44145299bc7add54874dfb. For the first empirical study, the data set analyzed is available in the <https://github.com/tmsalab/hmcdm>. For the second empirical study, the data set analyzed is available from the fourth author on reasonable request, E-mail: xzhang1@hku.hk.

Declarations

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

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication The author(s) agreed to publish the article.

References

- Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16. <https://doi.org/10.1007/BF02294143>
- Bartolucci, F., Farcomeni, A., & Pennoni, F. (2013). *Latent Markov models for longitudinal data*. Chapman and Hall/CRC Press.
- Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, *31*, 419–456.
- Bianconcini, S., & Bollen, K. A. (2018). The latent variable-autoregressive latent trajectory model: A general framework for longitudinal data analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 791–808.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.
- Bishop, J., Geiser, C., & Cole, D. A. (2015). Modeling latent growth with multiple indicators: A comparison of three approaches. *Psychological Methods*, *20*(1), 43–62. <https://doi.org/10.1037/met000018>
- Bock, D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51.
- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Wiley-Interscience.
- Bolsinova, M., & Tilmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 13–38.
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455. <https://doi.org/10.2307/1390675>
- Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612. <https://doi.org/10.1007/s11336-010-9178-0>
- Chen, Y., & Culpepper, S. A. (2020). A multivariate probit model for learning trajectories: A fine-grained evaluation of an educational intervention. *Applied psychological measurement*, *44*(7–8), 515–530.
- Collins, L. M., Graham, J. W., Rousculp, S. S., & Hansen, W. B. (1997). Heavy caffeine use and the beginning of the substance use onset process: An illustration of latent transition analysis. In K. Bryant, K. M. Windle, & S. West (Eds.), *The Science of Prevention: Methodological Advances from Alcohol and Substance Use Research* (pp. 79–99). American Psychological Association.
- Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software*, *36*, 1–34.
- De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*, 102. <https://doi.org/10.3389/fpsyg.2019.00102>
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353. <https://doi.org/10.1007/BF02295640>
- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Lawrence Erlbaum Associates.
- Ferrer, E., & McArdle, J. J. (2010). Longitudinal modeling of developmental changes in psychological research. *Current Directions in Psychological Science*, *19*(3), 149–154.
- Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553.
- Frederiksen, N., Glaser, R., Lesgold, A., & Shafto, M. (1990). *Diagnostic monitoring of skill and knowledge acquisition*. Lawrence Erlbaum Associates.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. CRC Press.
- Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436. <https://doi.org/10.1007/BF02295430>
- Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21–35.
- Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. <https://doi.org/10.1037/a0038889>

- Han, Y., Liu, H., & Ji, F. (2021). A sequential response model for analyzing process data on technology-based problem-solving tasks. *Multivariate Behavioral Research*. Advanced Online. <https://doi.org/10.1080/00273171.2021.1932403>
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593–1623.
- Jeon, M., De Boeck, P., Luo, J., Li, X., & Lu, Z.-L. (2021). Modeling within-item dependencies in parallel data on test responses and brain activation. *Psychometrika*, 86, 239–271.
- Jiao, H., & Lissitz, R. W. (2018). *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective*. Information Age Publishing.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258–272.
- 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. <https://doi.org/10.1037/a0014877>
- Kohli, N., & Harring, J. R. (2013). Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research*, 48(3), 370–397.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.
- LeFevre, J. A., Fast, L., Skwarchuk, S. L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development*, 81(6), 1753–1767. <https://doi.org/10.1111/j.1467-8624.2010.01508.x>
- Leszczensky, L., & Wolbring, T. (2022). How to deal with reverse causality using panel data? Recommendations for researchers based on a simulation study. *Sociological Methods & Research*, 51(2), 837–865.
- Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. CRC Press.
- Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198–1202.
- Man, K., & Harring, J. R. (2021). Assessing preknowledge cheating via innovative measures: A multiplegroup analysis of jointly modeling item responses, response times, and visual fixation counts. *Educational and Psychological Measurement*, 81(3), 441–465. <https://doi.org/10.1177/0013164420968630>
- Man, K., Harring, J. R., Jiao, H., & Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement*, 43(8), 639–654. <https://doi.org/10.1177/0146621618824853>
- Man, K., Harring, J. R., & Zhan, P. (2022). Bridging models of biometric and psychometric assessment: A three-way joint modeling approach of item responses, response times, and gaze fixation counts. *Applied Psychological Measurement*, 46(5), 361–381.
- Mayer, L. S. (1986). On cross-lagged panel models with serially correlated errors. *Journal of Business & Economic Statistics*, 4(3), 347–357.
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, 60, 577–605.
- McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. American Psychological Association.
- Meijering, B., & Van Rijn, H. (2009). Experimental and computational analyses of strategy usage in the time-left task. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 31, No. 31).
- Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, 68(2), 197–219. <https://doi.org/10.1111/bmsp.12042>
- Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research*, 51(2), 606–626.
- Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, 24(6), 882–891.
- Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55(2), 463–469.
- Ouyang, X., Zhang, X., & Zhang, Q. (2022). Spatial skills and number skills in preschool children: The moderating role of spatial anxiety. *Cognition*, 225, 105165. <https://doi.org/10.1016/j.cognition.2022.105165>
- Paek, I., Park, H.-J., Cai, L., & Chi, E. (2014). A comparison of three IRT approaches to examinee ability change modeling in a single-group anchor test design. *Educational and Psychological Measurement*, 74, 659–676. <https://doi.org/10.1177/0013164413507062>
- Pan, Y., & Zhan, P. (2020). The impact of sample attrition on longitudinal learning diagnosis: A prolog. *Frontiers in Psychology*, 11, 1051. <https://doi.org/10.3389/fpsyg.2020.01051>
- Peugh, J., & Enders, C. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74, 525–556.
- Pokropek, A. (2011). Missing by design: Planned missing-data designs in social science. *ASK. Research & Methods*, 20, 81–105.
- Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling*, 55(4), 361–382.
- Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer.
- Rubin, D. (1976). Inference and missing data. *Biometrika*, 63, 581–592.
- Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2, e55.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2), 100.
- Siegler, R. S. (1989). Hazards of mental chronometry: An example from children's subtraction. *Journal of Educational Psychology*, 81, 497–506.
- Toh, S., & Hernán, M. A. (2008). Causal inference from longitudinal studies with baseline randomization. *The International Journal of Biostatistics*, 4(1), 22. <https://doi.org/10.2202/1557-4679.1117>
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181–204.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287–308.
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73(3), 365–384.
- van der Maas, H. L., & Jansen, B. R. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, 85(2), 141–177.
- von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, 76, 318–336. <https://doi.org/10.1007/s11336-011-9202-z>
- Wang, S., & Chen, Y. (2020). Using response times and response accuracy to measure fluency within cognitive diagnosis models. *Psychometrika*, 85(3), 600–629.

- Wang, C., & Nydick, S. W. (2020). On longitudinal item response theory models: A didactic. *Journal of Educational and Behavioral Statistics*, 45(3), 339–368. <https://doi.org/10.3102/1076998619882026>
- Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018a). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*, 43(1), 57–87.
- Wang, S., Zhang, S., Douglas, J., & Culpepper, S. (2018b). Using response times to assess learning progress: A joint model for responses and response times. *Measurement: Interdisciplinary Research and Perspectives*, 16(1), 45–58.
- Wang, S., Zhang, S., & Shen, Y. (2020). A joint modeling framework of responses and response times to assess learning outcomes. *Multivariate Behavior Research*, 55(1), 49–68. <https://doi.org/10.1080/00273171.2019.1607238>
- Woolf, B. P. (2009). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufman Publishers.
- Zhan, P. (2021). Refined learning tracking with a longitudinal probabilistic diagnostic model. *Educational Measurement: Issues and Practice*, 40(1), 44–58.
- Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, 71, 262–286.
- Zhan, P., Jiao, H., Liao, D., & Li, F. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics*, 44(3), 251–281. <https://doi.org/10.3102/1076998619827593>
- Zhan, P., Jiao, H., Man, K., Wang, W.-C., & He, K. (2021). Variable speed across dimensions of ability in the joint model for responses and response times. *Frontiers in Psychology*, 12, 469196.
- Zhan, P., Man, K., Wind, S., & Malone, J. (2022). Cognitive diagnosis modeling incorporating response times and fixation counts: Providing comprehensive feedback and accurate diagnosis. *Journal of Educational and Behavioral Statistics*, 47(6), 736–776. <https://doi.org/10.3102/10769986221111085>
- Zhang, S., & Wang, S. (2019). Modeling learner heterogeneity: A mixture learning model with responses and response times. *Frontiers in Psychology*, 9, 2339.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.