MULTIVARIATE BEHAVIORAL RESEARCH
2024, VOL. 59, NO. 1, 78–97

A Time-Varying Dynamic Partial Credit Model to Analyze Polytomous and Multivariate Time Series Data

Sebastian Castro-Alvarez[a] , Laura F. Bringmann[a,b] , Rob R. Meijer[a] , and Jorge N. Tendeiro[c]

[a]Department of Psychometrics and Statistics, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands; [b]Interdisciplinary Center Psychopathology and Emotion Regulation (ICPE), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [c]Office of Research and Academia-Government-Community Collaboration, Education and Research Center for Artificial Intelligence and Data Innovation, Hiroshima University, Japan

ABSTRACT

The accessibility to electronic devices and the novel statistical methodologies available have allowed researchers to comprehend psychological processes at the individual level. However, there are still great challenges to overcome as, in many cases, collected data are more complex than the available models are able to handle. For example, most methods assume that the variables in the time series are measured on an interval scale, which is not the case when Likert-scale items were used. Ignoring the scale of the variables can be problematic and bias the results. Additionally, most methods also assume that the time series are stationary, which is rarely the case. To tackle these disadvantages, we propose a model that combines the partial credit model (PCM) of the item response theory framework and the time-varying autoregressive model (TV-AR), which is a popular model used to study psychological dynamics. The proposed model is referred to as the time-varying dynamic partial credit model (TV-DPCM), which allows to appropriately analyze multivariate polytomous data and nonstationary time series. We test the performance and accuracy of the TV-DPCM in a simulation study. Lastly, by means of an example, we show how to fit the model to empirical data and interpret the results.

KEYWORDS

Item response theory; time series; psychological dynamics; non-linear trends; splines

Intensive longitudinal methods such as experience sampling or ecological momentary assessment have allowed researchers to study and unravel the psychological dynamics of individuals (Hamaker et al., 2015; Hamaker & Wichers, 2017). These methods consist of assessing individuals repeatedly during short periods of time. In particular, popular intensive longitudinal designs require participants to fill in short questionnaires of about 10 times a day for 5–7 days (Vachon et al., 2019). As a result, psychological time series commonly have between 50 and 100 time points. However, analyzing this kind of data has proven to be a challenging task.

Intensive longitudinal data are complex data with strong dependencies between the measurements due to their closeness in time. Because of this, researchers have applied extensions of the autoregressive model to analyze this kind of data (e.g., Asparouhov et al., 2018; Chatfield, 2003; Hamilton, 1994; Kuppens et al., 2010; Shumway & Stoffer, 2017; Song & Zhang, 2014; Walls & Schafer, 2006). The simplest autoregressive model used to analyze intensive longitudinal data is the autoregressive model of order 1 (AR(1); Chatfield, 2003; Hamilton, 1994), which regresses the dependent variable on a lagged version of itself to represent the relation between two consecutive observations of the dependent variable. This model has been extended, for example, to multilevel and multivariate settings (Bringmann et al., 2013), to account for measurement error (Schuurman & Hamaker, 2019; Schuurman et al., 2015; Song & Zhang, 2014), and to model unequally spaced measurements (i.e., continuous-time modeling, Crayen et al., 2017; Voelkle & Oud, 2013; Voelkle et al., 2012). Furthermore, a comprehensive framework to analyze intensive longitudinal data, known as dynamic structural equation modeling, was recently proposed by Asparouhov et al. (2018).

However, one of the shortcomings of these current methods is that most of these approaches require the

data to be continuous, which is not always the case. In particular, to study psychological dynamics, researchers tend to either use visual analogue scales or Likert scales (Vachon et al., 2019). While the former are continuous variables, which are suitable for the mentioned methods, the latter, strictly speaking, are ordinal categorical variables. This is a limitation, especially if there are not many response categories and if the distributions of the item responses are heavily skewed (Vogelsmeier et al., 2021). Furthermore, despite some few exceptions, most of the available statistical methods used to analyze intensive longitudinal data do not account for measurement error, which is likely to be present when measuring psychological constructs (Schuurman et al., 2015). On top of that, in many intensive longitudinal studies, multiple items are used to measure a unique construct such as positive or negative affect (e.g., Hamaker et al., 2018; Krieke et al., 2016) and composite scores are computed before fitting the model. However, ignoring the nature of the variables and the factor structure of the data might lead to biased estimates (Dolan, 1994; McNeish & Wolf, 2020). Hence, measurement models for categorical intensive longitudinal data are needed.

A useful statistical theory that can help to overcome these drawbacks is the item response theory framework (IRT; Embretson & Reise, 2013). In general, IRT models are latent variable measurement models that relate the categorical responses of a set of items to one or multiple latent continuous variables that represent unobservable psychological traits or ability levels (Hambleton & Swaminathan, 1985; Rijn et al., 2010) such as positive affect. Well-known IRT models are, for example, the Rasch model (von Davier, 2016) and the 2-parameter logistic model (van der Linden, 2016) for dichotomous responses, and the partial credit model (Masters, 2016) and the graded response model (Samejima, 1997) for ordered categorical responses. Additionally, IRT as a psychometric theory also allows taking an in-depth look at the quality of the psychological tests and measures. Within IRT, the standard error of measurement differs across scores depending on the characteristics of the items and the latent ability level of the participant (Embretson & Reise, 2013) and measurement precision can be determined conditional on the latent construct. This means that the quality of the measurements might vary across individuals, given their level on the latent construct.

Although IRT models have been largely developed within educational cross-sectional settings, dynamic IRT models for intensive longitudinal data have also been proposed in the recent years (e.g., Hecht et al., 2019; Kropko, 2013; Rijn et al., 2010; Wang et al., 2013). On the one hand, Rijn et al. (2010) proposed a Rasch model and partial credit model for intensive longitudinal data within the state space modeling framework, which is estimated by means of a Kalman Filter. On the other hand, the approaches by Kropko (2013, item response theory models for time series), Wang et al. (2013, dynamic Rasch model for educational data), and Hecht et al. (2019, continuous time Rasch model) are implemented within the Bayesian framework. The models proposed by Rijn et al. (2010) and Kropko (2013) are of special interest for us as they were developed to analyze psychological time series of one individual. However, these approaches are still limited as they (a) are not suitable for non-stationary time series, (b) have not been systematically tested in simulation studies, (c) lack user-friendly tutorials to be used by practitioners, and (d) do not use the core features of IRT modeling (e.g., item characteristic curves and item information functions) that allow assessing the quality of the scales.

In this article, we, therefore, propose the time-varying dynamic partial credit model (TV-DPCM), which is an item response theory (IRT) model suitable to analyze multivariate time series data of polytomous responses. With this new method, we aim to offer a flexible tool that allows modeling non-linear trends and studying the psychometric properties of the scales used in intensive longitudinal data studies. Also, to facilitate its use by practitioners, we share all the code needed to fit the model in the following git repository: https://github.com/secastroal/DIRT. In particular, the TV-DPCM is useful to analyze intensive longitudinal data of one individual, when a set of Likert scale items that measure the same construct are repeatedly used to measure one participant. The TV-DPCM extends the partial credit model (PCM; Masters, 2016) by assuming that the latent variable follows a time-varying autoregressive model (TV-AR; Bringmann et al., 2017).

The article is organized in the following sections. Firstly, we introduce the TV-DPCM in detail. This section also covers a brief introduction of the generalized additive model framework. Secondly, we conducted a "proof of concept" simulation to test the performance of the model under diverse conditions, while varying, for example, the number of time points and the size of the true autoregressive effect. Thirdly, we present an empirical application of the model to experience sampling data of self-esteem, which aims to exemplify how to use and interpret the results obtained by means of fitting the TV-DPCM. Lastly,

we discuss our findings and how the TV-DPCM can contribute to a better understanding of measurement in intensive longitudinal research. Moreover, we provide some ideas for future methodological research for intensive longitudinal data based on IRT.

The time-varying dynamic partial credit model

As mentioned before, the TV-DPCM integrates the partial credit model (PCM; Masters, 2016) and the time-varying autoregressive model (TV-AR; Bringmann et al., 2017). Briefly, the PCM is an IRT model for polytomous data, which can be seen as an extension of the Rasch model (Embretson & Reise, 2013; Masters, 2016; Ostini & Nering, 2006). This means that the PCM holds most of the assumptions and properties of the Rasch model such as the assumption of unidimensionality, local independence, and the separability of the person and the item parameters. On the other hand, the TV-AR is a dynamic model for non-stationary time series that models the parameters of the standard autoregressive model based on the generalized additive model (Bringmann et al., 2017; Wood, 2017). Within the TV-AR, both the intercept and the autoregressive effect are allowed to smoothly vary over time. In our implementation, we only allowed the intercept to smoothly vary over time. By combining these two approaches, we get the TV-DPCM, in which the measurement model is given by the PCM and the dynamic latent process is described by a TV-AR model.

The basis: the partial credit model

To start, we first introduce the PCM (Masters, 2016), which is an IRT model for polytomous items. The motivation to develop this model was to allow analyzing test items that required multiple sequential steps to find the correct answer, where partial credit is given for completing each of the steps (Embretson & Reise, 2013). Evidently, this model was proposed within an educational assessment context, however, it is also appropriate, and it has been widely used to analyze items with ordered response options as found in attitudes and personality tests (Embretson & Reise, 2013).

The PCM is commonly described as a "divide-by-total" (Thissen & Steinberg, 1986) or "direct" (Embretson & Reise, 2013) model because the probability to endorse a certain response option is directly defined as the ratio of the probability of that response option to the sum of the probabilities of all possible response options. Consider that we have a test with I Likert-scale items that is used to measure, for example, positive affect. The items are scored from 0 to m_i, with $i = 1, ..., I$; which means that item i has $K_i = m_i + 1$ response categories (items might differ in the number of response options). Then, the probability to select response option x of the i-th item given the latent trait of the j-th person, θ_j, can be written as:

$$P(X_i = x|\theta_j) = \frac{\exp\left[\sum_{k=0}^{x}(\theta_j - \delta_{ik})\right]}{\sum_{v=0}^{m_i} exp\left[\sum_{k=0}^{v}(\theta_j - \delta_{ik})\right]}, \qquad (1)$$

where δ_{ik} is the step parameter, also known as threshold parameter, of the k-th category of the i-th item. These threshold parameters δ_{ik} represent the level on the latent continuum at which the probabilities of selecting the response options k and $k-1$ are equal. An example of an item with five response options is presented in Figure 1. This shows how the probability of endorsing each response option depends on the level of the latent ability of the participant. Therefore, persons with lower levels of the latent trait are more likely to select the response option 0 of this item

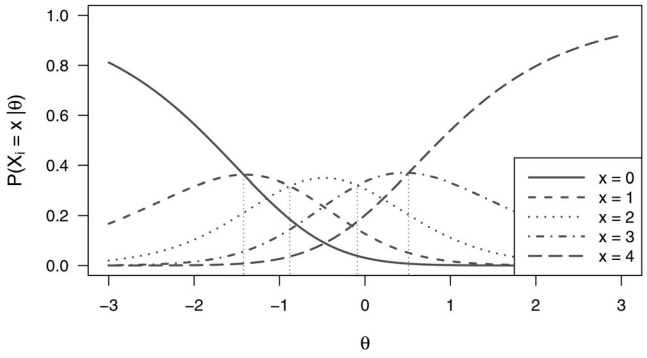

Figure 1. Item characteristic curves based on the PCM of an item with five response options and threshold parameters −1.42, −0.88, −0.09, and 0.51. The location of the threshold parameters is shown with the vertical dotted gray lines, which also correspond with the intersection between the curves of adjacent response options.

(when θ is lower than -1.42). Notice that for notational convenience, when $x = 0$, the summation in the numerator is defined as 0 so that the exponential evaluates to 1. Thus, when there are only two response options (correct or incorrect), the PCM simplifies into the Rasch model.

Moreover, it is important to highlight some of the assumptions and properties of the PCM. First, regarding the assumptions, in a similar way as the most widespread IRT models, the PCM assumes that unidimensionality and local independence hold (Embretson & Reise, 2013). Unidimensionality means that the model assumes that all the items in the test measure a unique latent construct (e.g., positive affect or neuroticism). On the other hand, the assumption of local independence implies that the responses to any pair of items are independent after controlling for the latent variable. Secondly, the PCM also keeps two important properties that are shared, in particular, with the Rasch model: The separability of the person and the item parameters and sufficient statistics. The former property means that each type of parameters can be conditioned out from the estimation of the other. The latter property means that the raw scores are sufficient statistics for the person parameters, so all persons with the same sum score are assumed to display the same value on the latent trait under study.

A straightforward extension: modeling a dynamic latent process

Now, in the context of studying psychological time series, a straightforward extension of the PCM model is to add an autoregressive structure at the latent level. This has been suggested by Rijn et al. (2010) within the state-space modeling framework and by Kropko (2013) within the Bayesian framework. However, to the best of our knowledge, in none of these studies nor in any other studies, the models have been systematically tested in a simulation study. In this article, we further extend this model (see following subsection) and assess its performance in a simulation study. Thus, the model changes as follows:

$$P(X_i = x|\theta_t) = \frac{\exp\left[\sum_{k=0}^{x}(\theta_t - \delta_{ik})\right]}{\sum_{v=0}^{m_i} exp\left[\sum_{k=0}^{v}(\theta_t - \delta_{ik})\right]}. \quad (2)$$

Notice, that the latent variable θ now has a subscript t, which indicates time. In this case, when there are repeated measurements from one individual, the latent variable does not represent the latent trait of a person but the latent state dispositions of the individual at each measurement occasion. In other words, the latent state disposition represents the attitude or emotion of the person in the situation where the measurement took place. Moreover, we assume that these latent state dispositions follow an autoregressive process of lag-order 1, which is:

$$\theta_t = \alpha + \varphi\theta_{t-1} + \varepsilon_t, \quad (3)$$

where α is the intercept of the process and φ is the autoregressive effect of lag-order 1 between consecutive measurement occasions. In particular, the lag-order indicates how many measurements in the past predict the current measurement. With a lag-order 1, only the immediately previous measurement is used to predict the current one. Moreover, the autoregressive effect represents the dependency between consecutive states. This effect is also known as the "inertia" parameter (Kuppens et al., 2010) because the larger this parameter is, the longer it takes the system to return to its equilibrium (i.e, its mean). Lastly, ε_t is the random innovation at time t. The innovations are the part of the current latent state that cannot be explained by the model. Yet, they still influence and are passed along to future states (Schuurman et al., 2015). The innovations are assumed to be normally distributed with mean 0 and variance Ψ.

By extending the PCM in this way, additional assumptions are made about the latent process. Firstly, this extension proposes a discrete-time model for the latent process. This means that the repeated measurements are assumed to be observed in equally spaced time intervals. If this condition is not satisfied, the autoregressive effect might be overestimated and lead to the wrong conclusions (Haan-Rietdijk et al., 2017). Secondly, the latent process is assumed to be stationary, which means that its means and its variances-covariances do not change over time (Chatfield, 2003). A necessary but not sufficient condition for stationarity in the autoregressive process in Equation 3 is that $|\varphi| < 1$. Lastly, it is assumed that item parameters (δ_{ik}) are also time invariant. In other words, in is assumed that longitudinal measurement invariance (Meredith, 1993; Meredith & Teresi, 2006) holds.

Dealing with change: the TV-DPCM

However, assuming stationarity might not be realistic in clinical practice. For example, consider a person that is under psychological treatment and fills in a daily diary questionnaire with Likert-scale items during the whole intervention. If the purpose is to monitor relevant psychological constructs for the intervention such as positive or negative affect, and if the intervention is effective, then, we would expect to

observe durable changes on the person's behavior and feelings (e.g., reduction of symptoms or increase in well-being). To allow for such change, we further extended the PCM to allow the latent dynamic process to be non-stationary. We called this extension the TV-DPCM, which aims to model the non-linear change of the latent variable while accounting for the measurement error of the psychological construct.

As with the previous extension, the TV-DPCM is described in two equations: The measurement equation and the structural equation. The measurement equation is the same as Equation 2. This equation models the relation between the observed responses and the latent construct based on the PCM. Then, the structural equation, which describes the latent dynamic process, is an extension of Equation 3 based on the TV-AR model (Bringmann et al., 2017). In this case, *only* the intercept α is allowed to vary over time[1]. To put it differently, with a time-varying intercept, the TV-DPCM is able to model latent processes that are trend-stationary (i.e., the time series is stationary after detrending). Now, the structural equation is defined like this:

$$\theta_t = \alpha_t + \varphi\theta_{t-1} + \varepsilon_t, \tag{4}$$

where α has a subscript t, which indicates that the intercept changes over time. This change is assumed to be described by a smooth function (see the following section).

Moreover, based on the time-varying intercept and the autoregressive effect, it is possible to derive the model-implied mean and variance of the dynamic process in Equation 4 (Bringmann et al., 2017; Chatfield, 2003; Giraitis et al., 2014). Firstly, in a TV-AR, the intercept does not have a clear interpretation and what describes the trend of the time series is, in fact, the mean of the dynamic process. Because the intercept varies over time, the mean of the dynamic process also varies over time. Therefore, the mean of the dynamic process at time t can be defined as (see Bringmann et al., 2017):

$$\mu_t \approx \frac{\alpha_t}{1-\varphi}. \tag{5}$$

Notice that the approximation in Equation 5 applies as long as the change of the intercept is constrained to be gradual[2] (i.e., smooth). The time-varying mean is also known as the attractor (Giraitis et al., 2014). Furthermore, we can also derive the variance of the dynamic process. Because the autoregressive effect is time-invariant, then, the variance of the dynamic process is also assumed to be time-invariant, and it is shown to be as follows:

$$\sigma^2 \equiv \frac{\Psi}{1-\varphi^2}. \tag{6}$$

To summarize, herewith, we propose the TV-DPCM, which is a measurement model useful to analyze psychological time series of one individual. The model keeps most of the assumptions of the PCM and the TV-AR such as (a) unidimensionality, (b) local independence, (c) trend-stationarity of the latent process, and (d) equally spaced observations over time. In contrast, the separability of item and person parameters and the sufficient statistics property, which are properties of the PCM, do not hold for the TV-DPCM. Furthermore, it is important to highlight that the TV-DPCM also has some similarities with time-varying effects models (TVEM; Dziak et al., 2014; Tan et al., 2012). More specifically, the TV-DPCM can be seen as an ordinal TVEM for one individual.

Estimation: generalized additive models and Bayesian inference

As with other dynamic IRT models (Hecht et al., 2019; Kropko, 2013; Wang et al., 2013), we implemented the TV-DPCM within the Bayesian framework. This allows estimating all the parameters simultaneously and prior information can be incorporated. Additionally, to estimate the time-varying intercept, we make use of the generalized additive model (Wood, 2017). In what follows, we first do a brief introduction of the generalized additive model (GAM) framework and then we mention the suggested priors required to estimate the model.

Generalized additive models are flexible semiparametric models that define the relation between the dependent variable and the covariates based on "smooth functions" (Wood, 2017). They are specially useful to model nonlinear relationships while keeping a reasonable predictive power. A general representation of a GAM model, given one dependent variable and one covariate is:

[1]Ideally, both the intercept and the autoregressive effect should be allowed to vary over time, as proposed in the TV-AR model (Bringmann et al., 2017). By doing this, the model can handle different types of non-stationarity, where the means, the variances, and the autocorrelations change. However, we did not succeed on writing a working TV-DPCM model in Stan that also allowed the autoregressive effect to vary over time. Because of this, we settled with the simpler version in which only the intercept is allowed to vary over time.

[2]The change of the intercept is required to be gradual because an assumption used to derive Equation 5 is that μ_t must be approximately equal to μ_{t-1}. This is also why, in Equation 5, the approximation sign is used instead of the equal sign.

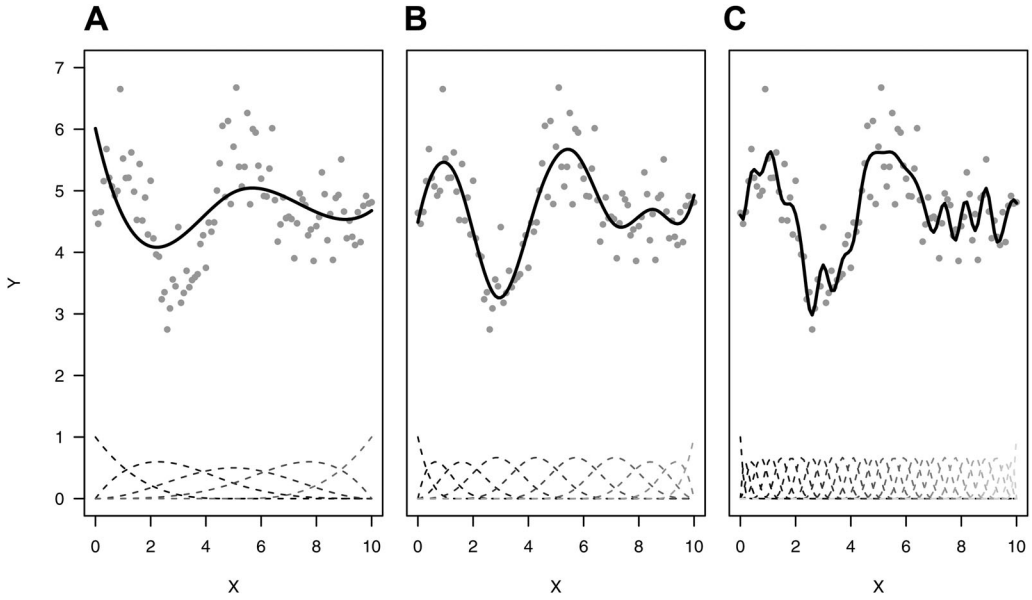

Figure 2. Predicted B-splines with different number of basis functions. The gray dots represent the observed data and the black line represents the predicted non-linear function. At the bottom of each plot, the basis B-splines functions are represented with dashed lines. The number of basis functions are 5 (A), 10 (B), and 30 (C).

$$y_i = f(x_i) + \epsilon_i, \tag{7}$$

where y_i is the dependent variable, x_i is a covariate, $f()$ is a smooth function, and ϵ_i are the independent and normally distributed random errors.

The smooth function is usually the weighted sum of some predefined "basis functions" and is represented as a linear model, as follows:

$$f(x) = \sum_{j=1}^{s} \beta_j b_j(x), \tag{8}$$

where $b_j()$, with $j = 1, ..., s$, is the j-th basis function, and β_j is the unknown weight for each function. Given this, in the TV-DPCM, the time-varying intercept α_t is modeled as a smooth function of time:

$$\alpha_t = f(t) = \sum_{j=1}^{s} \beta_j b_j(t). \tag{9}$$

However, when using the GAM, one must decide on the type of smoother that is going to be used and how smooth the resulting fit has to be. In our implementation, we opted to use cubic B-splines (Kharratzadeh, 2017; Wood, 2017).[3] Without going into too much detail, the basis splines or B-splines are a popular smoother in the GAM literature for univariate analysis. B-splines have a polynomial degree p (order of the B-splines is $p+1$) and a set of q knots that are typically defined based on the percentiles of the predictor variable. Then, these knots are used to define $q+p-1$ basis functions for the B-splines. Each basis function consists of $p+1$ pieces of polynomials (except for the ones close to the borders), that are joined continuously at p interior knots and are differentiable $p-1$ times. For the remaining range of the covariate, the basis functions are 0. Most commonly, B-splines of order 4 (i.e., degree $p=3$), which are cubic B-splines, are used. To illustrate this, we simulated data based on cubic B-splines with 10 basis functions as shown in the middle panel of Figure 2. The 10 basis functions are depicted at the bottom of the graph. When these functions are weighted by the β_j coefficients and summed together, they result in the nonlinear trend (solid black line) that describes the data.

Figure 2 also shows what can happen when too little or too many basis functions are used. Panel A presents the results from a cubic B-splines with 5 basis functions and panel C presents the results from a cubic B-splines with 30 basis functions. While using too little basis functions can result in underfitting, using too many can result in overfitting the data. Because of this, when using GAM, researchers usually use a larger number of basis functions than they would think are needed but impose a penalization on the selected smoother (Bringmann et al., 2017; Wood, 2017). For our implementation, to penalized the cubic B-splines, we used a random-walk prior for the β_j coefficients (Kharratzadeh, 2017). This means:

[3]We also wrote an alternative version of the model in JAGS (Depaoli et al., 2016), which can use other kind of smoothers such as thin plate or penalized P-splines based on the *mgcv* (Wood, 2017) package.

$$\beta_1 \sim \mathcal{N}(0,1), \qquad \beta_j \sim \mathcal{N}(\beta_{j-1},\tau), \qquad \tau \sim \mathcal{N}(0,1), \tag{10}$$

where τ is the smoothness hyperparameter. The reasoning of Kharratzadeh (2017) to use this prior is based on the fact that the sum of the basis functions ($b_j(t)$) is equal to 1. If all the β_j coefficients are equal, then, the resulting B-spline is a constant function of value β_j. Therefore, the closer the β_j coefficients are to each other, the smoother the spline function is.

Lastly, to estimate the TV-DPCM within the Bayesian framework, we used relatively informative prior distributions for the different parameters. The following priors were used in both the simulation study and the empirical application. Starting with the threshold parameters δ_{ik}, we used, as it is common in the IRT literature, a standard normal prior (Fox, 2010). For the random innovations ε_t, we first sampled a starting value from the standard normal, which was later scaled in the computation of θ_t given Equation 4. Then, the prior for the scaling factor of the innovations (i.e., the standard deviation σ) was a normal distribution with mean 1 and standard deviation 1, which was truncated to be positive. Finally, for the autoregressive effect φ, we used an uniform distribution between −1 and 1 as prior.

Simulation study

In this section, we present the design and results of the simulation study that we conducted with the TV-DPCM. The purpose of this simulation was to assess the performance, in terms of convergence and recovery of the population parameters, of the TV-DPCM under common settings seen in the literature.

Data simulation and design

Data were simulated based on the TV-DPCM model assuming that the time varying intercepts α_t followed a sinusoidal trend. An example of the simulated latent dynamic process and its trend is presented in Figure 3. The same trend was used for all the conditions but it was adjusted to the length of the time series. Moreover, we also kept the variance of the innovations fixed (at 1) and the number of response categories per item (5) equal across all conditions. Regarding the threshold parameters, these were randomly generated in such a way that they were ordered within an item. For example, the threshold parameters for an item with 5 response options in the simulation could be: −1.42, −0.88, −0.09, and 0.51 (recall Figure 1 and that the threshold parameters represent where in the latent continuum the item characteristic curves of adjacent response options intersect). Lastly, the latent state disposition of the first measurement occasion was randomly generated for each replication and each condition from a normal distribution with mean $\alpha_1/(1-\varphi)$ and standard deviation $\sqrt{\Psi/(1-\varphi^2)}$. For details on the generation of these parameters, see the code shared on the GitHub repository of this manuscript.

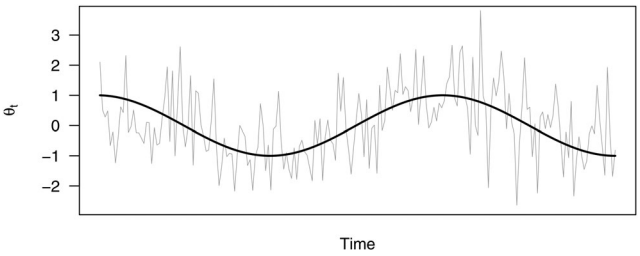

Figure 3. True latent dynamic process (in gray) and its trend (in black) of simulated data.

Next, for the simulation design, we manipulated four factors. Firstly, the number of time points were varied between 100, 200, 300, and 500. These number of time points were chosen based on previous simulations with $N = 1$ time series (Bringmann et al., 2017; Schuurman et al., 2015). In fact, based on preliminary simulations with the TV-DPCM, we do not expect the model to perform well with under 200 time points. Secondly, the number of items was either 3 or 6 items. The reason for this is that scales used in ESM studies tend to be short in order to reduce participants' burden. Next, the size of the autoregressive effect was varied between 0, 0.25, and 0.5, which is similar to the values used in simulations with the VAR model with measurement error (Schuurman & Hamaker, 2019; Schuurman et al., 2015). Lastly, the proportion of missing observations was either 0% or 30%. To recreate the missing data patterns that are commonly seen in ESM data, where participants either fill in the complete questionnaire or do not fill it in at all, we randomly sampled 30% of the time points and removed all the observations in those time points. Basically, the simulated missing data mechanism was missing completely at random with the constraint that the observation of the first time point was never removed. The conditions with missing data aimed to test the model under realistic circumstances, as the percentage of missing measurements usually ranges between 20% and 40% (Vachon et al., 2019). To summarize, the simulation had a $4 \times 2 \times 3 \times 2$ fully crossed design, in which we ran 200 replications per condition (i.e., a total of 9,600 analyses).

The models were estimated within a Bayesian framework through the Hamiltonian Monte Carlo algorithm as implemented in Stan (Carpenter et al., 2017). We ran three chains per analysis, each with 2,000 iterations, 500 of which were used for warm-up[4]. To run the analyses, we also adjusted other parameters of the Hamiltonian Monte Carlo algorithm such as the *delta* and the *maximum treedepth* (Stan Development Team, 2022). We increased parameter *delta* from 0.8 (default) to 0.99 and the *maximum treedepth* from 10 (default) to 15, as this was required to facilitate model convergence.

The simulation of the data, the estimation of the model, and the analysis of the results were performed in R (R Core Team, 2022) with the R packages: rstan (Stan Development Team, 2020) and bayesplot (Gabry & Mahr, 2021). Analyses were run on a high performance computing cluster with Intel Xeon E5 2680v3 CPU (2.5 GHz). The maximum RAM usage for an analysis was approximately 500MB.

Output variables

To assess the performance of the TV-DPCM, we focused on the convergence of the model and the quality of the estimates. In relation to model convergence, we relied on the convergence checks provided in Stan for the Hamiltonian Monte Carlo algorithm. According to these checks, an analysis diverged if the Gelman-Rubin statistic ($\hat{R}$; Gelman & Rubin, 1992) for any of the parameters was larger than 1.05, if there was any divergent transition after warm-up (Stan Development Team, 2022), or if any Bayesian Fraction of Missing Information (BFMI; Betancourt, 2017) was too low. Stan also provides other diagnostic checks about the efficiency of the algorithm that indicate if the maximum tree depth was exceeded or if the effective sample sizes (ESS) were too low[5]. While the latter checks were tracked, no action was taken if, for example, the ESS of an analysis was too low, as these problems do not jeopardize the quality of the estimates and they are usually solved by increasing the number of iterations.

To assess the quality of the estimates, we looked at different accuracy statistics such as bias, absolute bias (abbias), relative bias (rbias), and root mean squared error (RMSE). Suppose that we focus on the set of parameters Θ (e.g., the thresholds, the latent states, or the autoregressive effect) and we run a simulation with M replications per condition. Given a condition c where there are N_c parameters Θ_n with $n = 1, \ldots, N_c$, and their estimates for the m-th replication are $\hat{\Theta}_{nm}$, with $m = 1, \ldots, M$, then, these accuracy statistics are defined as follows:

$$bias = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N_c}\sum_{n=1}^{N_c}(\hat{\Theta}_{nm} - \Theta_n)\right], \tag{11}$$

$$abbias = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N_c}\sum_{n=1}^{N_c}|\hat{\Theta}_{nm} - \Theta_n|\right], \tag{12}$$

$$rbias = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N_c}\sum_{n=1}^{N_c}\frac{\hat{\Theta}_{nm} - \Theta_n}{\Theta}\right], \tag{13}$$

$$RMSE = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{N_c}\sum_{n=1}^{N_c}(\hat{\Theta}_{nm} - \Theta_n)^2}. \tag{14}$$

For parameters such as the item thresholds, the latent state dispositions, and the attractor, we did not compute the relative bias because some of the true values of these parameters were 0 or very close to 0. As a result, the computed relative bias reached infinity or was extremely large, which made the measure unusable. Hence, for these parameters we computed the correlation between the true and the estimated parameters as well as the RMSE. In contrast, with parameters such as the autoregressive effect and the innovation variance, it was possible to compute the relative bias in most of the conditions. Additionally, we also inspected the coverage proportion of the credibility intervals as well as their average width for all the parameters.

Results

In total, 148 analyses of the 9,600 diverged. All the divergent analyses were due to the presence of divergent transitions after warm-up (as indicated by the convergence checks in Stan). Figure 4 presents the percentage of convergent replications per condition. This shows that most of the divergences occurred in the conditions with 100 time points and when the true autoregressive effect was the largest. These results indicate that, in general, at least 200 time points seem to be required to fit the TV-DPCM.

[4]We conducted preliminary simulations analyses with the model to ascertain that this number of total and warm-up iterations was enough to obtain reliable samples from the posterior distributions.

[5]As suggested by an anonymous reviewer, we rerun a small part of the simulation to save the typical bulk and tail effective sample sizes estimated for fitting the TV-DPCM with the MCMC setup used for the simulation and empirical example of this study. Plots that summarize the computed bulk and tail effective sample sizes are presented on Figures S7 through S9 of the supplementary material.

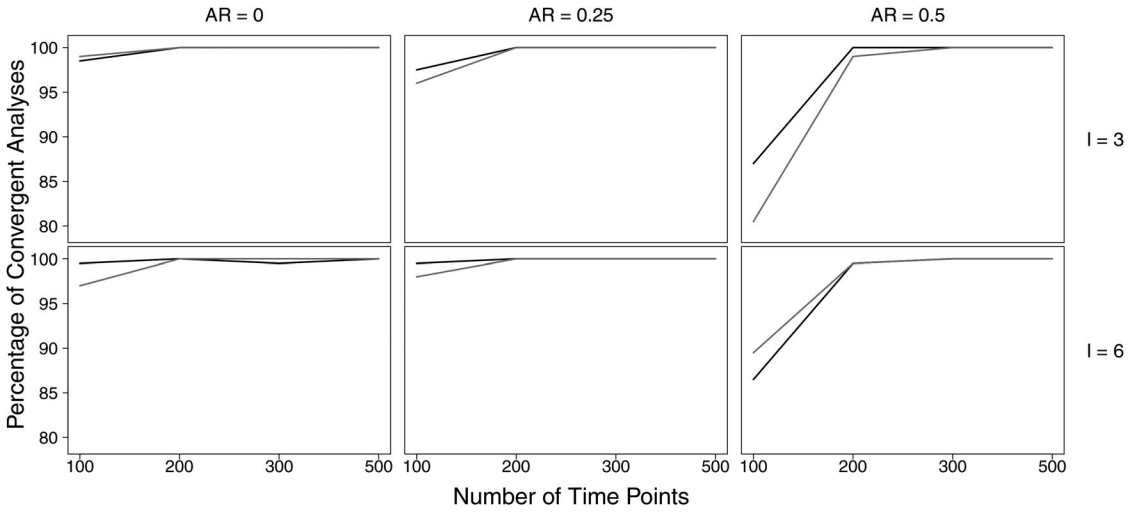

Figure 4. Percentage of analyses that converged per condition.

Next, to assess the quality of the estimates of the relevant parameters such as the thresholds, the latent state dispositions, and the autoregressive effect, we looked at the different accuracy statistics per each set of parameters. Hereby, we present in detail the results from the threshold parameters and the autoregressive effect. For the other parameters, we summarize the main findings and add supporting Figures in the supplementary materials. Figure 5 shows the average coverage proportions, the average width of the credibility intervals, the average correlation, and the average bias across conditions for the threshold parameters. The intervals around these averages indicate the interquartile range of the measure over the 200 replications per condition. Starting with the width of the credibility intervals across conditions, panel B of Figure 5 shows the credibility intervals shrank when there were more time points. This was expected, as usually with IRT models, the estimation of the item parameters improves when the number of participants (time points in the TV-DPCM) increases and vice versa. Secondly, regarding the coverage proportion, the panel A shows that on average 80% of the credibility intervals included the true parameter. It seemed that the average coverage was slightly lower and spread more when there were more time points. This can be explained by the fact that for some analyses, there were large biases and all the threshold parameters were completely over- or underestimated. This in combination with narrow credibility intervals resulted in lower coverage rates. Thirdly, the correlation between the true and the estimated thresholds (panel C) was on average above 0.9 across all conditions and it approached 1 when the number of time points increased. The presence of missing data worsened the correlation between the true and the estimated thresholds in relation to the conditions without missing values. However, these differences became smaller as the number of time points increased. Lastly, the average bias of the threshold parameters was close to 0 across all conditions (panel D). Similar figures for the absolute bias and RMSE of the threshold parameters are included in the supplementary material, which show that these measures became smaller as the number of time points increased. Overall, the accuracy of the estimates of the threshold parameters improves, as evidenced with the average correlation and average width of the credibility intervals, when the number of time points is larger than 200.

Regarding the autoregressive effect, Figure 6 presents the average coverage proportion, credibility interval width, absolute bias, and relative bias of this parameter. On average, the coverage proportion of the autoregressive effect was close to 100% across all conditions (panel A). In contrast, the width of the credibility interval clearly depended on the number of time points and the percentage of missing values as shown in panel B. Furthermore, we present the absolute bias instead of the correlation, as computing the correlation was not adequate or informative. Panel C shows that the average absolute bias range between 0.15 and 0.05 across conditions, decreasing when the number of time points increased or when there were no missing values. Regarding the relative bias (panel D), the average is only presented for the conditions when the true autoregressive effect was different from 0. In general, the relative bias of the autoregressive effect was on average 0 across all conditions. However, when there were 100 time points, 3 items, no missing values, and an autoregressive effect of

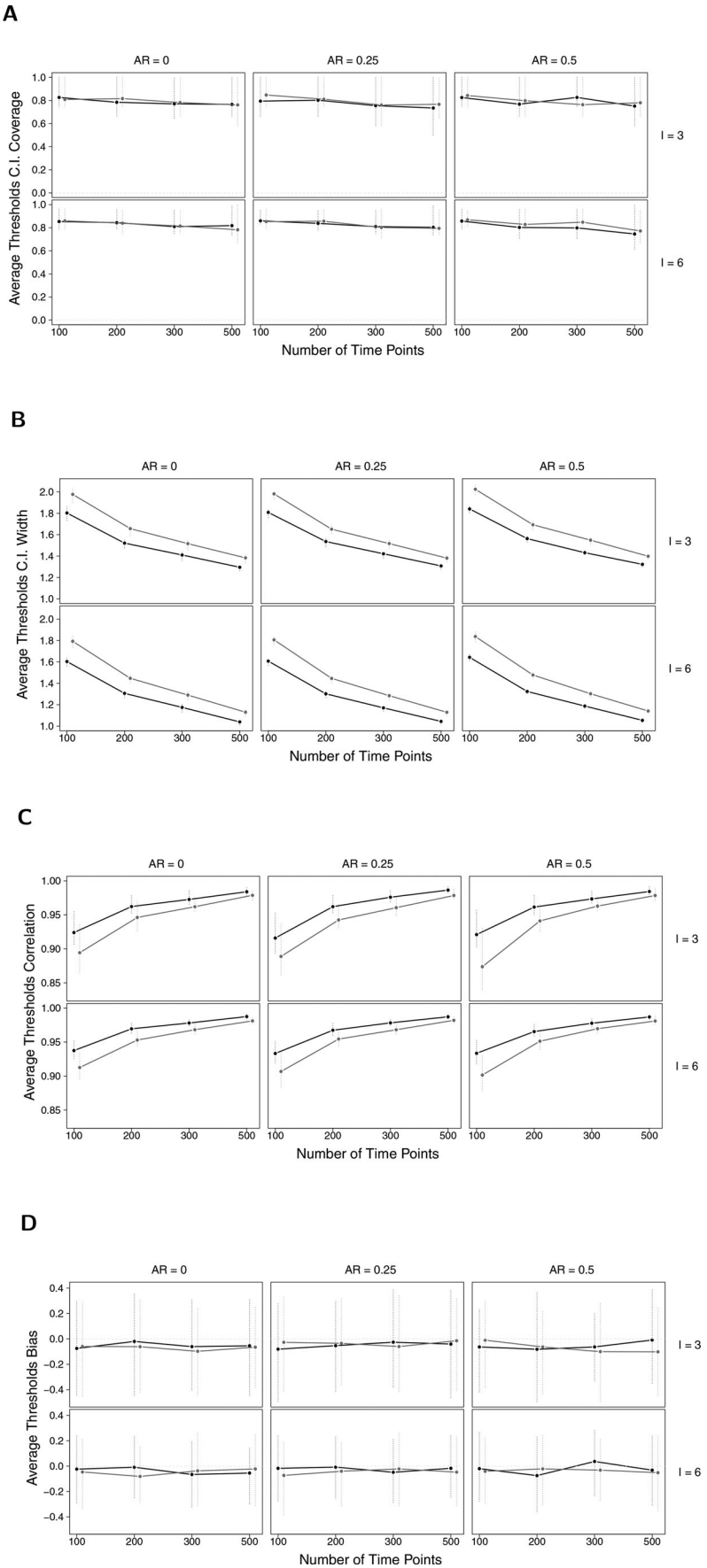

Figure 5. Parameter recovery and accuracy statistics of the threshold parameters. The black lines represent the conditions where there were no missing data and the gray lines represent the conditions with 30% missing values. The vertical dotted lines around the dots represent the interquartile range per condition. (A) Coverage proportion, (B) average width of the credibility interval, (C) average correlation between the true and the estimated thresholds, and (D) average bias per condition.

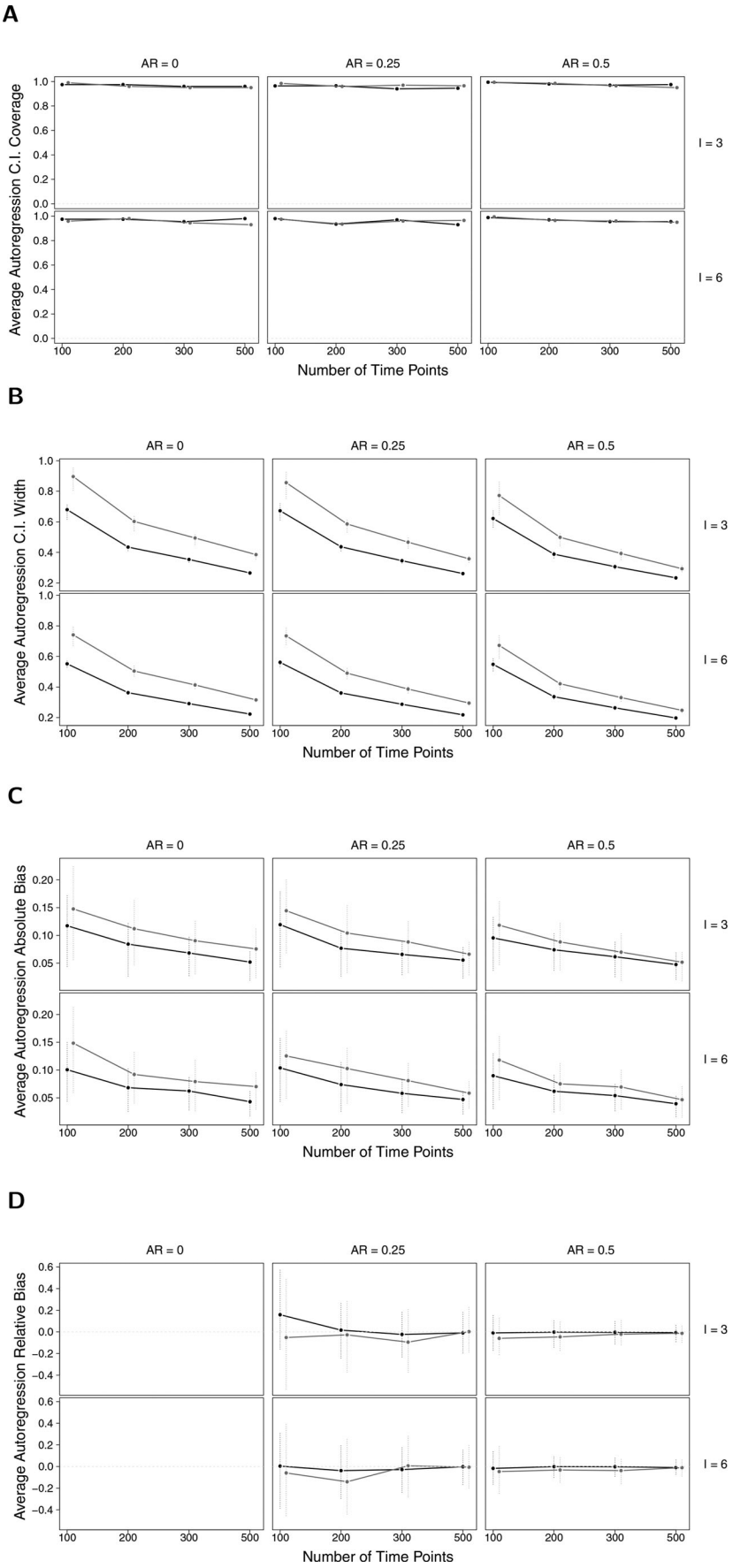

Figure 6. Parameter recovery and accuracy statistics of the autoregressive effect. The black lines represent the conditions where there were no missing data and the gray lines represent the conditions with 30% missing values. The vertical dotted lines around the dots represent the interquartile range per condition. (A) Coverage proportion, (B) average width of the credibility interval, (C) average absolute bias, and (D) average relative bias per condition.

Table 1. Summary of the recovery of the other parameters of the TV-DPCM.

Parameter	Results summary
Latent state dispositions	The width of the credibility intervals shrank when the number of items increased. Also, the average correlation increased when the number of items increased, when the size of the autoregressive effect was larger, and when there were no missing values.
Attractor	Similarly as with the threshold parameters, increasing the number of time points resulted in a slight decrease of the coverage rate, a shrinkage of the credibility intervals, and a raise of the average correlation.
Variance of the innovations and variance of the dynamic process	Just as with the autoregressive effect, the width of the credibility intervals and the average absolute bias decreased when the number of time points increased. However, these parameters tend to be underestimated as their estimates were between 20% (conditions with 100 time points) to 10% (conditions with 500 time points) lower than the true parameter according to the relative bias.

0.25, the TV-DPCM tended to overestimate the autoregressive effect about 20% above its true value.

In relation to the other parameters of interest, such as the latent states, the attractor, the variance of the innovations, and the total variance of the dynamic process, we briefly summarize the findings about the recovery of these parameters in Table 1. (Figures for these parameters are included in the supplementary material.) In general, the average coverage rate per condition of these parameters was between 80% and 90%, and the coverage percentage did not seem to be influenced by the manipulated factors. Moreover, the recovery of the attractor showed similar results as the ones seen for the threshold parameters. Similarly, the results of the variance of the innovations and the variance of the dynamic process in relation to the width of the credibility intervals and the mean absolute bias were similar to the results observed for the autoregressive effect.

Summary

To conclude, this simulation study showed that the TV-DPCM performs well at recovering its parameters across most of the conditions. In general, the accuracy of the estimates of the TV-DPCM improves when the number of time points increases. The results suggest that at least 200 time points are required for the model to converge and to accurately estimate the parameters. However, given the width of the credibility intervals of some parameters, we actually suggest 300 time points as a minimum to estimate the TV-DPCM. Still, in some cases, the model might over- or underestimate some of the parameters of interest. While in such cases the estimates are biased and the coverage of the credibility intervals is poor, the overall pattern is still well recovered as indicated by the high correlations. Regarding the number of items, it seems the TV-DPCM can be estimated with as little as 3 items, with the caveat that the credibility intervals can be very wide, specially in combination with the presence of missing data.

Empirical example: using the TV-DPCM to analyze self-esteem

To exemplify how to use and interpret the results of the TV-DPCM, in this section, we analyzed mood data from one participant. These data, collected between August 2012 and April 2013, were retrieved from Kossakowski et al. (2017) and were previously analyzed by, among others, Wichers and Groot (2016). The data come from a 57 year-old male (at the time of the study) who had been diagnosed with major depressive disorder. The participant completed up to 10 semi-random assessments per day for 239 days. During this period, the participant also followed a blind gradual reduction of their anti-depressant medication dosage. In what follows, the items of interest and the data collection procedure are described in detail. Then, the data pre-processing procedures are presented. Finally, the TV-DPCM is adjusted to the data in order to study the psychological dynamics of self-esteem and the performances of the items of the ESM questionnaire.

Data collection and procedure

As mentioned before, the participant filled in an ESM questionnaire up to 10 times a day for 239 days. The questionnaire was programmed at random moments within 90-minute intervals that were set between 07:30 AM and 10:30 PM. After the beep signal, the participant had a 10-minute window to complete the questionnaire, which consisted of 50 momentary assessment items that measured different emotions (e.g., feeling enthusiastic or feeling lonely), self-esteem, and descriptions of the situation such as

whether the participant was alone or doing something. Furthermore, additional items were used at certain beep signals to measure, for example, sleep quality and depressive symptoms. These items were filled up on a daily or weekly basis. All the momentary assessment items were measured on a 7-point Likert scale from "not feeling the state" to "feeling the state very much". The participant completed a total of 1473 assessments (i.e., on average 6.2 assessments per day). Moreover, the study was divided in 5 phases (Kossakowski et al., 2017): (1) A baseline period of four weeks, (2) a double-blind period without dosage reduction of two weeks, (3) a double blind period with gradual dosage reduction of eight weeks, (4) a post-assessment period of 8 wk, and (5) a follow-up period of twelve weeks. For this empirical example, we fitted the TV-DPCM to the items of self-esteem[6] during phases 1 and 2 (286 complete beeps).

Data pre-processing

The items of self-esteem were *I like myself* (Self-like), *I am ashamed of myself* (Ashamed), and *I doubt myself* (Self-Doubtful)[7]. The items *Ashamed* and *Self-Doubtful* were reverse-coded to have high scores on the scale represent high levels of self-esteem. Also, given that not all the response categories were selected and that some were selected too few times, several response categories were collapsed. For the item *Self-like*, the response categories lower than 3 and the response categories larger than 5 were collapsed and recoded into response categories 1 and 3, respectively. Also, response category 4 was recoded as 2. For the items *Ashamed* and *Self-Doubtful*, response categories lower than 5 (after reversed coding) were collapsed into response category 1 and response categories 6 and 7 were recoded to 2 and 3, respectively. Therefore, the responses were changed from a 7-point Likert scale to a 3-point Likert scale.

Moreover, it is important to note that the TV-DPCM is a discrete time model. This means that the model requires that the time interval between consecutive observations is the same for the whole duration of the data collection. This was clearly not the case with the data at hand due to the random beeps, the missing data, and the overnight time between days. One way to address this issue within the Bayesian framework is to include missing values in order to make the time intervals between observations approximately the same (Asparouhov et al., 2018). This approach has been shown to be useful to deal with unequal time intervals and the results from these kind of analyses are comparable with results from continuous time models (Haan-Rietdijk et al., 2017). Given this, we also implemented this approach in the TV-DPCM analysis of the self-esteem items. For this, we divided the days in 90-minute time windows. As a result, there is a total of 16 time windows per day, 6 of which were always missing because they happened during the night. Observations within any of these time windows were considered as a representation of the state of self-esteem of the participant for that time point. When no observations were available, "missing values" were included in the date set. By doing this, we added 380 rows of missing values, for a total number of 666 time windows. The time series of the observed mean scores, after recoding and after including rows of missing values, is presented in Figure 7. The mean scores ranged between 1 and 3.

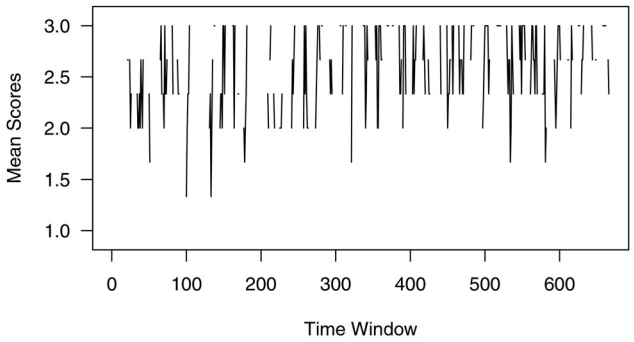

Figure 7. Observed mean scores of the self-esteem items.

Fitting the TV-DPCM

To fit the TV-DPCM to the data, we used the same setup for the Hamiltonian Monte Carlo algorithm as we did in the simulation study. This means that we ran three chains in parallel, each with 2000 iterations, 500 of which were discarded as warm-up, and we kept the same values for the *adapt_delta* (0.99) and *max_treedepth* (15) parameters. To check convergence of the model, we examined the diagnostics provided in Stan for the HMC algorithm. According to these

[6]We also fitted the TV-DPCM to the items of positive affect, negative affect, and mental unrest, which were other psychological constructs of interest in Wichers and Groot (2016). In the end, we deemed that the analysis of the items of self-esteem was more appropriate to illustrate the model because with this set of items the model fitted best. To assess the model fit, we used posterior predictive model checking methods (PPMC) for the TV-DPCM that we were developing simultaneously. This showed that the best fitting model was when analyzing the items of self-esteem during phases 1 and 2. Results of the fitting model to the items of self-esteem across all phases are reported in the supplementary material.

[7]A fourth item of self-esteem was *I can handle anything*. However, this item was excluded because the scale did not seem to be measuring an unidimensional construct when this item was included according to the preliminary version of the PPMC methods.

Table 2. Estimated parameters of the TV-DPCM.

	Median	SD	C.I.	ESS
$\hat{\delta}_{11}$	−2.55	0.55	(−3.63,−1.46)	4474
$\hat{\delta}_{12}$	1.46	0.44	(0.6,2.34)	3110
$\hat{\delta}_{21}$	−1.44	0.63	(−2.69,−0.21)	4784
$\hat{\delta}_{22}$	−1.60	0.48	(−2.55,−0.67)	3429
$\hat{\delta}_{31}$	0.55	0.44	(−0.32,1.41)	3041
$\hat{\delta}_{32}$	2.54	0.47	(1.63,3.48)	3177
$\hat{\varphi}$	0.47	0.12	(0.22,0.69)	758
$\hat{\Psi}$	1.90	0.56	(1.03,3.2)	1212
$\hat{\sigma}^2$	2.48	0.64	(1.47,4.02)	1401

Note. C.I. = 95% central credible interval.

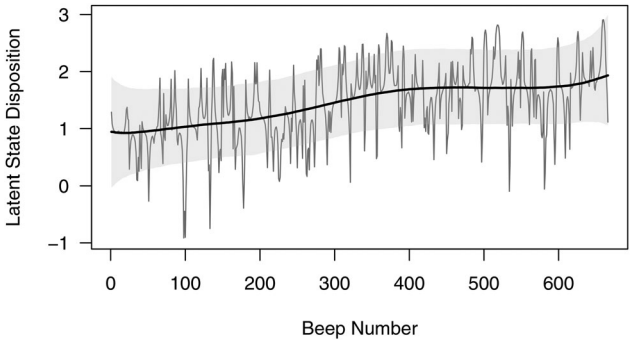

Figure 8. Estimated latent state dispositions for each beep (observed and missing) are represented with the gray line. The trend of the dynamic process or attractor is represented with the black line alongside with its 95% credibility interval band in light gray.

diagnostics, we found no evidence of divergence. Graphical diagnostics for some selected parameters are presented in the supplementary materials.

Table 2 shows the estimated values (i.e., the median of the posterior distribution), the standard deviation of the posterior distribution, the credibility interval, and the effective sample size of the threshold parameters, the autoregressive effect, the variance of the innovations, and the total variance of the dynamic process. Note that the threshold parameters are ordered within items 1 and 3 but not for item 2. This means that there is a "reversal" for item 2. Hence, the probability to select response category 2 is always lower than the probability to select either response category 1 or 3 across the latent continuum (see Figure 9). Next, the estimated autoregressive effect was 0.47, which implies that there is a medium-strong dependency between consecutive states of self-esteem. Thus, when the person experienced a high level of self-esteem at a certain time point, it was likely that they would keep experiencing high levels of self-esteem for the next measurement. Lastly, the variance of the innovations and the variance of the dynamic process were 1.9 and 2.48, respectively. The first one indicates the variability of the state of self-esteem that cannot be explained by the previous state of self-esteem. The latter represents the total variance of the states of self-esteem across the time series.

The estimates of the latent state dispositions and the time-varying attractor are presented in Figure 8. To facilitate the interpretation, these estimates were previously divided by the standard deviation of the dynamic process (i.e., $\hat{\sigma}$). By doing this, a latent state disposition of 1 means that the latent state of the individual at a certain time point is one standard deviation above the expected mean score on the test. Thus, Figure 8 shows that the latent state dispositions varied between −0.91 and 2.91. The time varying mean or attractor with its credibility interval band shows a slight increasing trend over time[8]. This implies that, on average, at the beginning of the study the mean of the latent states of self-esteem was about one standard deviation above the expected mean score of the questionnaire. Moreover, the mean of the latent states of self-esteem increased in such a way that by the end of the second phase, the mean of the latent states was close to two standard deviations above the expected mean score of the questionnaire.

Importantly, one of the key features of IRT modeling is that IRT models allow studying the properties of the items and the test. In this context, IRT provides the item characteristic functions (ICFs), the item information functions (IIFs), and the test information function (TIF). For the TV-DPCM, we can compute and plot these functions because the model assumes that the item parameters do not change over time (longitudinal measurement invariance holds). Therefore, these functions are defined given the latent state disposition (θ_t) at a certain time point t, namely the states of self-esteem of the individual. Figures 9–11 present the ICFs, the IIFs, and the TIF for the three items of self-esteem, respectively. Just as before, to facilitate the interpretation, the estimated latent state dispositions and the estimated thresholds were divided by the standard deviation of the dynamic process $\hat{\sigma}$. Regarding the ICFs, for the items *Self-like* and *Self-Doubtful*, the curves for each response category are nicely ordered and each of the response options gets to have the highest response probability at some point in the latent continuum. On the other hand, for the item *Ashamed* there is a reversal (i.e., the threshold parameters are not ordered for this item). As a consequence, there is no point on the latent continuum where the response category 2 has the highest response probability. Additionally, when inspecting

[8]This trend must be interpreted with caution given the width of the credibility intervals of the attractor parameter, which can also suggest that the real trend is stable.

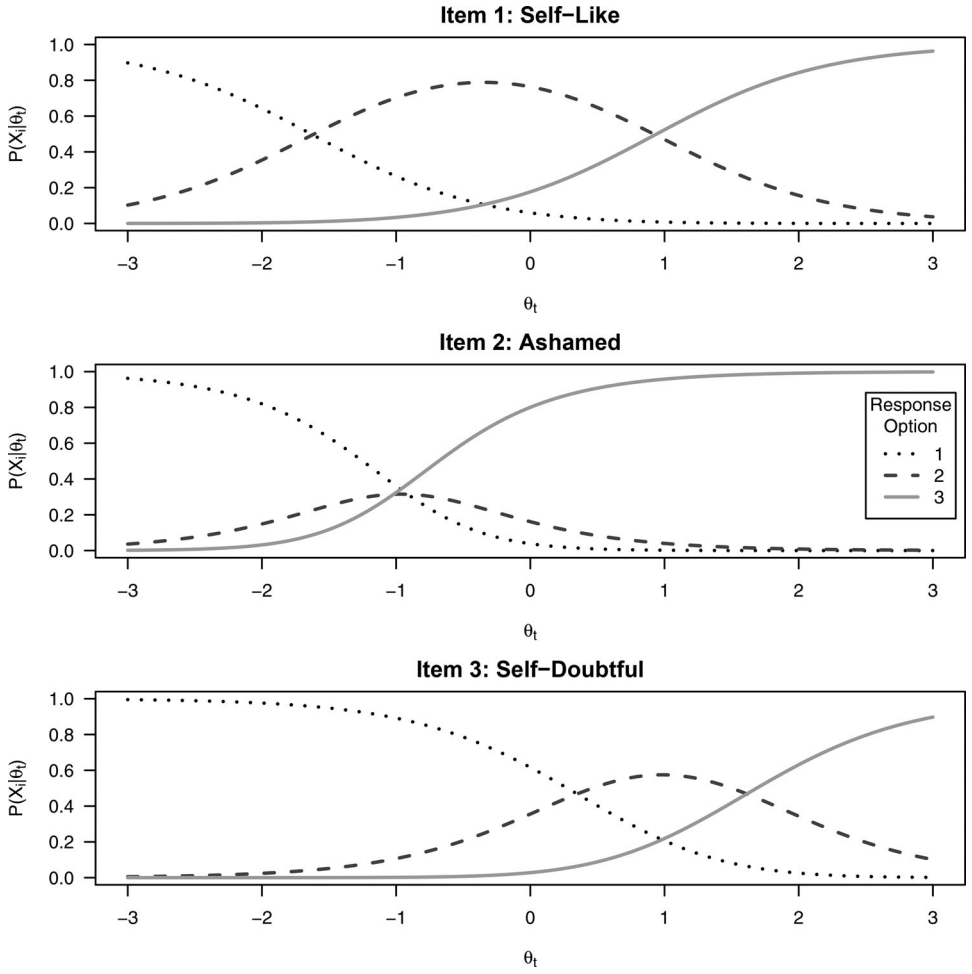

Figure 9. Item characteristic functions for the items of self-esteem.

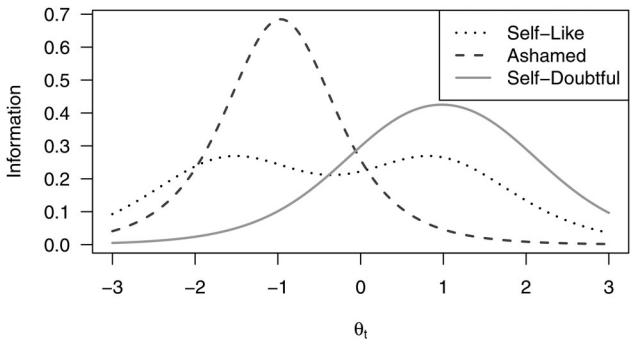

Figure 10. Item information functions of the items of self-esteem.

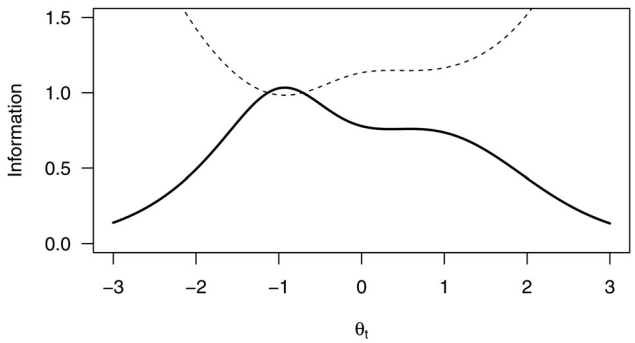

Figure 11. Test information function. The solid black line represents the test information function, which is the sum of the item information functions. The dashed line represents the standard measurement error of the scale given the level of the self-esteem.

the IIFs, we can see at what levels of the latent continuum the items are more or less informative. Thus, the item *Ashamed* seems to be more useful at measuring lower states of self-esteem and the item *Self-Doubtful* seems to be more useful at measuring higher states of self-esteem. In contrast, the item *Self-Like* is less informative than the other two items. Nonetheless, it seems it is useful to distinguish between very high and very low states of self-esteem but it is not informative in the middle levels of self-esteem. Lastly, the TIF shows that, overall, these three items are the most informative when measuring lower levels of self-esteem (solid line). However, during the study the participant mostly experienced medium and high levels of self-esteem, which means that their self-esteem was measured with high levels of standard

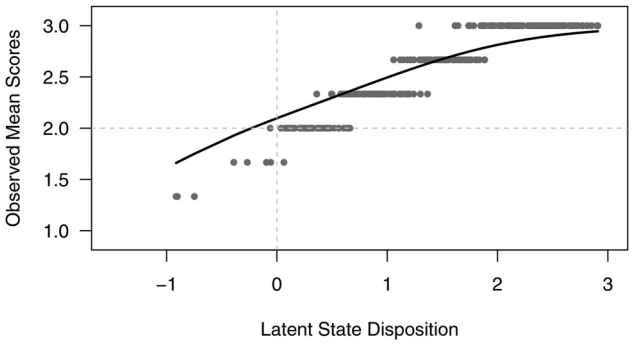

Figure 12. Comparison between the estimated latent state dispositions and the observed mean scores (gray dots). Also, the expected mean scores given the model in relation to the latent state dispositions are represented by the solid black line.

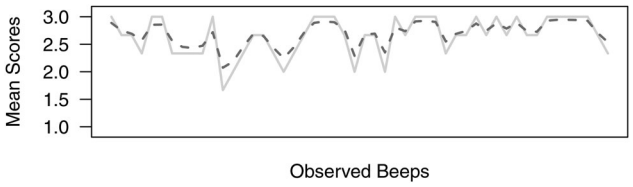

Figure 13. Comparison between the expected mean scores and the observed mean scores for the last 50 observations of phase 2. The observed mean scores are represented with the solid light gray line and the expected mean scores given the model are represented with the dashed gray line.

measurement error (dashed line). This indicates that more items would be needed to accurately measure the whole spectrum of the participant's self-esteem.

Finally, we also computed the expected mean scores given the model, which can be interpreted as estimates of the *true scores* (Embretson & Reise, 2013), to compare them with the observed mean scores. This is shown in Figures 12 and 13. Figure 12 shows the nonlinear relation between the estimated latent state dispositions and the observed mean scores. It also displays the expected mean scores given the model (black line) for the observed range of the latent state dispositions. This plot evidences that the observed mean scores are not sufficient statistics for the latent state dispositions. Moreover, Figure 13 shows the observed mean scores against the expected mean scores for the last 50 observed beeps. The trajectory of the observed mean scores is closely followed by the trajectory of the expected mean scores. This shows the high predictive value of the TV-DPCM.

Discussion

In this article, we presented an extension of the PCM to analyze psychological time series, namely, the TV-DPCM. This proposed model integrates the PCM (Masters, 2016) and the TV-AR (Bringmann et al., 2017) to allow studying the quality of the measures of psychological constructs when measured intensively on one participant. We tested the performance of the model in a simulation study while controlling for the number of time points, the number of items, the size of the autoregressive effect, and the presence of missing data. We also illustrated, by means of an empirical example, how to estimate the model and interpret its results. Overall, the TV-DPCM seems to be a promising tool to further understand how psychological measurement works on intensive longitudinal settings.

In general, the simulation study indicated that the model requires a large number of time points (more than 200) to converge and to deliver accurate estimates. This is in line with other results from simulation studies with autoregressive models for one individual (Bringmann et al., 2017; Schuurman et al., 2015). In fact, the TV-DPCM might also require more time points due to the increased complexity of the model. Given that the credibility intervals of the estimates tend to be wide even with 200 time points, we actually suggest 300 observed time points as a minimum to have less uncertainty about the results and we discourage researchers to use the model with less than 200 time points. Nonetheless, the TV-DPCM seems to be able to accurately recover its parameters across most of the tested conditions.

In relation to the empirical example, we showed how the TV-DPCM allows making a rich interpretation of the scales used in intensive longitudinal settings. By using all the features provided by the IRT framework while accounting for the time dependencies of the data, we were able to take a closer look at the properties of the items and the scale of self-esteem from the empirical data. The information provided by the ICCs, the IIFs, and the TIF allows assessing the quality of the items and the scale, which can give researchers the opportunity to make adjustments for future applications of their experience sampling questionnaires. In this particular example, we noticed that probably more items that measure medium and high levels of self-esteem were needed to reliably measure this individual's self-esteem. However, these results should be interpreted carefully as the proportion of missing values in the empirical data was considerably higher than the proportion studied in the simulation study. As shown in the simulation study, the presence of missing values have an effect on the width of the credibility intervals and the absolute bias of some parameters.

Even though the TV-DPCM can be a useful tool to gather relevant information about the measures in intensive longitudinal data, which can help to improve the scales used, the model still has its limitations. First, as shown in the empirical example, the model requires several steps of data manipulation such as reverse coding and collapsing response options. These steps are required to facilitate the interpretation of the latent variable and to be sure that thresholds are interpretable. Just as in the PCM, all responses options need to be observed to be able to estimate the parameters of the items in the TV-DPCM. If this is not the case, collapsing and recoding some response options becomes necessary, which reduces the variability of the observed data. While this might be a limitation of the model, it also represents a general challenge for researchers that are interested in studying psychological dynamics. This suggest that more research is needed in relation with the wording and the number of response options of the Likert-scale items used in intensive longitudinal settings. In other words, research focused on testing and improving the questionnaires used in intensive longitudinal settings are lacking.

Secondly, while the TV-DPCM is flexible enough to handle (non-linear) trend-stationary time series, which is a specific kind of non-stationarity, the model cannot handle other types of non-stationarity. Initially, a more flexible extension would be to allow the autoregressive effect to also smoothly vary over time (Bringmann et al., 2017). By doing this, the model would be able to handle time series with time varying variances and autocorrelations. While such an extension is the most reasonable step forward, its implementation is not necessarily straightforward. In spite of our efforts, we did not succeed in writing a working model with a time-varying autoregressive effect. One of the challenges that we faced when we tried was that we were not able to bound the B-spline of the autoregressive effect within -1 and 1. Similarly, the TV-DPCM assumes that longitudinal measurement invariance (Meredith, 1993; Meredith & Teresi, 2006) holds. This assumption implies that the items have the same meaning and the same relation with the latent variable for the whole duration of the study. However, if measurement invariance does not hold, the parameters of the items might change, namely item parameter drift (Donoghue & Isham, 1998), and then the latent state dispositions from different measurement occasion would not be comparable. Given this, it would be necessary to extend the model to handle item parameter drift or at least develop statistics to test whether there is item parameter drift on some items. Currently, one way to study measurement non-invariance on intensive longitudinal data from multiple participants has been proposed by Vogelsmeier et al. (2021) with the latent Markov latent trait analysis. Yet, given the complexity of this model, its use by practitioners might be limited and more research is needed to set guidelines in terms of minimum sample size, number of measurement occasions, or number of response options, are lacking.

Thirdly, the simulation study showed that the TV-DPCM requires more than 200 time points to perform well. This is considerably above the typical length of the time series observed in intensive longitudinal research of psychological dynamics (Vachon et al., 2019). To overcome this, future research can try to extend the model to multilevel settings. Moreover, the simulation also showed that the credibility intervals of most parameters tend to be very wide even with large samples, meaning that there is high uncertainty around the point estimates. This can be a drawback in practical settings as it can undermine efforts of doing statistical inference with the TV-DPCM. The most straightforward approach for practitioners to address this issue would be to increase the number of items. This would reduce the uncertainty around the point estimates, especially for the parameters concerning the latent dynamic process. Yet, increasing the number of items that measure the same construct is not an easy task to achieve in most intensive longitudinal data settings.

Lastly, in the empirical example, we showed that the TV-DPCM has a reasonable predictive value when comparing the expected mean scores and the observed mean scores. However, this is no guarantee that the model fits the empirical data well. For this, goodness of fit statistics should be developed for the TV-DPCM model and in general for the methods used to analyze intensive longitudinal data. Within the Bayesian framework, a method to assess the goodness of fit of a model is based on posterior predictive model checking methods (Gelman et al., 1996). These methods have also been developed for Bayesian IRT models (Li et al., 2017; Sinharay et al., 2006) but they need to be extended for the TV-DPCM to account for the time dependencies present in intensive longitudinal data. In fact, this is ongoing research that we expect to be also useful for the TV-DPCM and other IRT models for intensive longitudinal data. Furthermore, another challenge of Bayesian analyses is the specification of priors. Prior misspecification can not only affect the quality of the estimates but also the effectiveness of Bayesian fit statistics, such a posterior predictive model checking methods (Ames, 2018). In this study, we used typical priors used for IRT models. However,

further research is required to study prior sensitivity analysis of the TV-DPCM.

To conclude, bringing IRT with all its features to intensive longitudinal research is a great opportunity for the field. It is not only a tool that can help assessing the quality of the scales used in intensive longitudinal data but it can also provide insight in how to improve these scales. As a result, researchers might be able to make better inferences and comprehend better the psychological dynamics of the individuals.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. Dr. Jorge N. Tendeiro was receiving funding from the Japanese JSPS KAKENHI (21K20211) grant and dr. Laura F. Bringmann was receiving funding from the Netherlands Organization for Scientific Research Veni Grant (NWO-Veni 191G.037) at the moment of publication.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by a grant.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Ames, A. J. (2018). Prior sensitivity of the posterior predictive checks method for item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *16*(4), 239–255. https://doi.org/10.1080/15366367.2018.1502026

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. betancourt2018conceptual. https://doi.org/10.48550/arxiv.1701.02434

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological Methods*, *22*(3), 409–425. https://doi.org/10.1037/met0000085

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS One*. *8*(4), e60188. https://doi.org/10.1371/journal.pone.0060188

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Chatfield, C. (2003). *The analysis of time series: An introduction* (6th ed.). Chapman and Hall/CRC. https://doi.org/10.4324/9780203491683

Core Team, R. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Crayen, C., Eid, M., Lischetzke, T., & Vermunt, J. K. (2017). A continuous-time mixture latent-state-trait Markov model for experience sampling data: Application and evaluation. *European Journal of Psychological Assessment*, *33*(4), 296–311. https://doi.org/10.1027/1015-5759/a000418

Depaoli, S., Clifton, J. P., & Cobb, P. R. (2016). Just Another Gibbs Sampler (JAGS). Flexible software for MCMC implementation. *Journal of Educational and Behavioral Statistics*, *41*(6), 628–649. https://doi.org/10.3102/1076998616664876

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*(2), 309–326. https://doi.org/10.1111/j.2044-8317.1994.tb01039.x

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*(1), 33–51. https://doi.org/10.1177/01466216980221002

Dziak, J. J., Li, R., Zimmerman, M. A., & Buu, A. (2014). Time-varying effect models for ordinal responses with applications in substance abuse research. *Statistics in Medicine*, *33*(29), 5126–5137. https://doi.org/10.1002/sim.6303

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Lawrence Erlbaum. https://doi.org/10.4324/9781410605269

Fox, J.-P. (2010). *Bayesian item response modeling*. Springer. https://doi.org/10.1007/978-1-4419-0742-4

Gabry, J., Mahr, T. (2021). *Bayesplot: Plotting for Bayesian models*. https://mc-stan.org/bayesplot/

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–807. https://www.jstor.org/stable/24306036?seq=1#metadata_info_tab_contents

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Giraitis, L., Kapetanios, G., & Yates, T. (2014). Inference on stochastic time-varying coefficient models. *Journal of Econometrics*, *179*(1), 46–65. https://doi.org/10.1016/j.jeconom.2013.10.009

Haan-Rietdijk, S., de, Voelkle, M. C., Keijsers, L., & Hamaker, E. L. (2017, November). Discrete- vs. continuous-time modeling of unequally spaced experience sampling method data. *Frontiers in Psychology*, *8*, 1849. https://doi.org/10.3389/fpsyg.2017.01849

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. *Multivariate Behavioral Research*, *53*(6), 820–841. https://doi.org/10.1080/00273171.2018.1446819

Hamaker, E. L., Ceulemans, E., Grasman, R. P., & Tuerlinckx, F. (2015). Modeling affect dynamics: State of the art and future challenges. *Emotion Review*, *7*(4), 316–322. https://doi.org/10.1177/1754073915590619

Hamaker, E. L., & Wichers, M. (2017). No time like the present. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177/0963721416666518

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory : principles and applications*. Boston: Kluwer-Nijhoff Pub. https://doi.org/10.1007/978-94-017-1988-9

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press. https://doi.org/10.1515/9780691218632

Hecht, M., Hardt, K., Driver, C. C., & Voelkle, M. C. (2019). Bayesian continuous-time Rasch models. *Psychological Methods*, *24*(4), 516–537. https://doi.org/10.1037/met0000205

Kharratzadeh, M. (2017). *Splines in Stan*. https://mc-stan.org/users/documentation/case-studies/splines_in_stan.html

Kossakowski, J. J., Groot, P. C., Haslbeck, J. M. B., Borsboom, D., & Wichers, M. (2017). Data from critical slowing down as a personalized early warning signal for depression. *Journal of Open Psychology Data*, *5*(1), 1. https://doi.org/10.5334/jopd.29

Krieke, L. V. D., Jeronimus, B. F., Blaauw, F. J., Wanders, R. B. K., Emerencia, A. C., Schenk, H. M., Vos, S. D., Snippe, E., Wichers, M., Wigman, J. T. W., Bos, E. H., Wardenaar, K. J., & Jonge, P. D. (2016). HowNutsAreTheDutch (HoeGekIsNL): A crowdsourcing study of mental symptoms and strengths. *International Journal of Methods in Psychiatric Research*, *25*(2), 123–144. https://doi.org/10.1002/mpr.1495

Kropko, J. (2013). *Dynamic measurement of political phenomena: Item response theory for time-series data* (Tech. Rep.). Columbia University. https://nanopdf.com/download/item-response-theory-for-time-series-data_pdf#

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*(7), 984–991. https://doi.org/10.1177/0956797610372634

Li, T., Xie, C., & Jiao, H. (2017). Assessing fit of alternative unidimensional polytomous IRT models using posterior predictive model checking. *Psychological Methods*, *22*(2), 397–408. https://doi.org/10.1037/met0000082

Masters, G. N. (2016). Partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Volume 1: Models* (pp. 109–126). CRC Press. https://doi.org/10.1201/9781315374512

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11 Suppl 3), S69–S77. https://doi.org/10.1097/01.mlr.0000245438.73837.89

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models* (144th ed.). SAGE.

Rijn, P., V., Dolan, C. V., & Molenaar, P. C. M. (2010). State space methods for item response modeling of multisubject time series. In P. C. M. Molenaar & K. M. Newell (Eds.), *Individual pathways of change: Statistical models for analyzing learning and development* (pp. 125–151). American Psychological Association. https://doi.org/10.1037/12140-008

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85–100). Springer.

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, *24*(1), 70–91. https://doi.org/10.1037/met0000188

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n = 1 psychological autoregressive modeling. *Frontiers in Psychology*, *6*, 1038. https://doi.org/10.3389/fpsyg.2015.01038

Shumway, R. H., Stoffer, D. S. (2017). *Time series analysis and its applications: With R examples*. Springer International Publishing. https://doi.org/10.1007/978-3-319-52452-8

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321. https://doi.org/10.1177/0146621605285517

Song, H., & Zhang, Z. (2014). Analyzing multiple multivariate time series data using multilevel dynamic factor models. *Multivariate Behavioral Research*, *49*(1), 67–77. https://doi.org/10.1080/00273171.2013.851018

Stan Development Team. (2020). RStan: The R interface to Stan. http://mc-stan.org/

Stan Development Team. (2022). *Stan modeling language users guide and reference manual 2.29*. https://mc-stan.org

Tan, X., Shiyko, M. P., Li, R., Li, Y., & Dierker, L. (2012). A time-varying effect model for intensive longitudinal data. *Psychological Methods*, *17*(1), 61–77. https://doi.org/10.1037/a0025814

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577. https://doi.org/10.1007/BF02295596

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

van der Linden, W. J. (2016). Unidimensional logistic response models. In W. J. van der Linden (Ed.), *Handbook of item response theory: Volume 1: Models* (pp. 13–30). CRC Press. https://doi.org/10.1201/9781315374512

Voelkle, M. C., & Oud, J. H. (2013). Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes. *The British Journal of Mathematical and Statistical Psychology*, *66*(1), 103–126. https://doi.org/10.1111/j.2044-8317.2012.02043.x

Voelkle, M. C., Oud, J. H., Davidov, E., & Schmidt, P. (2012). An SEM Approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods*, *17*(2), 176–192. https://doi.org/10.1037/a0027543

Vogelsmeier, L. V., Vermunt, J. K., Keijsers, L., & De Roover, K. (2021). Latent Markov latent trait analysis for exploring measurement model changes in intensive longitudinal data. *Evaluation & the Health Professions*, *44*(1), 61–76. https://doi.org/10.1177/0163278720976762

von Davier, M.. (2016). Rasch model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Volume 1:* Models (pp. 31–48). CRC Press. https://doi.org/10.1201/9781315374512

Walls, T. A., & Schafer, J. L. J. L. (2006). *Models for intensive longitudinal data*. Oxford University Press.

Wang, X., Berger, J. O., & Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, *7*(1), 126–153. https://doi.org/10.1214/12-AOAS608

Wichers, M., & Groot, P. C.; Psychosystems, ESM Group, EWS Group. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and Psychosomatics*, *85*(2), 114–116. https://doi.org/10.1159/000441458

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). https://doi.org/10.1201/9781315370279