Published in final edited form as:
Behav Res Methods. 2024 March ; 56(3): 1506–1532. doi:10.3758/s13428-023-02107-3.

Dynamic structural equation models with binary and ordinal outcomes in M*plus*

Daniel McNeish[1], **Jennifer A. Somers**[2], **Andrea Savord**[1]

[1]Arizona State University, PO Box 871104, Tempe, AZ 85287, USA

[2]University of California, Los Angeles, Los Angeles, CA, USA

Abstract

Intensive longitudinal designs are increasingly popular, as are dynamic structural equation models (DSEM) to accommodate unique features of these designs. Many helpful resources on DSEM exist, though they focus on continuous outcomes while categorical outcomes are omitted, briefly mentioned, or considered as a straightforward extension. This viewpoint regarding categorical outcomes is not unwarranted for technical audiences, but there are non-trivial nuances in model building and interpretation with categorical outcomes that are not necessarily straightforward for empirical researchers. Furthermore, categorical outcomes are common given that binary behavioral indicators or Likert responses are frequently solicited as low-burden variables to discourage participant non-response. This tutorial paper is therefore dedicated to providing an accessible treatment of DSEM in M*plus* exclusively for categorical outcomes. We cover the general probit model whereby the raw categorical responses are assumed to come from an underlying normal process. We cover probit DSEM and expound why existing treatments have considered categorical outcomes as a straightforward extension of the continuous case. Data from a motivating ecological momentary assessment study with a binary outcome are used to demonstrate an unconditional model, a model with disaggregated covariates, and a model for data with a time trend. We provide annotated M*plus* code for these models and discuss interpretation of the results. We then discuss model specification and interpretation in the case of an ordinal outcome and provide an example to highlight differences between ordinal and binary outcomes. We conclude with a discussion of caveats and extensions.

Keywords

Intensive longitudinal data; Categorical data; Discrete data; DSEM; Time-series analysis

Research designs like experience sampling (Scollon et al., 2003), ambulatory assessment (Fahrenberg et al., 2007), daily diaries (Bolger et al., 2003), and ecological momentary assessment (Smyth & Stone, 2003) have sharply increased in popularity as smartphone and wearable technology have permitted data to be collected more intensively, more frequently,

✉Daniel McNeish, dmcneish@asu.edu.

Open science statement There was no preregistration for this paper because models were illustrative to demonstrate the method and contextualize the code and were not intended to address research hypotheses. All simulation code and simulation result files can be found on the Open Science Framework page associated with this project, located at https://osf.io/bx72m.

more naturally, and less invasively (Conner & Barrett, 2012; Hamaker & Wichers, 2017; Mehl & Conner, 2012; Nelson & Allen, 2018; Trull & Ebner-Priemer, 2014). Data collection with mobile or wearable devices reduces participant burden while also increasing ecological validity because responses are collected in real time rather than recalled after the fact (De Haan-Rietdijk et al., 2017). This higher ecological validity, ease of data collection, and reduced participant burden has resulted in intensive longitudinal designs becoming more commonplace in behavioral and health sciences, especially when studying affect, mood, interpersonal behaviors, or psychophysiology (Moskowitz & Young, 2006).

Intensive longitudinal designs produce dense datasets with many observations per person (roughly defined as data with 20 or more measurement occasions per person; Collins, 2006; Walls & Schafer, 2006), especially compared to traditional panel designs where each person is measured intermittently over a relatively longer timespan (e.g., Curran et al., 2010). Dense, frequently collected data facilitates research questions concerning within-person variability and how processes unfold moment to moment, which is opposed to the emphasis in panel data on between-person processes and mean changes across development (e.g., Ram & Gerstorf, 2009; Wang et al., 2012). Changes in data structure and corresponding changes in the types of questions researchers pose has led to rapid development of novel methodological approaches and software for intensive longitudinal data (e.g., Asparouhov et al., 2018; Driver et al., 2017; Hamaker et al., 2018; Ou et al., 2018).

As these models have increased in popularity, several didactic resources have been written to assist researchers embarking on these types of analyses (Asparouhov et al., 2018; Li et al., 2022; McNeish & Hamaker, 2020; Sadikaj et al., 2021; Zhou et al., 2021). These resources cover much ground given that intensive longitudinal models have many nuances with which researchers must familiarize themselves such as Bayesian estimation, new assumptions, and different parameters that may be included to capture conceptual differences in densely collected data. Even though many intensive longitudinal studies include outcomes that are binary behavioral items or single ordinal Likert items (e.g., Berli et al., 2021; DeMartini et al., 2022; Kiekens et al., 2020), discussions in existing didactic resources focus primarily on continuous outcomes. Categorical outcomes are either not mentioned (Li et al., 2022; McNeish & Hamaker, 2020), mentioned briefly in the discussion or limitations (Sadikaj et al., 2021, p. 38; Zhou et al., 2021, p. 243), or mentioned as straightforward extension of the continuous case (Asparouhov et al., 2018, p. 362–363; Asparouhov & Muthén, 2019, p. 135–136; Asparouhov & Muthén, 2020, p. 285–286).

We understand the viewpoint that – to more technical audiences – extensions to categorical outcomes may be considered relatively trivial statistically. For instance, as one option, researchers can assume that a normal distribution underlies the categorical variable such that the observed categorical responses are essentially a discretization of an unobserved continuous process. Then, one can apply the principles of a model for continuous outcomes to the underlying normal distribution instead of the raw categorical outcome.

Although this extension may be straightforward statistically, the corresponding changes in model fitting and interpretation when extending the model to categorical outcomes this way are not necessarily trivial for empirical researchers using these models to inform their conclusions (e.g., how to interpret coefficients and covariate effects, how to center or standardize categorical variables, working with a latent normal distribution rather than an observed variable). Essentially, the practical aspects of extending the model to categorical variables are less intuitive than the statistical aspects. Anecdotally, we as authors were motivated by this exact issue when encountering intensive longitudinal data with categorical outcomes. The statistical foundations of extending the model to categorical outcomes were relatively straightforward (assuming some familiarity with probit regression). However, we had more difficulty with interpretation and connecting the results to the original research questions.

Therefore, the goal of this paper is to provide a didactic resource devoted solely to modeling categorical outcome variables for empirical researchers whose experiences may match our own. Specifically, we discuss how to fit dynamic structural equation models (DSEMs) for intensive longitudinal data with categorical outcomes in M*plus* and provide software code and guidance on interpreting the results. To outline the structure of this paper, we first provide an overview of the foundations of intensive longitudinal data and the DSEM approach as implemented in M*plus* to model moment-to-moment dynamics. This overview focuses on the simpler case of continuous outcomes. We then provide an overview of probit modeling for categorical data. This approach is used within M*plus* to accommodate binary and ordinal outcomes and is central to proper interpretation of coefficients and centering, even though psychologists are not always familiar with concepts related to probit models (e.g., Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018). We highlight the connection between the probit and continuous models to motivate why the technical literature considers this extension statistically straightforward. We follow with a description of a motivating dataset from an empirical intensive longitudinal study on people with binge eating disorder that has both binary and ordinal outcomes. Four example analyses and associated code are then provided using this dataset and potential differences in model misspecification and interpretation between models for binary and ordinal outcomes are discussed. We conclude with a discussion of extensions and limitations.

Intensive longitudinal data and DSEM

Intensive longitudinal data

In contrast to the relatively recent increase in popularity of intensive longitudinal data in behavioral sciences, fields like finance and climatology have an extensive history of modeling dense, frequently collected data. In these fields, this type of data is more commonly referred to as *time-series data* and a common theme of time-series models is to determine how the preceding state of the system affects the subsequent state (Hamaker et al., 2018). A common way to accomplish this goal – both in historic applications of time-series models and in more recent extensions in behavioral sciences – is through *autoregressive* models where the outcome variable is predicted from itself at one or more earlier time-points. Each respective previous time-point is referred to as a *lag*. For instance, a lag-1 model uses the immediately preceding time-point as a predictor, a lag-2 model uses the two immediately preceding time-points as predictors, etc. The effect of the prior state of the system on its current state is referred to as carryover, inertia, or autoregression.

A common time-series model for continuous outcomes is the lag-1 autoregressive model, which can be written as

$$y_t = \alpha + \phi y_{t-1} + e_t \tag{1}$$

where y_t is the outcome at time t, α is the intercept of the time series, ϕ is an autoregression coefficient capturing the carryover effect from the first lag y_{t-1}, and e_t is the residual at time t which is normally distributed with a mean of 0 and constant variance σ^2. Autoregressive models of the form in Eq. (1) make a *stationarity* assumption such that the mean, variance, and autocorrelations of the outcome do not systematically change over time. This implies that the time-series is *mean-reverting* (Stroe-Kunold et al., 2012), meaning that the outcome variable may be higher or lower than the mean momentarily, but the expected value of the time series is not a function of time (i.e., the outcome does not systematically increase or decrease over time).

Time-series models have historically been applied to data from a single entity like a stock or weather from a particular location (i.e., $N = 1$ models). However, a challenge of time-series models in behavioral sciences is that intensive longitudinal data are typically collected for multiple people such that $N > 1$. Time-series data with $N > 1$ can be supported with bottom-up or top-down approaches (Liu, 2017). Bottom-up approaches are more idiographic and models are first fit to each person's data separately. Then, similarities between the dynamics of different people are sought either with automatic searches (e.g., group iterative multiple model estimation, GIMME; Gates & Molenaar, 2012) or by constraining parameters across people (Hamaker et al., 2003). Conversely, top-down approaches are more nomothetic and fit a model with the same functional form to all people but permit between-person variability in the parameters. Different dynamics across people are accomplished with random effects such that the parameter is modeled as a distribution of values rather than a single value. In the common context where the distribution of the parameter is assumed to be normal, there is a fixed effect capturing the average parameter value across all people and a variance capturing the heterogeneity of the person-specific parameter values across people.

Top-down models have historically been fit in the mixed effect framework (e.g., Bolger & Laurenceau, 2013; Walls & Schafer, 2006), but mixed effect models possess weaknesses for intensive longitudinal data in some contexts (McNeish & Hamaker, 2020). First, mixed effect models can be challenging with unequal intervals between measurement occasions (which are an intentional feature of some research designs like ecological momentary assessment to prevent participants from anticipating the next measurement occasion or experience sampling). Second, person-mean centering is applied to disaggregate within-person from between-person processes (e.g., Curran & Bauer, 2011; Hamaker & Grasman, 2015) but centering in mixed effect models is based on observed person-means, which is known to be produce estimates that are susceptible to Nickell's bias (Nickell, 1981) and Lüdtke's bias (Lüdtke et al., 2008) with intensive longitudinal data. Nickell's bias results in underestimated autoregressive effects with observed person-mean centering because the error term is not necessarily independent of the lagged predictor (i.e., endogeneity is present). This bias is unaffected by the number of people but diminishes as the number of time points approaches infinity. Lüdtke's bias notes that the observed person means are susceptible to measurement error, unreliability, and missing data; so, using the observed person-mean can lead to bias to the extent that the observed person mean does not reflect the true person mean. Third, standard mixed effect models assume that all variables are manifest and do not directly allow measurement models or invariance testing (e.g., Castro-Alvarez et al., 2022; McNeish et al., 2021; Vogelsmeier et al., 2022).

Dynamic structural equation models

To address some weaknesses of mixed effect models for intensive longitudinal data, the DSEM framework was recently introduced and incorporated into the M*plus* software program (Asparouhov et al., 2017, 2018). Throughout this paper, we focus on application of DSEM in M*plus* although the same models can be fit in general Bayesian software like `Stan, JAGS,` or `Win-BUGS` (Hamaker et al., 2023; Li et al., 2022). The `brms` R package (which interfaces with `Stan` to fit multivariate multilevel models) can also support many types of models in the DSEM framework, particularly with observed variables (Williams et al., 2020, p. 989; also see ten Brink et al., 2021 for an application of DSEM with the `brms` package).

DSEM integrates (a) *time-series analysis* to allow lagged relations for modeling autoregressive effects between densely collected repeated measures and Kalman filters for unequally spaced intervals between observations (b) *multilevel modeling* to accommodate repeated measures nested within multiple individuals and to allow individual differences in parameters with a top-down approach, and (c) *structural equation modeling* to permit multivariate models, latent variables, and full structural models that allow any between-person variable (including latent variables) to be a predictor, mediator, or outcome. DSEM allows users to leverage aspects of these three approaches to meet the demands of the intended model. In other words, standard mixed effect models that can be fit in software like SAS `PROC MIXED,` SPSS `MIXED,` or the `Ime4`R package are a special case of the broader DSEM framework, meaning that the DSEM results will be the same when the additional features are not needed but that DSEM can extend beyond capabilities of standard mixed effect models (Savord et al., 2023).

In DSEM, a Kalman filter can help address issues related to unequal intervals between observed data that are present with traditional mixed effect models (Kim & Nelson, 1999). The Kalman filter originated in aerospace engineering and is based on hidden Markov models. When applied to time-series analysis, the general idea is that researchers specify the largest interval in which only one observation can occur. The Kalman filter makes predictions of the value within each interval based on previous values. For intervals with observed data, the predictions are updated with the observed information. For intervals without observed data (e.g., missing data, no response was solicited), the Kalman filter retains its prediction from the previous interval and continues to the next interval without updating. Kalman filters have been integrated in M*plus* but typically must be manually programmed for models fit in general Bayesian software.

With regard to centering, Nickell's and Ludtke's biases that emerge with observed centering in traditional mixed effect models can be avoided in DSEM by centering around *latent* means that can incorporate measurement error and unreliability into the measurement process (e.g., Asparouhov & Muthén, 2019). Latent centering has been found to be effective when the number of time-points per person is ten or more (Gistelinck et al., 2021). Issues in mixed effect models pertaining to multiple outcomes and latent variables can also be accommodated in DSEM by incorporating principles of structural equation modeling, which is inherently multivariate and naturally accommodates measurement models for latent variables.

Though many of the existing papers in the short history of the DSEM literature have focused on continuous outcomes, it is possible to incorporate categorical outcomes as well. When fitting categorical models in M*plus*, a probit link is used rather than a logit link as is commonly used in logistic regression in behavioral sciences. The next section familiarizes readers with the concepts behind probit regression before we discuss extending DSEM to binary outcomes using a probit link.

Overview of probit models

Probit regression falls under the umbrella of the generalized linear model for modeling non-normal outcomes. Similar to other generalized linear models like logistic regression or Poisson regression, the relationship between predictors and the outcome is linear after applying a link function. In logistic regression, the coefficients are linear in the log-odds; in Poisson regression, the coefficients are linear in the natural log of the count. Probit models relate the predictors to the outcome linearly through the *standard normal cumulative distribution function* (denoted $\Phi(\cdot)$).

This sounds like a mouthful, but the concept is less daunting than the statistical verbiage used to describe it. To simplify the explanation with an example, imagine a model for a binary outcome y with a single continuous predictor x_1. A probit model would be written as $\Pr(y = 1 \mid x_1) = \Phi(\beta_0 + \beta_1 x_1)$. Notice that the model still contains a linear regression equation, but it appears within the $\Phi(\cdot)$ function. This equation would be translated as, "the probability that y equals 1, given the value of the predictor x_1, is equal to the standard normal cumulative distribution function of a point defined by an intercept (β_0) plus the coefficient associated with x_1 (β_1) times the value of x_1 ". In English, "standard normal cumulative distribution function" is the area under a Z-distribution to the left of a particular Z-score. Essentially, $\Phi(\cdot)$ is shorthand for "area under a Z-distribution from $-\infty$ to the number in parentheses". β_0, β_1, and x_1 determine the Z-score to appear in the parentheses, which is then converted to probability by taking the area under a Z-distribution to the left of the Z-score implied by β_0, β_1, and x_1.

Imagine that $\beta_0 = -1$ and $\beta_1 = 1.5$. The intercept coefficient β_0 corresponds to the Z-score associated with the predicted probability that $y = 1$ when $x_1 = 0$. To get the predicted probability from the probit coefficient of −1, we calculate the area of a Z-distribution from $-\infty$ to −1. Statistically, this would be written as $\Phi(\beta_0 + \beta_1 x_1) = \Phi(-1 + 1.5 \times 0) = \Phi(-1) = 0.16$. Thearea to the left of −1 on a Z-distribution is about 0.16, which can be looked up in a Z table or calculated using a software utility like the NORM.S.DIST function in Excel or the `pnorm` function in R. The predicted probability that $y = 1$ when $x_1 = 0$ would therefore be 0.16. This is shown visually in the left panel of Fig. 1.

The probit model is linear in the Z-score associated with the predicted probability: a 1-unit change in x_1 leads to a β_1-unit change in the Z-score associated with the predicted probability. The value of β_1 in this example was 1.5, so the Z-score associated with the predicted probability of y increases by 1.5 units if $x_1 = 1$ compared to when $x_1 = 0$. As shown in the right panel of Fig. 1, the predicted probability would be $\Phi(-1 + 1.5 \times 1) = \Phi(0.50) = 0.69$. The regression coefficients determine the Z-score in parentheses and the $\Phi(\cdot)$ converts this Z-score to a predicted probability based on the area under the curve to the left of the Z-score.

Similar to logistic regression, probit regression is nonlinear when converted to the probability scale. For instance, in the previous paragraph we saw that the predicted probability that $y = 1$ when $x_1 = 0$ was $\Phi(-1) = 0.16$ and the predicted probability that $y = 1$ when $x_1 = 1$ was $\Phi(0.50) = 0.69$, a difference of 53 percentage points. If x_1 were equal to 2, the Z-score associated with the predicted probability of $y = 1$ would increase by another 1.5 points (the value of β_1), making the predicted probability $\Phi(2.0) = 0.98$. The difference in predicted probabilities from $x_1 = 1$ to $x_1 = 2$ is only 29 percentage points. Because probabilities are bounded between 0 and 1, linear changes on the Z-score scale do not correspond to linear changes on the probability scale. Although the model is linear with respect to the Z-score on the probit scale, it is nonlinear on the probability scale.

As will become relevant later, probit models can also be parameterized to model the probability that the outcome is 0 instead of 1 by multiplying the right-hand side of the model equation by −1. That is, $\Pr(y = 0 \mid x_1) = \Phi(-\beta_0 - \beta_1 x_1)$. In this case, $-\beta_0$ is often referred to as the *threshold* (usually denoted by τ_0) such that the threshold corresponds to the Z-score associated with the predicted probability that $y = 0$ when $x_1 = 0$. To use the same coefficients as the earlier example such that $\tau_0 = -\beta_0 = 1$ and $\beta_1 = 1.5$, we can calculate the predicted probability that $y = 0$ when $x_1 = 0$ by $\Phi(\tau_0 - \beta_1 x_1) = \Phi(1 - 1.5 \times 0) = \Phi(1) = 0.84$. This is represented by the white space under the curve in the left panel of Fig. 1. Similarly, the predicted probability of $y = 0$ when $x_1 = 1$ would be $\Phi(\tau_0 - \beta_1 x_1) = \Phi(1 - 1.5 \times 1) = \Phi(-0.50) = 0.31$. This is represented by the white space under the curve in the right panel of Fig. 1. These values complement the values above and merely show how to reparametrize the model to recover the probability that the outcome is 0 rather than 1.

Thresholds and intercepts are related but serve slightly different purposes. In the binary case where there are only two response categories, the difference is minimal. However, differences are more pronounced with three or more response categories, which we will discuss later in the section on ordinal outcomes.

An alternative conceptualization of probit models

Another way to think about the probit model is to imagine that there is an underlying normal distribution that only manifests into categorical responses (Agresti, 2012). In this way, the underlying nature of the variable changes imperceptibly but the realized values of a categorical variable only change once some threshold has been passed. From this perspective, the underlying normal distribution is the true interest and the categorical responses are simply an imprecise reflection of the underlying normal distribution (Long, 1997, p. 116). The normal distribution is not just a link for computing predicted probabilities but instead is assumed to be an unobserved process that manifests categorical responses. For example, if participants are asked if they smoked a cigarette, the observed information may be “Yes” or “No”, but one could imagine a continuous (but harder to measure) underlying process like “motivation to smoke” or “nicotine withdrawal” driving this decision and participants will choose to smoke once some threshold on the underlying process has been exceeded.

Reconsidering Fig. 1, this would mean that the data only provide information on 0s and 1s even though the real process is continuous. Any person whose normal distribution value lies to the left of the threshold (the vertical dashed line) in the grey area of Fig. 1 responds as a “1” and anyone whose normal distribution value lies to the right of the threshold in the white area of Fig. 1 responds as a “0”. We do not see the normal distribution values in the data, we only see the 0s and 1s. Therefore, the observed categorical responses are a rough ordinal approximation of the more articulate - but unobserved - normal distribution. Of course, this perspective may not apply universally because some outcomes truly are discrete processes. For instance, it may make less sense to consider the outcome of a coin flip as having an underlying normal process dictating the result. Nonetheless, this perspective is often appropriate in behavioral research where complex - but unobserved or difficult to measure- processes drive discretely measured behavior.

This perspective is consistent with how the probit model was developed in toxicology by Bliss (1935), who studied pest death (a binary variable) as a function of pesticide dose. The vital functioning of an insect is actually a continuous process and the amount of pesticide consumed by an insect continuously changes vital functioning (e.g., consuming some pesticide can have detrimental but non-lethal effects like diminished organ functioning). However, it is not feasible to precisely gauge insect vital functioning, so the effect of pesticide consumption on vital functioning is most easily observable as a dichotomous process (i.e., alive vs. dead) based on whether some threshold has been crossed (i.e., a lethal dose of pesticide).

The assumption that a normal process underlies the manifest categorical variable is why categorical models are often stated to be a straightforward extension of continuous models in the intensive longitudinal literature. This latent underlying normal process can be modeled instead of the manifest categorical data that were collected from participants, which permits the principles of continuous outcomes to be applied, even if the manifest data are not themselves continuous. In other words, the latent normal process is substituted for the manifest categorical variable to replace the difficult aspects of modeling categorical data with the more tractable principles of modeling continuous data. As we discuss in more detail shortly, assuming an underlying normal process facilitates the statistical aspects of the model because the normal process - even if latent - may be easier mathematically and computationally than an observed categorical variable. However, this affects substantive considerations because the coefficients and conclusions will pertain to the abstract concept of the unobserved, underlying normal process rather than the observed categorical data.

Although probit and logistic regression models can often be applied to the same data with similar results, a benefit of the probit model with Bayesian estimation (which is used for DSEM in M*plus*) is that using $\Phi(\cdot)$ as the link function makes computation with a Gibbs sampler much simpler with normal priors compared to logistic regression (e.g., Agresti & Hitchcock, 2005; Albert & Chib, 1993). This can make estimation with Bayesian Markov Chain Monte Carlo (MCMC) fast and efficient, even with categorical outcomes (Asparouhov & Muthén, 2021). As a result, as of M*plus* Version 8.8 released in April 2022, DSEM with categorical outcomes can only be fit with a probit link function. However, other more general software may permit users to use a logit (or other) link function if interpretations provided on other link functions are preferred.

Motivating data

The motivating data are an intensive longitudinal study supported by the National Institutes of Health's Science of Behavior Change initiative (Eisenberg et al., 2018; Nielsen et al., 2018; Scherer et al., 2022). Complete details about data collection and the research design are reported on the study's dedicated page on ClinicalTrials.gov.[1] To summarize key characteristics, one of the study's focal populations was people with binge eating disorder. The study therefore sampled 50 overweight/obese adults $\left(27 \leq \mathrm{BMI} \leq 45\ \mathrm{kg/m^2}\right)$ who met DSM- 5 criteria for binge eating disorder. All 50 participants had access to a mobile intervention app each day during a 28-day observation period and were asked to engage with the app to learn and apply techniques useful for modifying health behavior. The number of steps taken (as measured by pedometers) was collected daily for these participants as a measure of health behavior as was whether the participant engaged with the mobile intervention app each day of the study (i.e., intervention adherence is time-varying). Each morning, questions about participants' feelings related to binge eating behavior were solicited (e.g., "On a scale of 1 to 10, how motivated are you to avoid binge eating today?"), and at the end of the day, questions about participants' temptation related to binge eating behavior were also solicited (e.g., "Tempting food made it difficult for me to not binge eat today").

In the sections that follow, we provide example analyses using this data. The first three examples use binary intervention adherence as the outcome variable. We start with the simplest case of a model with no covariates to cover the foundational aspects of a time-series model with categorical outcomes. Then, we extend the model to include a time-varying covariate to assess whether binge eating avoidance affects adherence (both within-person or between-person). In the third example, we discuss a model for outcomes that have systematic trends over time. Lastly, we discuss a model an ordinal outcome ("temptation")

[1] https://www.clinicaltrials.gov/ct2/show/NCT03774433?term=marsch&draw=2&rank=3.

based on a single Likert response that includes a time-varying covariate. Each example discusses the statistical model, the M*plus* code, and interpretation.

Example 1: Unconditional probit DSEM for a binary outcome

We start with an example of an $N > 1$ multilevel autoregressive probit model for a binary outcome. In the motivating data, the daily measure of intervention adherence takes values of 0 (no adherence) or 1 (adherence) and is measured 28 times per person over the course of the study. There is very little missing data on Adherence (0.1%) because failure to participate in the study on a particular day is coded as a 0 for non-adherence rather than missing. The only two instances of missing data (out of 1400 possible observations) were due to technical difficulties with the mobile application that precluded participants from signing in. This variable will serve as the outcome such that the main interest is the moment-to-moment dynamics of adherence to the intervention. As some exploratory data analysis, the mean proportion of days where participants adhered to the intervention was 0.77. There was notable variability across people and the range of treatment adherence across all 50 people was 0.07 to 1.00 with an interquartile range of [0.64, 0.96] and a median of 0.89. Figure 2 shows trace plots of intervention adherence for four participants with adherence rates near the mean.

Model equation and path diagram

The naïve approach to writing out a multilevel autoregressive probit model for binary time-series data would be

$$\Pr\left(Adhere_{ti} = 1\right) = \Phi(\alpha_i + \phi_i Adhere_{t-1,i}) \tag{2}$$

where the probability of adhering to the intervention at time t $(t = 1, \ldots, 28)$ for person i $(i = 1, \ldots, 50)$ is equal to the standard normal cumulative distribution function (the Φ operator) of a function defined by a standard lag-1 autoregressive model with a person-specific intercept for person i (α_i) plus the autoregressive effect (ϕ_i) of adherence for the same person from the previous day. When specifying the model this way, $\Phi(a_1)$ would correspond to the predicted probability that person i adheres at time t if they did not adhere at time $t - 1$ (i.e., $Adhere_{t-1,i} = 0$). The predicted probability that person i adheres at time t if they did adhere at time $t - 1$ (i.e., $Adhere_{t-1,i} = 1$) would then be $\Phi(a_1 + \varphi_i)$. As in typical probit models, the idea is that predicted probabilities of adherence are determined through a Z distribution. Though straightforward conceptually, setting up the model with $Adhere_{t-1,i}$ uncentered and in its raw binary form renders the parameter estimates susceptible to bias and can conflate within-person and between-person processes (Asparouhov & Muthén, 2019; Hamaker & Grasman, 2015). That is, the difference between momentarily adhering and habitually adhering would not be distinguishable and these both effects would be combined into single, blended estimate (Hoffman, 2019).

With continuous outcomes, this issue can be addressed by *latent centering* the lagged predictor such that the person-specific intercept would be subtracted from the lagged predictor (i.e., $Adhere_{t-1,i} - \alpha_i$; Yaremych et al., 2022). This would rescale the lagged predictor such that 0 would indicate that the lagged predictor is equal to the person's mean, positive values would indicate the lagged predictor is above the person's mean, and negative values would indicate the value is below the person's mean. This helps to isolate the within-person effect because the person mean is factored out, leaving only momentary deviations from the person's typical behavior. It also treats the person mean as a latent variable to account for possible measurement error or unreliability in the observed data (e.g., Lüdtke et al., 2008). Although straightforward to apply for continuous variables, this approach is problematic for categorical variables because the latent person mean (as captured by α_i) is on a different scale than the lagged predictor. That is, $Adhere_{t-1,\,i}$ is on raw scale and can only take values of 0 or 1 whereas α_i is a parameter on a Z-scale given that it appears after the link function.

Instead, the model can be reparametrized by working with the unobserved normal process underlying the binary Adherence variable, which we refer to as $y*$. With this parameterization, values for Adherence would manifest depending on $y*$ such that

$$Adhere_{ti} = \begin{cases} 1 \text{ if } y_{ti}^{*} > 0 \\ 0 \text{ if } y_{ti}^{*} \leq 0 \end{cases} \tag{3}$$

Using this connection, the multilevel autoregressive probit model - with the lagged predictor centered - could be written as,

$$\begin{aligned} &\Pr\left(Adhere_{ti} = 1\right) = \Phi\left(y_{ti}^{*}\right) \\ &y_{ti}^{*} = y_{ti}^{*(w)} + \alpha_i \\ &y_{ti}^{*(w)} = \phi_i y_{t-1,i}^{*(w)} + e_{ti} \end{aligned} \tag{4a}$$

The difference between Eq. (2) and Eq. (4a) is that Eq. (4a) is modeling the underlying normal process of adherence $(y*)$ rather than directly modeling raw binary Adherence. The linear predictor in the argument of $\Phi(\cdot)$ in Eq. (4a) therefore has four changes:

1. The underlying normal process $(y*)$ is decomposed into a latent person mean (α_i) that captures habitual behavior and a within-person component $(y_{ti}^{*(w)})$ that captures momentary deviations from the latent person mean.

2. The lagged predictor $(Adhere_{t-1,\,i})$ is replaced by $y_{t-1,i}^{*(w)}$. This allows the lagged predictor to be latent-centered such that $y_{ti}^{*(w)} = y_{ti}^{*} - \alpha_i$ by rearranging the second expression in Eq. (4a) because the lagged predictor and the latent mean are now on the same scale.

3. ϕ_i becomes a person-specific *tetrachoric* autocorrelation (Asparouhov & Muthén, 2019), which is a correlation in the underlying normal processes of two binary variables rather than a correlation between the raw binary variables themselves (e.g., a phi correlation).

4. The linear predictor now has an error term, e_{ti}. For identifiability, a typical assumption for e_{ti} is that it follows a standard normal distribution such that $e_{ti} \sim \mathcal{N}(0,\ 1)$.

Equation (4a) is the *within*-person equation, which models moment-to-moment dynamics. Multilevel autoregressive models also have a *between-person* model for associations among variables that are constant over time (i.e., that are time-invariant). Every parameter with an i subscript in the within-person model becomes an outcome in the between-person model. The between-person model associated with Eq. (4a) could be written as,

$$\begin{aligned} &\alpha_i = -\tau_0 + u_{0i} \\ &\phi_i = \gamma_{10} + u_{1i} \\ &\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \\ \sigma_{10} & \sigma_{11} \end{bmatrix}\right) \end{aligned} \tag{4b}$$

The first expression is a little unorthodox because M*plus* uses a latent intercept (α) but defines the fixed effect in terms of a threshold (τ).[2] The threshold parameterization is useful for outcomes with three or more categories (discussed in detail later), which makes the threshold parameterization more generalizable. However, it can sometimes be unintuitive with binary outcomes, so it is important to remember this distinction where interpreting M*plus* output for a model with binary outcomes.

With this in mind, the person-specific intercept (α_i) is modeled by the opposite of the fixed effect for the threshold $(-\tau_0)$ to capture the average intercept across people, plus a person-specific random effect (u_{0i}) that captures the deviation of person i's intercept from the average intercept. The second expression shows that the person-specific tetrachoric autocorrelation (ϕ_i) is equal to a fixed effect (γ_{10}) that represents the average tetrachoric autocorrelation across all people plus a person-specific random effect for tetrachoric autocorrelation (u_{1i}) that captures the deviation of person i's tetrachoric autocorrelation from the overall average tetrachoric autocorrelation. The last expression specifies the distributional assumptions for the random effects, which states that the person-specific intercepts and tetrachoric autocorrelations are distributed multivariate normal with a mean vector of $\mathbf{0}$ and a covariance matrix $\mathbf{\Sigma}$. The σ_{00} term represents the variance of the person-specific intercepts, σ_{11} represents the variance of the person-specific tetrachoric autocorrelations, and σ_{10} represents the covariance between the random intercepts and random tetrachoric autocorrelations (e.g., to assess whether there is a systematic relationship between where person's baseline probability of adhering and the strength of the person's tetrachoric autocorrelation).

Equations (4a) and (4b) can be combined into one complete multilevel model such that,

[2] We thank Linda Muthén for clarifying and confirming this.

$$\begin{array}{ll}\text{Latent}-\text{Decomposition} & \begin{cases}\Pr[Adhere_{ti}=1]=\Phi\left(y_{ti}^{*}\right)\\ y_{ti}^{*}=y_{ti}^{*(w)}+\alpha_{i}\end{cases}\\ \text{Within}-\text{Person} & \begin{cases}y_{ti}^{*(w)}=\phi_{i}y_{t-1,i}^{*(w)}+e_{ti}\\ e_{ti}\sim\mathcal{N}(0,1)\end{cases}\\ \text{Between}-\text{Person} & \begin{cases}\alpha_{i}=-\tau_{0}+u_{0i}\\ \phi_{i}=\gamma_{10}+u_{1i}\\ u_{i}\sim\mathcal{N}(0,\ \Sigma)\end{cases}\end{array} \tag{5}$$

The first set of expressions clarifies that there is an assumed latent normal process $\left(y_{ti}^{*}\right)$ underlying the binary outcome Adhere and that this underlying process in made up of a within-person component $(y_{ti}^{*(w)})$ capturing momentary states and a between-person component (α_i) capturing habitual traits. The second set of expressions shows the within-person model for the momentary dynamics. The third set of expressions then show a top-down approach such that the average associations are explicitly modeled with fixed effects (captured by the τ and γ terms) and that there is between-person variability (captured by the $\mathbf{\Sigma}$ matrix) to allow heterogeneity in the parameter values across people.

A path diagram for the model in Eq. (5) is shown in Fig. 3. Rectangles indicate observed variables, circles represent latent variables, and triangles represent constants. The left panel shows the two aspects of the latent decomposition. First, a latent normal process is assumed to underlie the binary Adherence variable, which is shown by the wavy line between y_{ti}^{*} and $Adhere_{ti}$ (following the convention from de Boeck & Wilson, 2004). Second, a latent decomposition is then performed to partition y_{ti}^{*} into within-person $(y_{ti}^{*(w)})$ and between-person (α_i) components. The between-person component, α_i, is the latent person mean of y_{ti}^{*} and captures the trait-level information about Adherence (technically, the underlying normal process that manifests as Adherence) and only has an i subscript because it is time-invariant. The within-person component is the deviation of y_{ti}^{*} from the latent person mean, which represents the state-level information about Adherence (again, technically the underlying normally process that manifests as Adherence).

The top panel of Fig. 3 shows the within-person model. The within-person component of the underlying normal process $y_{ti}^{*(w)}$ is autoregressed on itself at the previous timepoint $y_{t-1,i}^{*(w)}$. The intercept of $y_{ti}^{*(w)}$ is fixed to zero because, on average, y_{ti}^{*} is expected to be at the person mean (i.e., the expected value of $y_{ti}^{*}-\alpha_i$ is zero). The autocorrelation path has a circle placed over it (following the notational convention from Curran & Bauer, 2007), which indicates that the path does not have a single value but instead is a latent variable with a distribution of values that vary across people. This latent variable - along with the latent person mean α_i – then become outcomes in the between-person model (the bottom panel of Fig. 3), where they have an average value across all people $(-\tau_0$ and $\gamma_{10})$, between-person variances $(\sigma_{00}$ and $\sigma_{11})$, and a between-person covariance (σ_{10}).

M*plus* code

The annotated code for fitting the full model in Eq. (5) in M*plus* is shown below. Latent centering through the unobserved normal process of the binary variable is the default approach in M*plus*, so the code is more compact than the explanation above. Text appearing after an exclamation point (!) is a comment that describes what each line does but is not needed to successfully run the code. In describing this example, we also will overview the basic setup of a DSEM model in M*plus*.

The `VARIABLE` statement specifies that Adherence is `CATEGORICAL,` the `TINTERVAL` option declares that we expect participants to be observed for each 1-unit interval of `day` (a Kalman filter is applied whenever no observation is observed to handle unequal intervals; if data were collected more or less frequently, the number in parentheses would change. This value is also sensitive to how time is coded.), the `CLUSTER` option specifies that repeated measures with the same `id` variable belong to the same person, the `MISS ING` option specifies that missing data are represented by a period, and the `LAGGED` option specifies that we want M*plus* to create a lag-1 predictor for Adherence. In the `ANALYSIS` statement, we identify that the model has a two-level structure (`day` clustered in `id`) with random slopes using `TYPE = TWOLEVEVEL RANDOM.` DSEM analyses can only be estimated with Bayesian MCMC and we specify that the algorithm should run for a minimum of 5000 iterations (maximum iterations can be changed by including a number in the `BITERATIONS` option outside of parentheses, the default is 50,000). More details on Bayesian estimation in M*plus* are provided in the next section.

```
CATEGORICAL=adhere; !specify variables that are categorical;
TINTERVAL = day(1); !specify the largest increment of time that can have
only one observation;
CLUSTER= id; !repeated measures are clustered within ID;
MISSING ARE .; !identify missing data code;
LAGGED= adhere(1); !create lag-1 predictor adherence;
ANALYSIS:
TYPE= TWOLEVEL RANDOM; !Two-Level Model with random effects;
ESTIMATOR=BAYES;!Lagged variables can only be estimated with Bayes in Mplus;
BITERATIONS= (5000); !Run at least 5000 iterations of the MCMC algorithm;
MODEL:
%WITHIN%
Phi|adhere ON adhere&1;
!Adherence yesterday is related to Adherence today, phi_i;
!The ampersand (“&”) in Mplus is a keyword for a lagged predictor;
!”phi|” adds a random effect so this effect is allowed to vary across people;
%BETWEEN%
adhere; !between-person variance in adherence intercept, sigma_00;
phi; ! between-person variance in tetrachoric autocorrelation, sigma_11;
phi WITH adhere; !between-person covariance of intercept and
autocorrelation, sigma_10;
```

```
[phi]; fixed effect of autocorrelation, gamma_10;
[adhere$1]; fixed effect of threshold tau_0;
OUTPUT: STDYX;
```

The `MODEL` statement outlines the desired associations between variables. For an $N > 1$ DSEM, there are two parts of the `MODEL` statement: `◦/◦ WITHIN◦` corresponds to the within-person model (i.e., Eq. (4a) and `◦BETWEEN÷` corresponds to the between-person model (i.e., Eq. 4b). An ampersand (&) following a variable name is used to denote the lag of the variable. So `adhere ON adhere&1` specifies that Adherence is regressed on lag-1 Adherence. To allow paths to vary across people, a vertical pipe (I) is placed before the path with an arbitrary label to name the latent variable being assigned to the path. For instance, `phi | adhere ON adhere & 1` specifies that the lag-1 autoregression is being assigned a latent variable named "phi" and the value of this path is a distribution rather than a single value. By default, M*plus* will include random intercepts for outcomes, so only random effects for slopes need to follow this convention. The default in M*plus* Version 8.1 or later is also to latent center any predictors not in a `WITHIN` or `BETWEEN` option in the `VARIABLE` statement, so the lagged predictor will automatically be latent centered in this code. Also note that the label for the raw binary variable is used in the code despite the fact that the model operates on the normal process assumed to underlie the binary variable. M*plus* will perform the necessary transformation behind the scenes and there is no code necessary to specify the model this way once the variable is included in the `CATEGORICAL` statement.

In the between-person model, fixed effects are placed in square brackets whereas variances and covariances are specified by plain variable names. M*plus* uses the outcome variable in the between-person model to represent the intercept variance. The fixed effect for the intercept is parameterized in terms of the threshold, so `[adhere $1]` corresponds to the threshold fixed effect (which needs to be multiplied by −1 to convert it back to the intercept). The dollar sign "$" in the square brackets is an M*plus* keyword corresponding to which threshold to estimate. With ordinal outcomes, there are multiple thresholds to consider and different numbers may appear here. In the simpler case of binary outcomes, there is only a single threshold to differentiates 0 s from 1s, so a 1 will always appear after the dollar sign for binary outcomes.

Model fitting

The code in the previous section was run in M*plus* Version 8.7 to fit the model in Eq. (5) to the motivating data. For readers looking for a brief overview of Bayesian MCMC with DSEM, we suggest p. 614 of McNeish and Hamaker (2020). This code uses the M*plus* default settings of two chains using the potential scale reduction method (Gelman & Rubin, 1992) with a stringent threshold of $\hat{R} \leq 1.10$ for all parameters (Brooks & Gelman, 1998, p. 442) to determine convergence. The code sets a minimum of 5000 iterations before the chains are allowed to stop, which is not an M*plus* default and is explicitly set in the code. By default, M*plus* discards the first half of iterations as burn-in and posteriors are based on the second half of iterations. Prior distributions were set to the M*plus* defaults, which are $\mathcal{N}(0, \infty)$ for the autocorrelation fixed effect, $\mathcal{N}(0,\ 5)$ for the threshold fixed effect, and

$\mathscr{W}^{-1}(\mathrm{I}_2, 3)$ for the random effect covariance matrix. M*plus* output for each example analysis is provided on the Open Science Framework page for this project. [3]

$\mathscr{W}^{-1}$ is the inverse Wishart distribution, which is a multivariate distribution applied to all elements of a matrix simultaneously as opposed to placing priors on each element of the matrix individually. The benefit of the inverse Wishart prior is that its support is restricted to positive definite matrices, which avoids potential nonpositive definite issues that can arise with placing priors on individual elements (e.g., all covariance matrices are square and symmetric but not all square symmetric matrices are covariance matrices). Using an inverse Wishart whose first argument is an identity matrix and whose second argument (the degrees of freedom) is the dimension of the matrix plus one yields a marginal distribution for random effect correlations that is uniform over [−1, 1], which is uninformative and gives equal consideration to any admissible random effect correlation (Asparouhov & Muthén, 2010).

Bayesian estimation is preferred for DSEM for computational purposes because maximum likelihood and other frequentist methods often encounter convergence issues or are intractable with many latent variables (Asparouhov et al., 2018). This reflects a “Bayes as Computational Frequentism” approach whereby computational advantages of MCMC motivate Bayesian methods rather than a philosophical Bayesian approach where subjective beliefs are explicitly built into the model (Levy & McNeish, 2022). Because the motivation for Bayesian methods is computationally motivated rather than philosophically motivated, the interpretation of the results minimally deviates from a model estimated with frequentist methods and the M*plus* output changes little when the estimation method is changed.

Results and interpretation

The posterior distribution medians and 95% credible intervals for the unconditional probit DSEM model are shown in Table 1. Throughout this paper we refer to the posterior distribution medians as “estimates” to correspond to how M*plus* labels the output but note that Bayesian MCMC provides an entire distribution of values for each parameter (which can be summarized by a measure of central tendency like the median) rather than a single point estimate for each parameter as with a frequentist estimator like maximum likelihood. Bayesian MCMC also provides credible intervals (CrI) rather than confidence intervals (CI) provided in frequentist analyses. The difference is that the goal of CIs is to determine limits within which the true population value would be found a certain percentage of the time if the study were infinitely repeated whereas CrIs summarize uncertainty in the parameter estimates for the given data and priors. CIs and CrIs both measure uncertainty and may be similar, but can differ non-trivially because they are quantifying uncertainty in different ways.

The first row in Table 1 shows that the fixed effect for the threshold is $\tau_0 = -1.28$. Given the latent centering, probit assumptions (i.e., the left panel of Fig. 3), and M*plus* intercept parameterization; the interpretation of the output can be a little nuanced. Based on our

[3] The Open Science Framework project link is https://osf.io/bx72m

previous discussion of probit model, the intuitive interpretation would be to use τ_0 to calculate the average predicted probability of adherence such that $\Phi(-\tau_0) = \Phi(1.28) = 0.90$. However, it is important to note that DSEM in M*plus* uses a *theta* parameterization for probit models, which means that the within-person residual variance is constrained to 1 but that there is no set constraint on the variance of the underlying process $y*$. This is relevant in DSEM because there are almost always predictor variables and between-person variances. So, to the extent that predictors explain variance in the outcome and there is non-null between-person variance, the variance of $y*$ will be greater than 1. If the variance of $y* > 1$, then $y*$ will not necessarily follow a standard normal distribution, which has ramifications for calculating predicted probabilities because $\Phi(\cdot)$ is only appropriate for standard normal distributions.

Nonetheless, there is a straightforward solution - namely, $\Phi(\cdot)$ can still be used but its argument (the number within parentheses) needs to be divided by the standard deviation $y*$, denoted $SD(y*)$. Of course, the next question is how to calculate $SD(y*)$. The standard deviation of $y*$ is $\sqrt{Var\left(y^{*(w)}\right) + Var\left(y^{*(b)}\right)}$, the square root of the sum of the within-person and between-person underlying process variances. These numbers do not appear directly in the M*plus* output, but they can be obtained without too much effort. The easiest way to calculate $Var(y^{*(w)})$ from the M*plus* output is $Var(y^{*(w)}) = 1/\left(1 - R_w^2\right)$ where R_w^2 is the within-person R^2 provided at the bottom of the standardized estimates in M*plus* (the standardized estimates must be requested in the `OUTPUT` statement). For $Var(y^{*(b)})$, the formula is similar but substitutes the intercept variance estimate in the numerator such that $Var\left(y^{*(b)}\right) = \sigma_{00}/\left(1 - R_b^2\right)$ where R_b^2 is the between-person R^2 provided at the bottom of the standardized estimates in M*plus*. If there are no between-person predictors, M*plus* will not report R_b^2 and this formula reduces to $Var\left(y^{*(b)}\right) = \sigma_{00}$. Together, this will account for all explained and unexplained sources of variance for $y*$ so that predicted probabilities can be accurately calculated.

In this example, $R_w^2 = 0.253$, so $Var(y^{*(w)}) = 1/\left(1 - 0.253\right) = 1.34$ and there are no between-person predictors, so $Var\left(y^{*(b)}\right) = 0.91$. Therefore, $SD(y*) = \sqrt{Var\left(y^{*(w)}\right) + Var\left(y^{*(b)}\right)} = \sqrt{1.34 + .91} = 1.50$. Finally, the interpretation of the threshold would be: for a person whose underlying normal process was at their person mean at time $t - 1$ (i.e., $y_{t-1,i}^{*(w)} = 0$), the average predicted probability of adherence is $\Phi[-\tau_0/SD(y*)] = \Phi(1.28/1.50) = \Phi(0.85) = 1.50$. This value is near the descriptive mean adherence proportion of 0.77. Essentially, latent centering results in the intercept referring to when the person was at their typical value for the underlying process of Adherence at time $t - 1$ rather than if Adherence itself was 0 at time $t - 1$. Note that the predicted probability calculation multiplies τ_0 by -1 because the interest is modeling the probability that Adherence =1, which requires the intercept rather than the threshold.

There are no p values in Bayesian models [4], but the Bayesian analog of the frequentist idea of statistical significance can be determined from inspecting whether the 95%CrI contains 0. On the probit scale, a 0 value for the intercept implies a predicted probability of 50%.

The 95%CrI for the threshold is [−1.69,−0.87], meaning that the predicted probability of Adherence for a person where $y_{t-1,i}^{*(w)} = 0$ is different from 50%. The 95%CrI for the intercept (rather than the threshold) could be determined by reversing the sign and order the upper and lower limits: [0.87, 1.69].

There is large between-person variability in the intercept - the estimate is 0.91 and the 95%CrI is [0.36,1.78], which clearly does not contain 0. Note that using an inverse Wishart prior distribution prohibits random effect variances from being negative, so assessing whether 0 is within the CrI is less objective for variance parameters because 0 will always be outside the CrI (e.g., McNeish & Hamaker, 2020, p. 614). Nonetheless, the lower bound of the CrI remains far from zero, so we are comfortable concluding that the intercept has non-negligible between-person variance despite the diminished utility of using CrIs for inference in this situation. Assuming normality of the random effects (as in Eq. 5) implies that most person-specific intercepts across people will fall between $1.28 \pm 1.96\sqrt{0.91} = [-0.59,\ 3.15]$. Again, these intercept values are on the probit scale, so we can convert them to predicted probabilities by evaluating the standard normal cumulative distribution function at these values, $\Phi(3.15/1.50) = 0.98$ and $\Phi(-0.59/1.50) = 0.35$, which corresponds to the descriptives whereby some people nearly always adhered to the intervention and others rarely did.

The autocorrelation fixed effect is 0.32 and zero is not within the CrI, meaning that there is a non-null autocorrelation between the underlying normal process of Adherence at time t and time $t - 1$. Additionally, note that the between-person variance in the tetrachoric autocorrelation is non-null and that zero is well outside the 95% CrI. This indicates that there is heterogeneity in the strength of the autocorrelation between the underlying process of Adherence at time t and time $t - 1$ across people. The random effect covariance between intercepts and tetrachoric autocorrelation is null, indicating that there does not appear to be a systematic relationship between the person-specific intercepts and person-specific tetrachoric autocorrelations.

Example 2: Probit DSEM for a binary outcome with a continuous covariate

The previous section covered the basics of DSEM in M*plus* for an unconditional model with a binary outcome and no covariates. Here, we cover considerations when a continuous covariate is included to predict the outcome. In the motivating data, each morning prior to participants having a chance to adhere to the intervention, participants were also asked on a 0–10 scale “how motivated are you to avoid binge eating today” where 10 indicated being the most motivated ($M = 7.38,\ SD = 2.44$). We can include this variable in the model to assess whether motivation to avoid binge eating predicts whether participants will adhere to the intervention. The covariate was measured at each day, so it is a *time-varying* covariate rather than a *time-invariant* covariant (the latter would be a variable that differs across people but is constant across time for a particular person). The distinction between time-varying and time-invariant covariates is based upon the frequency of data collection rather

[4] M*plus* does provide a column with a one-tailed *p* value in its default output. However, the interpretation of this value does not coincide with the interpretation provided by a traditional frequentist *p* value. The Bayesian *p* value reported in M*plus* corresponds to the proportion of the posterior distribution on the opposite side of 0 than the posterior summary (the “Estimate” column in M*plus*). For example, a value of 0.03 for a positive estimate would mean that 3% of the posterior distribution is below 0 (Muthén, 2010 p. 7).

than the variable itself. For instance, binge eating avoidance could be time-invariant if it were only collected once at baseline as opposed to being collected every day, as in the motivating study.

In multilevel time-series analysis (like DSEM), it is important to person-mean center time-varying covariates to disaggregate temporary states from stable traits. For example, imagine someone reports a “9” value for binge eating avoidance at day 2. Solely from this information, we could not determine if this person was generally avoidant and was reporting typical behavior or whether this person is reporting much higher avoidance relative to their typical behavior. Raw variable values conflate the state- and the trait-level information of a behavior into a single value (Hoffman & Walters, 2022). Person-mean centering can be used to calculate the mean of binge eating avoidance across all timepoints for each person as an estimate of trait-level avoidance. Each day’s difference from the person mean then provides a state-level estimate of avoidance relative to the person’s typical level. This changes the scaling of the time-varying covariate such that “0” means the person is at their typical value rather than at the lowest value on the scale. When person-mean centering covariates in multilevel time-series analysis, users have two options for how to calculate the person-mean - (a) observed-mean center around the descriptive mean or (b) latent-mean center treating the trait-level component as a latent variable. By default, M*plus* will latent-mean center covariates to accommodate measurement error, unreliability, or missing data.

Model equations and path diagram

When extending Eq. (5) to include a covariate, the model can be written as

$$
\begin{array}{ll}
\text{Latent-Decomposition} & \begin{cases} \Pr(Adhere_{ti} = 1) = \Phi\left(y_{ti}^{*}\right) \\ y_{ti}^{*} = y_{ti}^{*(w)} + \alpha_i \\ BEA_{ti} = BEA_{ti}^{(w)} + BEA_i^{(b)} \end{cases} \\
\text{Within- Person} & \begin{cases} y_{ti}^{*(w)} = \phi_i y_{t-1,i}^{*(w)} + \beta_i BEA_{ti}^{(w)} + e_{ti} \\ BEA_{ti}^{(w)} \sim \mathcal{N}(0, \omega) \\ e_{ti} \sim \mathcal{N}(0, 1) \end{cases} \\
\text{Between-Person} & \begin{cases} \alpha_i = -\tau_0 + \gamma_{01} BEA_i^{(b)} + u_{0i} \\ \phi_i = \gamma_{10} + u_{1i} \\ \beta_i = \gamma_{20} + u_{2i} \\ BEA_i^{(b)} = \gamma_{30} + u_{3i} \\ \begin{bmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \\ u_{3i} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & & & \\ \sigma_{10} & \sigma_{11} & & \\ \sigma_{20} & \sigma_{21} & \sigma_{22} & \\ 0 & 0 & 0 & \sigma_{33} \end{bmatrix} \right) \end{cases}
\end{array}
\tag{6}
$$

The latent decomposition in Eq. (6) includes the two expressions in Eq. (5) and now also includes the continuous covariate binge eating avoidance (abbreviated BEA) because the model will disaggregate its effect into within-person and between-person components. Similar to Eq. (5), $BEA_{ti}^{(w)}$ represents the state component of binge eating avoidance whereas $BEA_i^{(b)}$ represents the trait component. The within-person model in the second set of expressions in Eq. (6) is similar to Eq. (5) except that there is now a new effect β_i that captures how binge eating avoidance affects the underlying process of Adherence. Note that binge eating avoidance has a w superscript to indicate that it is the within-person component of binge eating avoidance that has been centered around its latent mean (i.e., the third expression could be rearranged such that $BEA_{ti}^{(w)} = BEA_{ti} - BEA_{i}^{(b)}$). Because binge eating avoidance is continuous, the within-person variance around the latent mean is quantified by ω.

The third set of expressions corresponds to between-person relationships. First, note that $BEA_{i}^{(b)}$ has its own equation in the between-person model because the person-mean is modeled as a latent variable (with a mean γ_{30} and a variance σ_{33}) rather than being a descriptive calculation from the data. Also note that $BEA_{i}^{(b)}$ also appears as a predictor of the intercept in the α_i equation, meaning that the baseline probability of adherence is being modeled as a function of the person's trait-level binge eating avoidance. This also means that $-\tau_0$ is conditional and corresponds to when $BEA_{i}^{(b)} = 0$.

Including binge eating avoidance in the within-person and between-person models allows the effect to be fully disaggregated so that the model can differentiate the effects of being *momentarily* avoidant and being *habitually* avoidant. The momentary effect of binge eating avoidance (β_i) is modeled as person-specific, meaning that the effect is allowed to be different across people. The random effect covariances of α_i and $BEA_{i}^{(b)}$ are constrained to 0 in Eq. (6) to avoid redundancy with parameters in the model (e.g., $BEA_{i}^{(b)}$ predicts α_i, so u_{3i} should not also covary with the random effect for the intercept u_{0i}).[5] All covariances between other random effects are estimated. Do note that random effect covariances are the least stable parameters to estimate (Kretzschmar & Gignac, 2019) and may present difficulties with modest samples or as models become more complicated.

Figure 4 shows a path diagram for the model in Eq. (6). The left panel shows the latent decomposition of Adherence (through its underlying normal process, y^*) and binge eating avoidance. All within-person components are constrained to have a mean of 0 because they are centered around the latent person mean (i.e., the average across all state components equals the trait component). The within-person component of binge eating avoidance is a predictor of the underlying process of Adherence, and the effect is modeled with a latent variable because the value of this path varies across people. [6] In the between-person model, the latent mean of binge eating avoidance predicts the person-specific intercept.

[5] Only the covariance between the intercept of the outcome and the trait-like component of the covariate $BEA_{i}^{(b)}$ must be constrained to 0. The other covariances involving $BEA_{i}^{(b)}$ could theoretically be estimated, but the full covariance would no longer be block diagonal, which is not supported by the Gibbs sampler in M*plus* (Asparouhov & Muthén, 2010). A random walk algorithm suggested by Chib and Greenberg (1998) can support arbitrary covariance structures and can be implemented in M*plus* by specifying `ALGORITHM=GIBBS(RW).` This algorithm does not support multivariate priors like inverse Wishart and can be less efcient that the default Gibbs sampler. When we applied this method, there was poor mixing even with millions of iterations, so we elected to use the M*plus* default sampler without estimating these two covariances.

[6] This model considers binge eating avoidance as a *contemporaneous* efect of Adherence such that the covariate collected at time t predicts an outcome also collected at time t. This was done because the covariate was collected before the outcome on each day, so there is no ambiguity about temporal precedence. However, covariates can also be *lagged* efects if the hypothesized effect is thought to take more time to unfold (e.g., binge eating avoidance yesterday predicts Adherence today) or to delineate between the cause and efect more clearly if one variable was not necessarily collected frst within time t. In such case, autoregression in the covariate may be added to the model.

M*plus* code and model fitting

The annotated code for fitting the conditional model in Eq. (6) is shown below. The code is similar to the code presented in Example 1 but has a few notable differences. First, the within-person model includes a second random slope for the effect of within-person binge eating avoidance on Adherence. Because `bea` is not included in a `WITHIN` option in the `VARIABLE` statement, it will automatically be latent centered. In the between-person portion of the `MODEL` statement, the latent person mean of binge eating avoidance predicts the person-specific intercept. The model was fit in M*plus* 8.7 using the same options described in the previous example.

Results and interpretation

The posterior medians and the 95% CrIs for the model are shown in Table 2. In this section, we focus on parts of the model that are new or have changed once the binge eating avoidance covariate was added to the model. The binge eating avoidance latent mean (γ_{30}) is 7.22, which means that the average person's mean of binge eating avoidance is 7.22 (on the original 1–10 scale of the outcome). There is variability in the latent person means as noted by the large variance ($\sigma_{33} = 3.06$), meaning that different people have different trait-level binge eating avoidance.

```
VARIABLE:
USEVARS ARE adhere bea
CATEGORICAL=adhere ; !specify variables that are categorical;
TINTERVAL = day(1); !specify the largest increment of time that can have
only one observation;
CLUSTER= id; !repeated measures are clustered within ID;
MISSING ARE .; !identify missing data code;
LAGGED= adhere(1); !create lag-1 predictor adherence;
ANALYSIS:
TYPE= TWOLEVEL RANDOM; !Two-Level Model with random effects;
ESTIMATOR=BAYES;!Lagged variables can only be estimated with Bayes in Mplus;
BITERATIONS= (5000); !Run at least 5000 iterations of the MCMC algorithm;
MODEL:
%WITHIN%
phi|adhere ON adhere&1;
!Adherence yesterday predicts Adherence today, phi_i;
!”phi|” adds a random effect so this effect is allowed to vary across people;
beta|adhere on bea; !BEA predicts Adherence at the same day, beta_i;
!”beta|” gives this path a random effect so it is allowed to vary across
people;
%BETWEEN%
adhere; !between-person variance in adherence intercept, sigma_00;
phi; ! between-person variance in autoregreesion, sigma_11;
beta; !between-person variance of BEA effect, sigma_22;
```

```
bea; !between-person variance of BEA latent mean, sigma_33;
adhere phi WITH adhere phi beta ; !between-person random effect covariances,
off-diagonal sigma elements;
[adhere$1]; !fixed effect of threshold, tau_0);
[phi]; !fixed effect of autoregression, gamma_10;
[beta]; !fixed effect of within-person BEA effect, gamma_20;
[BEA]; !intercept for latent mean of BEA (the mean of the person means),
gamma_30;
adhere on bea; !fixed effect of BEA latent mean on adherence intercept,
gamma_01;
OUTPUT: STDYX;
```

The effect of latent centered binge eating avoidance in the within-person model (γ_{20}) corresponds to the change in the predicted probability of adherence when a person is *momentarily* one point above the person mean (a 1-point increase has the same meaning as it did on the original 0–10 scale). This estimate is 0.06 but zero is within the CrI $(95\%\ \text{CrI} = [-0.07,\ 0.19])$, meaning that moment-to-moment changes in reported binge eating avoidance do not seem to meaningfully affect whether a person adheres to the intervention, on average. However, this effect varies across people $(\sigma_{22} = 0.08,\ 95\%\ \text{CrI} = [0.04,\ 0.15])$, so it is possible that momentarily higher binge eating avoidance does affect the probability of adherence for some people.

If we wanted to inspect the within-person effect of binge eating avoidance for each person, we can create a distribution of *plausible values* for each person (Asparouhov & Muthén, 2010). This is similar to factor scoring a latent variable in a frequentist setting whereby a value of the latent variable is predicted for each person. However, consistent with Bayesian philosophy, an entire distribution of possible factor scores is predicted (e.g., Rubin, 1996). This essentially treats the latent variable as a big multiple imputation problem where the latent variable is missing for all people (Enders, 2010; Mislevy & Sheehan, 1989). In M*plus*, a dataset of plausible values can be created by adding the following line of code to the input file,

```
SAVEDATA: FILE IS scores.dat;
SAVE ARE FSCORES (200);
```

This will create a dataset called “scores” (the name is arbitrary and can be changed) in the same folder as the input file (a path directory can also be specified to save the file elsewhere). To that dataset, M*plus* will save the mean, median, standard deviation, and 2.5 and 97.5 percentile of the distribution of plausible values for each person. The number in parentheses is the number of plausible values to create for each latent variable; we use 200 here and suggest using at least 100 plausible values. Increasing this number will increase the precision of the results, but the tradeoff is higher computational time. From this information, researchers can assess whether 0 is within the 95% CrI (formed by the 2.5 and 97.5 percentiles) to determine if the effect is non-null *for each person*. This information can be summarized in a caterpillar plot as in Fig. 5 below. We can see that the within-person binge eating avoidance effect is mostly null across the sample but there are four people in the data where the effect appears to be positive and non-null (though this may just be chance). If heterogeneity in this effect were a research interest, time-invariant predictors could be added to the β_i equation in the between-person model to explain the source of the heterogeneity.

If it is more desirable to plot the standardized values in a caterpillar plot, the within-person standardized estimates (as recommended by Schuurman et al., 2016) can be requested. Within-person standardization first standardizes coefficients for each person using the person-specific variances of the time-varying covariate and the outcome and then takes the mean across all values of all people to arrive at a single standardized value. The following M*plus* code at the end of the input file will output these standardized values,

```
OUTPUT: STDYX;
SAVEDATA: STDRESULTS ARE stand.dat;
```

This will create a dataset called “stand” (the name is arbitrary and can be changed) in the same folder as the input file (a path directory can also be specified to save the file elsewhere) with estimates and credible intervals for each person. For each person in the data, this file will have multiple rows – one for each within-person standardized effect in the within-person model and one for the R^2 of each within-person outcome. For the model in Eq. (6), this results in five rows per person – three within-person effects (ϕ_i, β_i, and ω) and two within-person R^2 values (Adhere and BEA). We do not report the caterpillar plot for the standardized effect because it largely mirrors Fig. 5.

Because we disaggregated binge eating avoidance, in addition to the within-person momentary effect, we also have the between-person effect to assess whether people who are habitually avoidant have different probabilities of intervention adherence. This is captured by γ_{01}, which is 0.23 and zero is not in the CrI $(95\%\ \mathrm{CrI} = [0.01,\ 0.43])$. The positive coefficient means that people who have higher trait-level values of binge eating avoidance are predicted to have a higher probability of adhering to the intervention.

Recall that probit models are linear in the underlying normal process associated with probabilities, but are nonlinear in the probability itself. Therefore, the intercept (and values of other covariates if there are more than one covariate) need to be included to accurately capture how changes in the underlying process change the predicted probability. For instance, the interpretation of this coefficient would *not* be that a 1 -unit increase in the latent mean of binge eating avoidance predicts an increase in the probability of adhere of $\Phi(\gamma_{01}) = \Phi(0.23) = 0.59$. Instead, the proper interpretation of a one-unit change in the latent mean of binge eating avoidance would depend on the value of binge eating avoidance because changes in probabilities are nonlinear.

The fixed effect for the intercept is $-\tau_0 = -0.25$, $R_w^2 = 0.353$, $R_b^2 = 0.144$, and $\sigma_{00} = 0.94$; meaning that $SD(y^*) = \sqrt{1.55 + 1.10} = 1.63$ and the predicted probability of adherence is

$\Phi(-0.25/1.63) = 0.44$, which falls well below the descriptive adherence probability of 0.77. This discrepancy is due to the fixed effect being conditional on the person mean being 0 whereas the average is closer to 7. Table 3 shows the predicted probabilities based on different person mean values of binge eating avoidance using the estimates in Table 2. A 1-point change in the person mean roughly equates to a 5% increase in the predicted probability for most of the scale but the increase tapers off towards the upper extreme. Note that when $BEA_i^{(b)}$ is at 7 and near the sample average, the predicted probability of adherence is 0.79 and near the sample average of 0.77. Given that the between-person effect of binge eating avoidance is non-null, there is a noticeable change in the predicted probability of adherence as $BEA_i^{(b)}$ changes. Overall, the model implies that increases in trait-level binge eating avoidance – but not increases in momentary state-level changes in binge eating avoidance – are related to higher intervention adherence.

Example 3: Probit residual DSEM for a binary outcome with a time trend

As noted at the beginning of this paper, autoregressive models assume *stationarity* which requires that the expected value does not systematically increase or decrease over time. In the motivating data, a violation of stationarity may occur if, for instance, people systematically adhered to the intervention less at the end of the study because they grew tired of participating. Systematic time trends (e.g., linear increases or decreases, cyclical changes like day of the week or time of day) in the raw outcome do not necessarily preclude using autoregressive models entirely because the assumption can still be satisfied as long as time trends are explicitly modeled.

Residual DSEM (RDSEM; Asparouhov et al., 2017) is one such model that can incorporate time trends for intensive longitudinal data. In an RDSEM, the autoregression occurs in the *residuals* of the outcome after accounting for a time trend rather than autoregressing the variable itself (Asparouhov & Muthén, 2020, 2022a). The location of the autoregression is moved to the residuals because– after accounting for systematic trends related to time – they are independent of systematic time trends. In other words, including time as a predictor produces detrended residuals, as long as the time trend has been properly modeled. This effectively relocates the stationarity assumption from the variable to the residuals. Therefore, the residuals can satisfy stationarity assumptions, even in the presence of a time trend in the variable.

In the context of the motivating data, the top panel of Fig. 6 shows a plot of the raw proportion of participants adhering to the intervention across the 28 days of the study. The grey dashed horizontal line is the mean across all days and the black line is a linear trend in the proportions. If the series were stationary and mean reverting, the solid black and dashed grey lines would overlap. The trend is not overwhelming, but the solid black line is discernibly different from the horizontal grey dashed line, so it may be worth considering a model that autoregresses detrended residuals to better adhere to stationarity. The bottom panel of Fig. 6 shows a plot of the detrended proportions with the linear effect of time removed. These data look more convincingly mean- reverting such that there are no systematic increases or decreases and the black and grey lines completely overlap. As an informal graphical explanation, detrending essentially rotates the entire data series in the top panel counterclockwise until the black line and grey line are indistinguishable.

This idea can be implemented as an RDSEM to model intensive longitudinal data that may have a systematic time trend. Equation (7) shows the model equations for an unconditional probit model with a linear trend for Day. The path diagram is shown in Fig. 7.

$$
\begin{array}{ll}
\text{Latent-Decomposition} & \begin{cases} \Pr[Adhere_{ti} = 1] = \Phi\left(y_{ti}^{*}\right) \\ y_{ti}^{*} = y_{ti}^{*(w)} + \alpha_i \end{cases} \\
\text{Within-Person} & \begin{cases} y_{ti}^{*(w)} = \gamma_{01} Day_{ti} + d_{ti} \\ d_{ti} = \phi_i d_{t-1,i} + e_{ti} \\ e_{ti} \sim \mathcal{N}(0, 1) \end{cases} \\
\text{Between-Person} & \begin{cases} \alpha_i = -\tau_0 + u_{0i} \\ \phi_i = \gamma_{10} + u_{1i} \\ u_i \sim \mathcal{N}(0, \Sigma) \end{cases}
\end{array}
\tag{7}
$$

The main difference in the RDSEM model in Eq. (7) occurs in the first and second expressions in the within-person model. Specifically, note that $y_{ti}^{*(w)}$ is no longer autoregressed on itself – instead, it is predicted by Day in the first expression. In Eq. (7), Day is modeled as a fixed effect, but it is possible to model this effect as person-specific if the time trend is thought to vary across people (adding a random effect to the time trend would allow the coefficient – but not the functional form – to change across people). Also note that the residual in the first expression is denoted by d rather than e, where d is the detrended residual associated with $y_{ti}^{*(w)}$. The detrended residual d appears as an outcome in the second expression where it is autoregressed on d at time $t - 1$. In the second expression, ϕ_i is the autocorrelation in the detrended residuals of $y_{ti}^{*(w)}$, rather than the autocorrelation of $y_{ti}^{*(w)}$ itself. The e term in the second expression is then the residual term in the autoregression equation for d. That is, the autoregression is on the detrended residuals of $y_{ti}^{*(w)}$ but this autoregression is not perfect and therefore has another residual e, which is assumed to follow a standard normal distribution. As a result, the variance of d is not fixed to a particular value and the residual variance of $y_{ti}^{*(w)}$ is instead a function of the strength of the autocorrelation, $1/\left(1 - \phi_i^2\right)$. This means that the model has a theta parameterization such that the residual variance is not constrained to be exactly equal to 1 (Asparouhov & Muthén, 2020, p. 285). Just like DSEM, this unconditional RDSEM could be also expanded to include covariates.

The code for fitting this model in M*plus* is similar to the code in Example 1 with a few changes. First, `day` is added to after a `WITHIN` = option in the `VARIABLE` statement because it is only used as a predictor in the within-person model and we do not need to disaggregate it (i.e., there are no trait and state components of time). The `%WITHIN%` section of the code changes slightly to

```
%WITHIN %
```

```
Phi| adherê ON adherê1;
adhere on day;
```

The caret (“^”) replaces the ampersand (“&”) in the code to tell M*plus* that the autoregression involves the *residual* of `adhere,` not `adhere` itself. The last line then includes a linear trend for `day` to model the possible trend over time. The full M*plus* input file is included on the Open Science Framework page associated with this paper.

After fitting the model with the same options as in Examples 1 and 2, the estimated fixed effect for the threshold τ_0 is now −1.48 rather than −1.28 as in the unconditional model in Example 1. $R_w^2 = 0.259,\ \sigma_{00} = 1.08$ and there are no between-person predictors, so $SD(y^*) = 1.56$. Converting this to a predicted probability yields $\Phi[-\tau_0 / SD(y^*)] = \Phi(1.48/1.56) = 0.83$. This is slightly higher than then 0.80 predicted probability in Example 1 because it corresponds only to the first day of the study rather than to the entire study window (similar to the trend line in the top panel of Fig. 6 at Day = 0). The linear effect of Day is $\gamma_{10} = -0.017$ which means that for each additional day of the study, the predicted probability of adherence on the y^* scale decreases by 0.017 and adherence is less likely as the study progresses. For example, the predicted probability of adherence on day 28 of the study (when coding the first day as day 0) is, $\Phi[(-\tau_0 + 27 \times \gamma_{10})/SD(y^*)] = \Phi[(1.48 + 27 \times -0.017)/1.56] = \Phi(1.02/1.56) = 0.74$, which corresponds to the trendline in the top panel of Fig. 6. The 95% CrI for the linear effect of Day barely does not include 0, $(95\%\ \text{CrI} = [-0.032,\ -0.002])$, so there is borderline evidence that the linear trend of Day is non-null and may need to be accounted to satisfy stationarity assumptions. Alternatively, including the linear trend of Day could be used as a sensitivity analysis to assess whether conclusions would change depending on potential stationarity violations.

Example 4: Cumulative probit DSEM for ordinal outcomes

When discussing binary probit DSEM in previous sections, we parameterized the fixed effect for the intercept as the opposite of the threshold such that $\alpha_i = -\tau_0 + u_{0i}$. This parameterization results in probit models that look similar to models for continuous outcomes (and matches M*plus* output). However, this parameterization does not extend to ordinal data with three or more categories, which require multiple thresholds. In this section, we discuss a more general parameterization of binary probit models and how this more general version can be extended to ordinal data. We also provide an example to discuss differences in interpretation when DSEM models are applied to ordinal versus binary outcomes.

General form of the binary probit model

In earlier sections, we discussed thresholds and intercepts with binary outcomes as essentially interchangeable aside from multiplying by −1. However, the intercept and threshold technically correspond to different quantities in a probit model (Long, 1997, p. 122). Specifically, the intercept is the mean of the underlying normal distribution y_{ti}^* whereas the threshold dictates which values of y_{ti}^* manifest as different observed values of y_{ti}. A model estimating the intercept and threshold(s) simultaneously is unidentified because there would be more latent quantities than observed information (Long, 1997, p. 123). Therefore, either the intercept or threshold must be constrained (often arbitrarily) for identification. Though many possibilities exist, there are two primary approaches.

In the first approach, the threshold is constrained to 0 and the fixed effect of the intercept is estimated. This is shown in the top panel of Fig. 8 using estimates from Table 1 where the intercept and threshold were 1.28 units apart. In the second approach, the intercept fixed effect could be constrained to 0 and the threshold is estimated. This is shown in the bottom panel of Fig. 8. Both approaches give equivalent predicted probabilities (Long, 1997, p. 122; the area to the left of τ in either panel of Fig. 8 is 0.20), they just represent different identification strategies because the scaling of the underlying normal process is arbitrary given that it is latent.

Up to this point, we blended these different approaches to simplify interpretations by essentially using the mirror image of the bottom panel of Fig. 8. That is, when discussing previous models, we kept the mean of the underlying normal distribution at 0 (i.e., Eq. 3) and estimated the intercept fixed effect (Eqs. 5–7) as the opposite of the threshold. This scaling approach makes the model look more like a continuous DSEM and simplifies calculations of the predicted probabilities. Although this is permissible to simplify interpretations with binary outcomes, the same simplification is not effective for ordinal models where multiple thresholds are present to establish three or more categories (Asparouhov et al., 2018, p. 363). For ordinal outcomes, the number of thresholds is equal to the number of categories minus one, which makes the distinction between the intercepts and thresholds more meaningful because multiple thresholds are present. The next subsection presents an example with an ordinal outcome to demonstrate these differences.

Extending probit models to ordinal outcomes

The motivating data contain a daily five-point Likert response to the prompt “Tempting food made it difficult for me to not binge eat today” (referred to as “Temptation” hereafter). Across all 50 participants and timepoints, the responses were roughly symmetric with a slight negative skew: 18% strongly disagree, 16% disagree, 24% neutral, 27% agree, and 16% strongly agree. Even though some studies suggest that five timepoints may be sufficient to treat a response as continuous (e.g., Rhemtulla et al., 2012; Robitzsch, 2020), we will model this item as ordinal to avoid any possible issues with treating Likert scales as continuous such as ceiling or floor effects, distributional assumption violations, or assumptions that sequential categories represent equal changes in the underlying construct (Hamaker et al., 2023; Haqiqatkhah et al., 2022; Liddell & Kruschke, 2018). We use the same binge eating avoidance covariate in the model as well to provide some guidance on covariate interpretation.

The major change for ordinal outcomes is that the thresholds appear in the definition of y_{ti}^{*} rather than in the α equation in the between-person model. So, whereas Eq. (3) used “0“ in the inequalities to specify how y_{ti}^{*} manifests into observed categories, Eq. (8a) uses thresholds:

$$Temptation_{ti} = \begin{cases} 5 \text{ if } y_{ti}^{*} \geq \tau_4 \\ 4 \text{ if } \tau_3 \leq y_{ti}^{*} < \tau_4 \\ 3 \text{ if } \tau_2 \leq y_{ti}^{*} < \tau_3 \\ 2 \text{ if } \tau_1 \leq y_{ti}^{*} < \tau_2 \\ 1 \text{ if } y_{ti}^{*} < \tau_1 \end{cases} \tag{8a}$$

This is the notable divergence between ordinal probit models and how we parameterized binary models in earlier sections. Whereas binary models have one threshold and one intercept, ordinal models have one intercept but multiple thresholds. It is less intuitive to determine constraints when there are multiple thresholds in ordinal models, so it is easier to constrain the lone intercept to have a fixed effect of 0 and estimate the thresholds.

Calculating predicted probabilities for each category is then expanded. The first and last categories are similar to the latent decompositions in Eqs. (5–7) in that they only have one term (e.g., any portion of y_{ti}^{*} below τ_1 manifests as Temptation =1, so there is no lower bound for values of y_{ti}^{*} that result in Temptation =1). However, middle categories require subtraction because thresholds are *cumulative*. That is, the total area to the left of τ_3 captures the probability of responding 1,2, or 3 rather than just the probability of responding 3. Therefore, the probability of all lower categories must be subtracted to isolate the probability of any middle category. More formally, the predicted probability calculations in ordinal models would be expanded to

$$\begin{aligned} &\Pr[Temptation_{ti} = 5] = \Phi\left(-\tau_4 + y_{ti}^{*}\right) \\ &\Pr[Temptation_{ti} = 4] = \Phi\left(\tau_4 - y_{ti}^{*}\right) - \Phi\left(\tau_3 - y_{ti}^{*}\right) \\ &\Pr[Temptation_{ti} = 3] = \Phi\left(\tau_3 - y_{ti}^{*}\right) - \Phi\left(\tau_2 - y_{ti}^{*}\right) \\ &\Pr[Temptation_{ti} = 2] = \Phi\left(\tau_2 - y_{ti}^{*}\right) - \Phi\left(\tau_1 - y_{ti}^{*}\right) \\ &\Pr[Temptation_{ti} = 1] = \Phi\left(\tau_1 - y_{ti}^{*}\right) \end{aligned} \tag{8b}$$

As in Eqs. (5–7), the underlying normal process y_{ti}^{*} is still partitioned into within-person and between-person components such that $y_{ti}^{*} = y_{ti}^{*(w)} + \alpha_i$. The only difference is that the fixed effect in the α_i expression is fixed to 0 given that the thresholds are estimated. However, there can still be between-person variance in the intercept by including a random intercept such that $\alpha_i = u_{0i}$. So, the *average* intercept is constrained to 0 but the intercept for person i is not necessarily 0. Additionally, time-invariant covariates can still be included to explain systematic between-person reasons why the intercept may differ across people. Note that between-person variance is modeled with one random effect on the intercept rather than through multiple random effects for each threshold, which makes an implicit assumption that the spacing between thresholds remains the same across all people. That is, the thresholds are not necessarily equidistant, but if the distance from τ_1 to τ_2 is 0.50 and the distance from τ_2 to τ_3 is 0.25, those relative distances will be maintained across all people (i.e., there are no i subscripts on the thresholds).

Otherwise, the rest of the model for an ordinal outcome is unchanged compared to a binary outcome. Models for binary and ordinal outcomes both operate on the underlying normal process y_{ti}^{*}, so defining how y_{ti}^{*} manifests into observed data is the only major change in the model specification between binary versus ordinal models and all other portions of the model are identical. So, if we are trying to model how binge eating avoidance affects the dynamics of the ordinal Temptation variable, the model would look almost identical to Eq. (6) for the binary Adherence variable. The only difference is that the first expression in the Latent Decomposition section would be expanded with Eq. (8b) and the $-\tau_0$ term would be removed from the between-person α_i expression based on the ordinal definition of y_{ti}^{*} in Eq. (8a). Similarly, the M*plus* code for an ordinal model with a disaggregated continuous covariate model looks no different from the code associated with the model in Eq. (6) other than changing the outcome variable (though not required, if being explicit, this model would have more thresholds denoted by including the outcome label followed by `$2,$3y,` and `$4` in square brackets). We do not provide the full code in text, but it is available on the Open Science Framework page for this paper. The results from fitting this model (with the same estimation criteria as previous examples) are shown in Table 4. To simplify the interpretation, we rescaled the binge eating avoidance covariate by subtracting 7.22 (its mean in Table 2). We still latent center the covariate, but this rescaling will make the thresholds correspond to the typical person rather than a someone whose latent person mean of binge eating avoidance is 0.

Results and interpretation

To highlight differences between this analysis and the earlier binary outcome analyses, there are now four thresholds which delineate which values of y_{ti}^{*} manifest as which discrete Likert responses. The estimated thresholds are $\tau_1 = -1.16,\ \tau_2 = -0.50,\ \tau_3 = 0.26$, and $\tau_4 = 1.28$. In this model, $R_w^2 = 0.203,\ R_b^2 = 0.018$, and $\sigma_{00} = 0.49$, meaning that $SD(y^*) = 1.32$. If applying the standard normal cumulative distribution function to these thresholds divided by 1.32 to rescale onto the standard normal metric, the corresponding predicted probabilities for the respective Likert categories are 19%, 16%, 22%, 26%, and 17%, which resemble the descriptive proportions of 18%, 16%, 24%, 27%, and 16%.

The top panel of Fig. 9 shows how y_{ti}^{*} is discretized to manifest into Likert responses, where the estimated thresholds are represented by vertical dashed lines. The mean of y_{ti}^{*} in the top panel of Fig. 9 is 0, which represents the average across all people. There is non-negligible between-person intercept variance ($\sigma_{00} = 0.49,\ 95\%\ \mathrm{CrI} = [0.27,\ 0.78]$), meaning that the location of the y_{ti}^{*} distribution shifts horizontally across people. For instance, the middle panel of Fig. 9 shows a hypothetical person whose intercept for Temptation is one standard deviation below the sample average (i.e., the mean of $y_{ti}^{*} = -0.70$). The threshold values have not changed, but y_{ti}^{*} has shifted left such that this person has a much higher probability of responding towards the lower end of the Likert scale (e.g., the area under the curve to the left of the leftmost threshold has expanded in the middle panel relative to the top panel). If interpreting the model this way, calculations for predicted probabilities must adjust values so that the mean of y_{ti}^{*} is rescaled to 0 so that $\Phi(\cdot)$ can be applied, similar to how we have needed to adjust for the variance not being equal to 1.

An alternative interpretation of the between-person intercept variability is that the mean of y_{ti}^{*} stays at 0, but all thresholds move to the *right* by 0.70 if the person that is one standard deviation below average on Temptation such that their spacing is preserved (i.e., moving the intercept to the left or all the thresholds to the right produces the same effect; this relates back to multiplying by −1 when converting between thresholds and intercepts in earlier sections). This is shown in the bottom panel of Fig. 9 (this panel is offset to align the thresholds). In other words, $-u_{0i}$ can be interpreted as a shift parameter on the thresholds (Asparouhov et al., 2018) and this shift occurs in the location, but not the scale (i.e., the variance is constant, but the means can change). Note that the probabilities are equivalent between the middle and bottom panels of Fig. 9, but the bottom panel has the advantage that no mean adjustments are required before applying $\Phi(\cdot)$ to compute predicted probabilities. The point of presenting these different interpretations is to emphasize that y_{ti}^{*} is *latent* and has no inherently meaningful scale. The parameter effects dictate the relative distance and spacing, but the anchor point is arbitrary and is only needed for identification (similar to constraining a latent variable variance or a factor loading in standard SEM).

Turning to covariate effects, the within-person effect of binge eating avoidance in Table 4 was non-null ($\gamma_{20} = -0.10,\ 95\%\ \text{CrI} = [-0.19,\ -0.01]$), so we will discuss its interpretation here. It is important to remember that, in the model, the covariates affect α_i and $y_{ti}^{*(w)}$, not τ. This implicitly assumes that all categories are affected equally (i.e., the spacing between thresholds is constant across all covariate values), which is sometimes referred to as the *parallel regression* assumption (McCullagh, 1980). Specifically, because γ_{20} is a within-person effect, its interpretation is: for every one-unit increase in binge eating avoidance above the latent person mean, $y_{ti}^{*(w)}$ is predicted to decrease by 0.10. Graphically, this means that $y_{ti}^{*(w)}$ shifts left as $BEA_{ti}^{(w)}$ increases such that responses on the lower end of the Likert scale become more probable. Alternatively, an equivalent effect would be to shift all thresholds to the right by 0.10.

Figure 10 shows this effect graphically. In the top panel, the grey distribution is $y_{ti}^{*(w)}$ when the person is at the latent person mean for the covariate and the black distribution is how $y_{ti}^{*(w)}$ shifts when the person is 1-unit above the covariate latent person mean. The alternative interpretation is shown in the bottom panel. The location of $y_{ti}^{*(w)}$ is constant (the solid black line), the grey dashed lines are the thresholds when the person is at their covariate latent person mean, and the black dashed lines represent how the thresholds shift when a person is one-unit above their covariate latent person mean. The predicted probabilities are identical in either case. The effect is not overwhelmingly large, but the response probabilities for the lower end of the scale increase as within-person binge eating avoidance momentarily increases. Substantively, the interpretation would be that momentarily higher feelings of binge eating avoidance are associated with being slightly less tempted by food.

Discussion

Many existing treatments of dynamic structural equation models consider categorical outcomes as a straightforward extension of models for continuous outcomes. Although this may resonate with statisticians and more technically inclined readers, the nuances of interpreting models with binary or ordinal outcomes – even if the statistical foundations are straightforward – warrant dedicated resources for empirical researchers for whom the connections may not be obvious. Our goal in this paper was to provide an accessible overview of some foundational models for multilevel time-series analysis with categorical outcomes in M*plus* to provide readers with a starting point for approaching these models, especially because binary behavioral items and Likert item responses are common outcomes in intensive longitudinal studies.

Of course, this topic is broader than what could be covered in detail in a single paper and there are caveats and extensions to what was discussed. For instance, these models are appropriate for processes that are mean reverting or that can made mean reverting after conditioning on time trends. Outcomes featuring rare binary events that are nearly always 1 or nearly always 0 will have trouble satisfying stationarity assumptions and may not be suitable for these models. This issue may also be problematic with binary outcomes and shorter time series (e.g., less than 20 observations per person) if there is not enough variability in the outcome to reasonably identify moment-to-moment dynamics (Asparouhov & Muthén, 2022b). The models described in this paper are also only appropriate for outcomes that do not have absorbing states such that the outcome is permitted to take on any possible value at any point in time. This could be problematic in areas like developmental psychology where children may not revert back to particular behaviors once a certain developmental milestone has been achieved. If a “1” response precludes a participant from later responding “0”, other models like dynamic latent class analysis or time-to-event models may be more appropriate.

M*plus* does have some built-in safeguards for mild violations of these assumptions such as removing inadmissible values from posterior distributions should they arise. Anecdotally, the M*plus* development team had noted that if the MCMC estimation converges, then the number of discarded iterations will tend to be small (e.g., less than 10%; Asparouhov, 2020, http://www.statmodel.com/discussion/messages/24588/27731.html?1580727445) and that highly non-stationary processes will simply fail to converge in most cases. For instance, in the analysis for Example 2 where the mean prevalence was 0.77 from 28 time-points per person, the output showed a warning that some iterations were discarded from the posterior (likely stemming from the small number of people with very high adherence rates and whose data were largely noninformative). However, in general, researchers considering using DSEM or RDSEM for binary outcomes should first ensure that that there is sufficient variability in the binary outcome to warrant such an analysis or that the design has enough time-points to allow for sufficient variability to be observed.

As extensions, we only covered continuous time-varying covariates but binary time-varying covariates may also be of interest. If the binary covariate has a contemporaneous effect, then the model would not look much different than the example presented in Example 2. The covariate could be latent centered such that the person-mean would be the person-specific proportion and the within-person and between-person components could be simultaneously included to disaggregate momentary and habitual effects of the covariate. If the binary covariate is lagged or if the model has two simultaneous binary outcomes, the situation becomes a little more complex because both binary variables would be assumed to have a continuous underlying process and both variables would be disaggregated into within and between components. This can be accommodated in M*plus* and an important point to remember when interpreting the output is that M*plus* parameterizes the model such that covariates predict the latent intercept even though the fixed effect is defined as a threshold. Therefore, covariate effects should be interpreted as effects on the intercept, not the threshold. Of course, as noted in Example 4, the scaling can be somewhat arbitrary so the covariate effects could also be conceptualized as effects on the thresholds if the sign of the effect were reversed.

In closing, there are many exciting opportunities to make new discoveries or to refine our understanding of existing theories of the mechanism of behavior using intensive longitudinal data. Recent software developments have made models for intensive longitudinal data more accessible to a broader audience of researchers, including new procedures that have recently been added to M*plus*. Though models for continuous outcomes have been covered in some detail recently, resources for applying these intensive longitudinal models to categorical outcomes have been far scarcer. We hope that this paper supplements existing didactic resources and gives empirical researchers a readable introduction to approaching these analyses when the focal outcome is binary or ordinal and helps to clarify concepts that are unique to models with categorical outcomes like probit links, underlying normal processes, cumulative distribution functions, and differences between intercepts and thresholds.

Disclosures

This work was partially supported by the National Institutes of Health (NIH) Science of Behavior Change Common Fund Program through awards administered by the National Institute for Drug Abuse (NIDA) (UH2/UH3DA041713). Jennifer Somers was supported as a postdoctoral fellow on NIMH T3215750.

References

Agresti A (2012). Categorical data analysis. Wiley.

Agresti A, & Hitchcock DB (2005). Bayesian inference for categorical data analysis. Statistical Methods and Applications, 14(3), 297–330.

Albert JH, & Chib S (1993). Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association, 88(422), 669–679.

Asparouhov T (2020, February 1). Problems computing standardized estimates [Discussion post]. Mplus Discussion Forum. http://www.statmodel.com/discussion/messages/24588/27731.html?1580727445. Accessed 31 Mar 2023.

Asparouhov T, & Muthén B (2010). Plausible values for latent variables using Mplus. Muthén & Muthén. https://www.statmodel.com/download/Plausible.pdf. Accessed 31 Mar 2023.

Asparouhov T, & Muthén B (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. Structural Equation Modeling, 26(1), 119–142.

Asparouhov T, & Muthén B (2020). Comparison of models for the analysis of intensive longitudinal data. Structural Equation Modeling: A Multidisciplinary Journal, 27(2), 275–297.

Asparouhov T, & Muthén B (2021). Expanding the Bayesian structural equation, multilevel and mixture models to logit, negative-binomial, and nominal variables. Structural Equation Modeling, 28(4), 622–637.

Asparouhov T, & Muthén B (2022). Residual structural equation models. Structural Equation Modeling, 30(1), 1–31. 10.1080/10705511.2022.2074422

Asparouhov T, & Muthén B (2022b). Practical aspects of dynamic structural equation models. Muthén & Muthén. http://www.statmodel.com/download/PDSEM.pdf. Accessed 31 Mar 2023.

Asparouhov T, Hamaker EL, & Muthén B (2017). Dynamic latent class analysis. Structural Equation Modeling, 24(2), 257–269.

Asparouhov T, Hamaker EL, & Muthén B (2018). Dynamic structural equation models. Structural Equation Modeling, 25(3), 359–388.

Berli C, Inauen J, Stadler G, Scholz U, & Shrout PE (2021). Understanding between-person interventions with time-intensive longitudinal outcome data: Longitudinal mediation analyses. Annals of Behavioral Medicine, 55(5), 476–488. [PubMed: 32890399]

Bliss CI (1935). The calculation of the dosage-mortality curve. Annals of Applied Biology, 22(1), 134–167.

Bolger N, & Laurenceau JP (2013). Intensive longitudinal methods: An introduction to diary and experience sampling research. Guilford Press.

Bolger N, Davis A, & Rafaeli E (2003). Diary methods: Capturing life as it is lived. Annual Review of Psychology, 54(1), 579–616.

Brooks SP, & Gelman A (1998). General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7(4), 434–455.

Bürkner PC, & Vuorre M (2019). Ordinal regression models in psychology: A tutorial. Advances in Methods and Practices in Psychological Science, 2(1), 77–101.

Castro-Alvarez S, Tendeiro JN, Meijer RR, & Bringmann LF (2022). Using structural equation modeling to study traits and states in intensive longitudinal data. Psychological Methods, 27(1), 17–43. [PubMed: 34014719]

Chib S, & Greenberg E (1998). Analysis of multivariate probit models. Biometrika, 85(2), 347–361.

Collins LM (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. Annual Review of Psychology, 57, 505–528.

Conner TS, & Barrett LF (2012). Trends in ambulatory self-report: The role of momentary experience in psychosomatic medicine. Psychosomatic Medicine, 74, 327–337. [PubMed: 22582330]

Curran PJ, & Bauer DJ (2007). Building path diagrams for multilevel models. Psychological Methods, 12(3), 283–297. [PubMed: 17784795]

Curran PJ, & Bauer DJ (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. Annual Review of Psychology, 62, 583–619.

Curran PJ, Obeidat K, & Losardo D (2010). Twelve frequently asked questions about growth curve modeling. Journal of Cognition and Development, 11, 121–136. [PubMed: 21743795]

De Boeck P, & Wilson M (2004). Explanatory item response models: A generalized linear and nonlinear approach. Springer.

De Haan-Rietdijk S, Voelkle MC, Keijsers L, & Hamaker EL (2017). Discrete- vs. Continuous-time modeling of unequally spaced experience sampling method data. Frontiers in Psychology, 8, 1849. [PubMed: 29104554]

DeMartini KS, Gueorguieva R, Taylor JR, Krishnan-Sarin S, Pearlson G, Krystal JH, & O'Malley SS (2022). Dynamic structural equation modeling of the relationship between alcohol habit and drinking variability. Drug and Alcohol Dependence, 233,109202. [PubMed: 35151022]

Driver CC, Oud JHL, & Voelkle MC (2017). Continuous time structural equation modeling with R package ctsem. Journal of Statistical Software, 77, 1–35.

Eisenberg IW, Bissett PG, Canning JR, Dallery J, Enkavi AZ, Whitfield-Gabrieli S, ... Poldrack RA (2018). Applying novel technologies and methods to inform the ontology of self-regulation. Behaviour Research and Therapy, 101, 46–57. [PubMed: 29066077]

Enders CK (2010). Applied missing data analysis. Guilford press.

Fahrenberg J, Myrtek M, Pawlik K, & Perrez M (2007). Ambulatory assessment--Monitoring behavior in daily life settings: A behavioral-scientific challenge for psychology. European Journal of Psychological Assessment, 23(4), 206–213.

Gates KM, & Molenaar PCM (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. NeuroImage, 65, 310–319.

Gelman A, & Rubin DB (1992). Inference from iterative simulation using multiple sequences. Statistical Science, 7(4), 457–472.

Gistelinck F, Loeys T, & Flamant N (2021). Multilevel autoregressive models when the number of time points is small. Structural Equation Modeling, 28(1), 15–27.

Hamaker EL, & Grasman RP (2015). To center or not to center? Investigating inertia with a multilevel autoregressive model. Frontiers in Psychology, 5, 1492. [PubMed: 25688215]

Hamaker EL, & Wichers M (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. Current Directions in Psychological Science, 26(1), 10–15.

Hamaker EL, Dolan CV, & Molenaar PCM (2003). ARMA-based SEM when the number of time points T exceeds the number of cases N: Raw data maximum likelihood. Structural Equation Modeling, 10, 352–379.

Hamaker EL, 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. [PubMed: 29624092]

Hamaker EL, Asparouhov T, & Muthén BO (2023). Dynamic structural equation modeling as a combination of time series modeling, multilevel modeling, and structural equation modeling. In Hoyle RH (Ed.), The Handbook of Structural Equation Modeling (2nd ed.). Guilford Press.

Haqiqatkhah MM, Ryan O, & Hamaker EL (2022). Skewness and staging: Does the floor effect induce bias in multilevel AR (1) models?. PsyArXiv, https://psyarxiv.com/myuvr/, November 26, 2022.

Hoffman L (2019). On the interpretation of parameters in multivariate multilevel models across different combinations of model specification and estimation. Advances in Methods and Practices in Psychological Science, 2(3), 288–311. [PubMed: 32719825]

Hoffman L, & Walters RW (2022). Catching Up on Multilevel Modeling. Annual Review of Psychology, 73, 659–689.

Kiekens G, Hasking P, Nock MK, Boyes M, & Kirtley O,... & Claes L (2020). Fluctuations in affective states and self-efficacy to resist non-suicidal self-injury as real-time predictors of non-suicidal self-injurious thoughts and behaviors. Frontiers in Psychiatry, 11, 214. [PubMed: 32265760]

Kim CJ, & Nelson CR (1999). State-space models with regime switching: Classical and Gibbs-sampling approaches with applications. MIT Press.

Kretzschmar A, & Gignac GE (2019). At what sample size do latent variable correlations stabilize? Journal of Research in Personality, 80,17–22.

Levy R, & McNeish D (2022). Perspectives on Bayesian inference and their implications for data analysis. Psychological Methods. 10.1037/met0000443

Li Y, Wood J, Ji L, Chow SM, & Oravecz Z (2022). Fitting multilevel vector autoregressive models in Stan, JAGS, and Mplus. Structural Equation Modeling, 29(3), 452–475. [PubMed: 35601030]

Liddell TM, & Kruschke JK (2018). Analyzing ordinal data with metric models: What could possibly go wrong? Journal of Experimental Social Psychology, 79, 328–348.

Liu S (2017). Person-specific versus multilevel autoregressive models: Accuracy in parameter estimates at the population and individual levels. British Journal of Mathematical and Statistical Psychology, 70(3), 480–498. [PubMed: 28225554]

Long JS (1997). Regression models for categorical and limited dependent variables. Sage.

Lüdtke O, Marsh HW, Robitzsch A, Trautwein U, Asparouhov T, & Muthén B (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. Psychological Methods, 13, 203–229. [PubMed: 18778152]

McCullagh P (1980). Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological), 42(2), 109–127.

McNeish D, & Hamaker EL (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. Psychological Methods, 25, 610–635. [PubMed: 31855015]

McNeish D, Mackinnon DP, Marsch LA, & Poldrack RA (2021). Measurement in intensive longitudinal data. Structural Equation Modeling, 28(5), 807–822. [PubMed: 34737528]

Mehl MR, & Conner TS (2012). Handbook of research methods for studying daily life. Guilford Press.

Mislevy RJ, & Sheehan KM (1989). Information matrices in latent-variable models. Journal of Educational Statistics, 14(4), 335–350.

Moskowitz DS, & Young SN (2006). Ecological momentary assessment: What it is and why it is a method of the future in clinical psychopharmacology. Journal of Psychiatry and Neuroscience, 31(1),13. [PubMed: 16496031]

Muthén B (2010). Bayesian analysis in Mplus: A brief introduction. Los Angeles, CA: Author. Retrieved from https://www.statmodel.com/download/IntroBayesVersion3.pdf. Accessed 31 Mar 2023.

Nelson BW, & Allen NB (2018). Extending the passive-sensing toolbox: Using smart-home technology in psychological science. Perspectives on Psychological Science, 13(6), 718–733. [PubMed: 30217132]

Nickell S (1981). Biases in dynamic models with fixed effects. Econometrica, 1417–1426.

Nielsen L, Riddle M, King JW, Aklin WM, Chen W, Clark D, ... Weber, W. (2018). The NIH Science of Behavior Change Program: Transforming the science through a focus on mechanisms of change. Behaviour Research and Therapy, 101, 3–11. [PubMed: 29110885]

Ou L, Hunter M, & Chow S-M (2018). dynr: Dynamic modeling in R. (R-package version 0.1.12–5). Retrieved from: https://cran.r-project.org/web/packages/dynr/. Accessed 31 Mar 2023.

Ram N, & Gerstorf D (2009). Time-structured and net intraindividual variability: Tools for examining the development of dynamic characteristics and processes. Psychology and Aging, 24,778. [PubMed: 20025395]

Rhemtulla M, Brosseau-Liard PÉ, & Savalei V (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychological Methods, 17(3), 354–373. [PubMed: 22799625]

Robitzsch A (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. In Frontiers in Education, 5, 589965.

Rubin DB (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91(434), 473–489.

Sadikaj G, Wright AG, Dunkley DM, Zuroff DC, & Moskowitz DS (2021). Multilevel structural equation modeling for intensive longitudinal data: A practical guide for personality researchers. In Rauthman JF (Ed.), Handbook of personality dynamics and processes (pp. 855–885). Elsevier.

Savord A, McNeish D, Iida M, Quiroz S, & Ha T (2023). Fitting the longitudinal actor-partner interdependence model as a dynamic structural equation model. Structural Equation Modeling, 30(2), 296–314.

Scherer D, Metcalf SA, Whicker CL, Bartels SM, Grabinski M, Kim SJ, Sweeney MA, Lemley SM, Lavoie H, Xie H, Bissett PG, Dallery J, Kiernan M, Lowe MR, Onken L, Prochaska J, Stoeckel L, Poldrack RA, MacKinnon DP, & Marsch LA (2022). Momentary influences on self-regulation in two populations with health risk behaviors: Adults who smoke and adults who are overweight and have binge-eating disorder. Frontiers in Digital Health, Section Connected Health, 4, 798895. 10.3389/fdgth.2022.798895

Schuurman NK, Ferrer E, de Boer-Sonnenschein M, & Hamaker EL (2016). How to compare cross-lagged associations in a multilevel autoregressive model. Psychological Methods, 21(2), 206–221. [PubMed: 27045851]

Scollon CN, Kim-Prieto C, & Diener E (2003). Experience sampling: Promise and pitfalls, strengths and weaknesses. Journal of Happiness Studies, 4, 5–34.

Smyth JM, & Stone AA (2003). Ecological momentary assessment research in behavioral medicine. Journal of Happiness Studies, 4(1), 35–52.

Stroe-Kunold E, Gruber A, Stadnytska T, Werner J, & Brosig B (2012). Cointegration methodology for psychological researchers: An introduction to the analysis of dynamic process systems. British Journal of Mathematical and Statistical Psychology, 65, 511–539. [PubMed: 22070760]

ten Brink M, Lee HY, Manber R, Yeager DS, & Gross JJ (2021). Stress, sleep, and coping self-efficacy in adolescents. Journal of Youth and Adolescence, 50(3), 485–505. [PubMed: 33141378]

Trull TJ, & Ebner-Priemer U (2014). The role of ambulatory assessment in psychological science. Current Directions in Psychological Science, 23, 466–470. [PubMed: 25530686]

Vogelsmeier LV, Vermunt JK, & De Roover K (2022). How to explore within-person and between-person measurement model differences in intensive longitudinal data with the R package lmfa. Behavior Research Methods. 10.3758/s13428-022-01898-1

Walls TA, & Schafer JL (Eds.). (2006). Models for intensive longitudinal data. Oxford University Press.

Wang LP, Hamaker E, & Bergeman CS (2012). Investigating inter-individual differences in short-term intra-individual variability. Psychological Methods, 17, 567–581. [PubMed: 22924600]

Williams DR, Martin SR, Liu S, & Rast P (2020). Bayesian multivariate mixed-effects location scale modeling of longitudinal relations among affective traits, states, and physical activity. European Journal of Psychological Assessment, 36(6), 981–997. [PubMed: 34764628]

Yaremych HE, Preacher KJ, & Hedeker D (2022). Centering categorical predictors in multilevel models: Best practices and interpretation. Psychological Methods. 10.1037/met0000434

Zhou L, Wang M, & Zhang Z (2021). Intensive longitudinal data analyses with dynamic structural equation modeling. Organizational Research Methods, 24(2), 219–250.

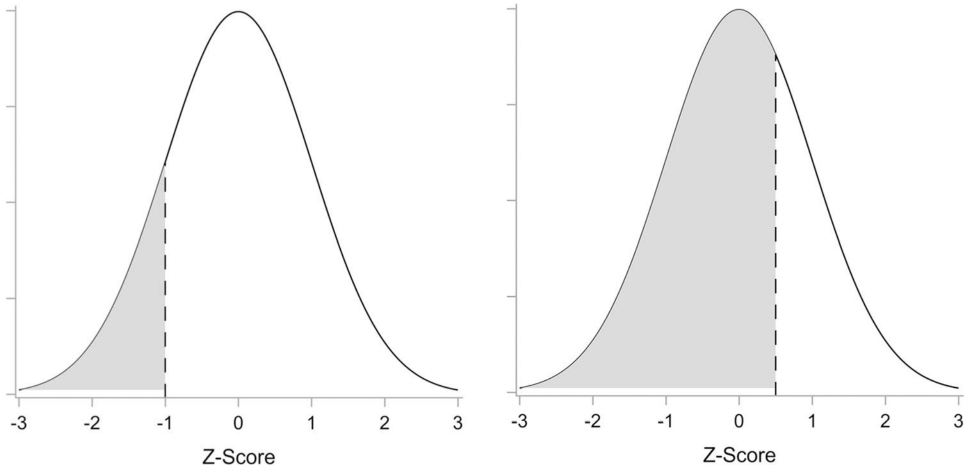

Fig. 1.

Plots showing how probit scale uses area under a standard normal distribution to determine predicted probabilities. The *left panel* shows the predicted probability of $y = 1$ as the *grey shaded area* to the left of a Z-score of −1. The *right panel* shows the predicted probability that $y = 1$ as the grey shaded area to the left of a Z-score of +0.50. The predicted probabilities for $y = 0$ based on thresholds are represented by the *white space under the curve*

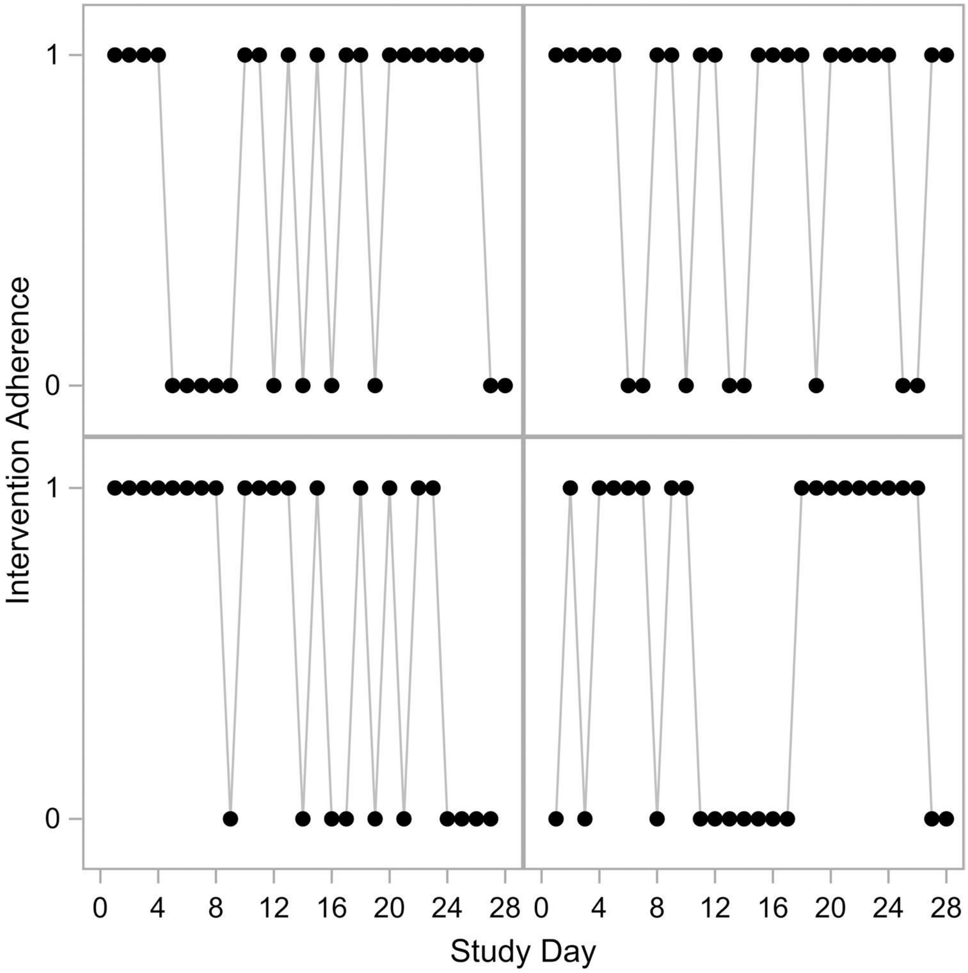

Fig. 2.

Trace plot for treatment adherence for four representative participants

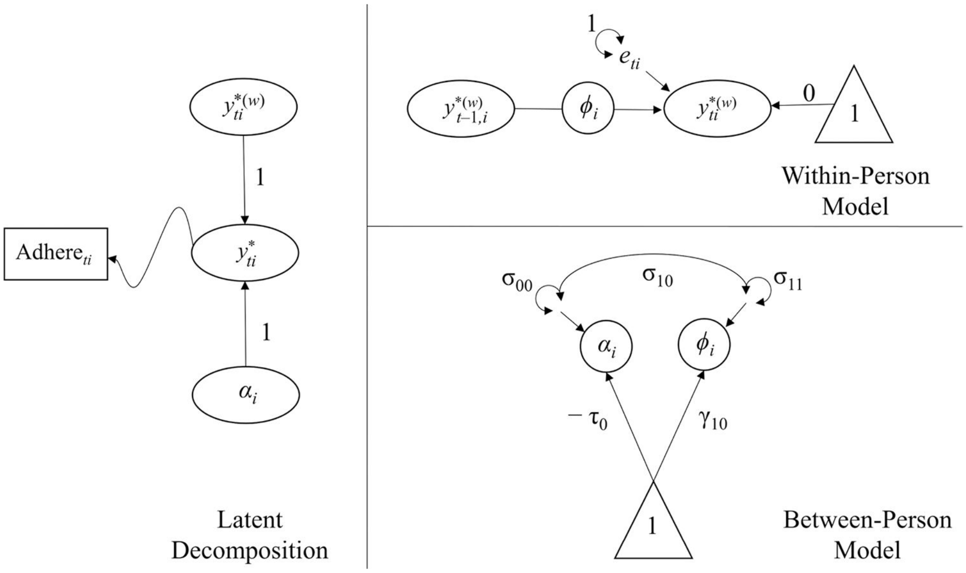

Fig. 3.

Path diagram for an unconditional probit DSEM. This model corresponds to the model shown in Eq. (5)

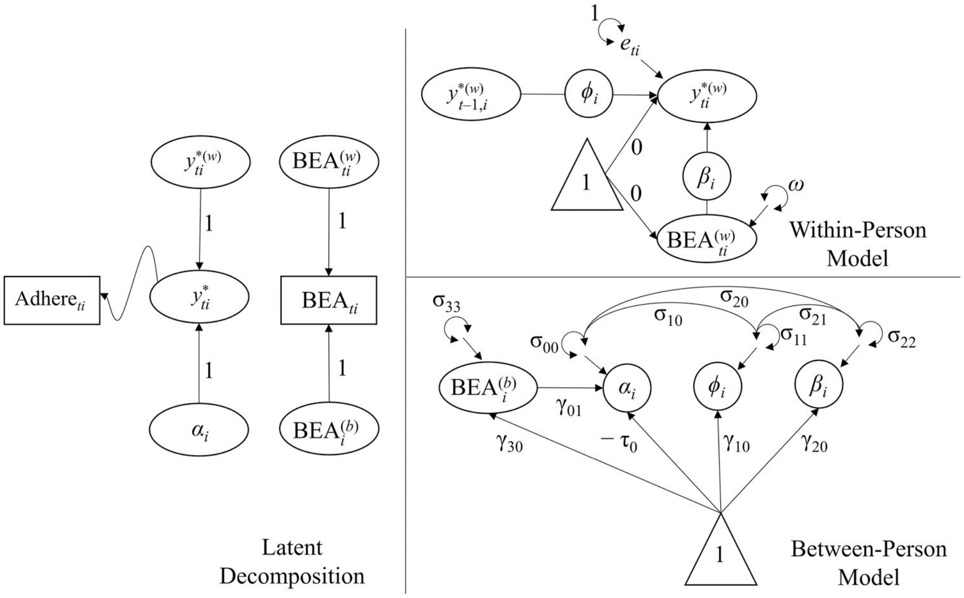

Fig. 4.

Path diagram of DSEM model with a latent-centered continuous covariate. This diagram corresponds to the model shown in Eq. (6)

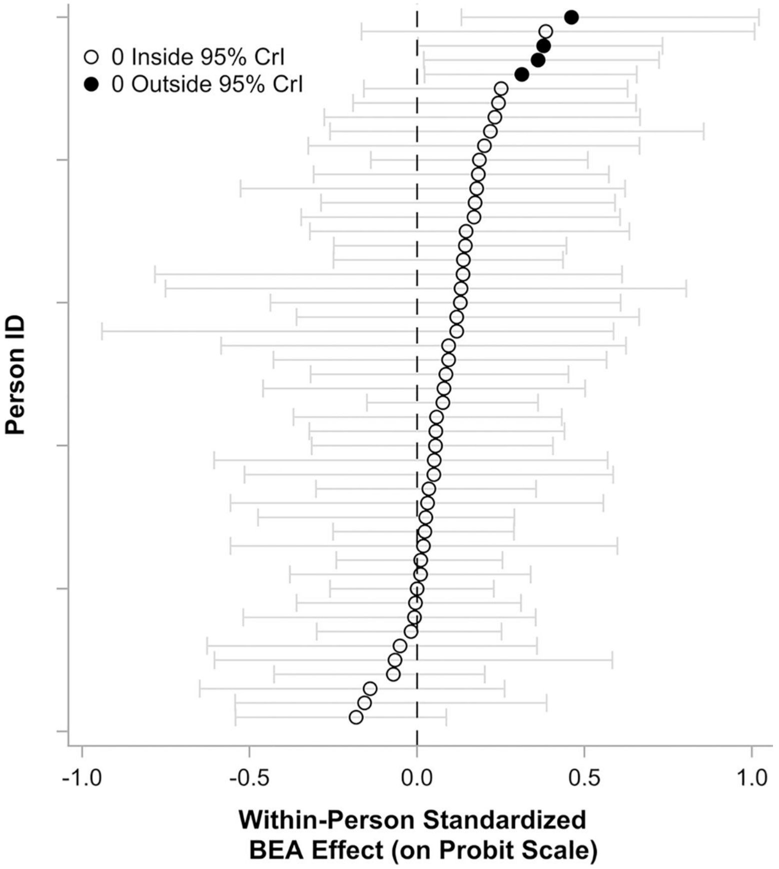

Fig. 5.

Caterpillar plot for within-person effect of binge eating avoidance for all 50 people in the study. *Solid markers* indicate that 0 was outside the person-specific 95% credible interval; *open markers* indicate that 0 was inside the person-specific 95% credible interval

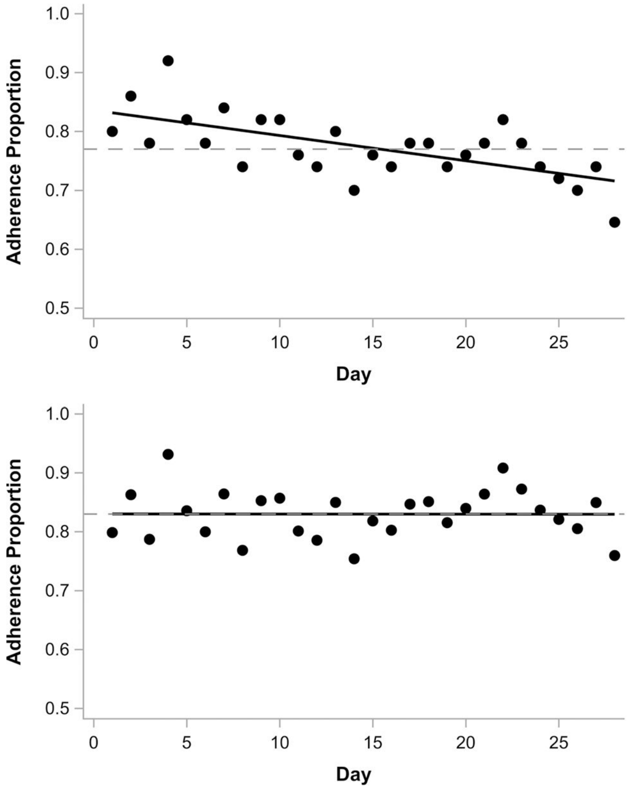

Fig. 6.

Average intervention adherence proportion across the 28 days of the study using the raw proportion (top panel) and the linearly detrended proportion (bottom panel). The horizontal dashed line is average across all days and the solid black line is the linear trend in the proportions

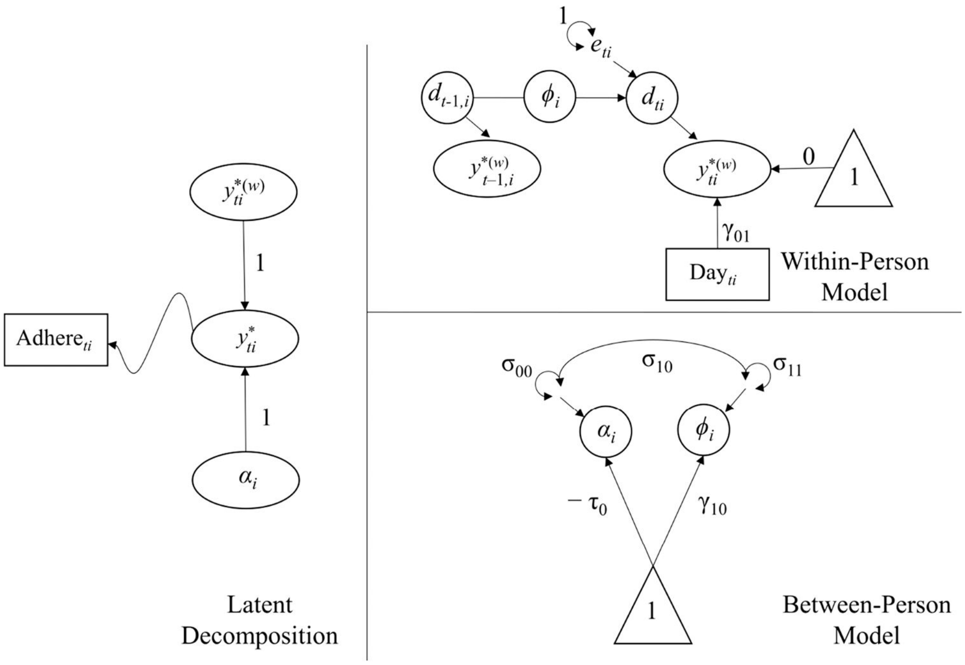

Fig. 7.

Path diagram for RDSEM model that includes a linear trend for Adherence across time. This diagram corresponds to the model in Eq. (7)

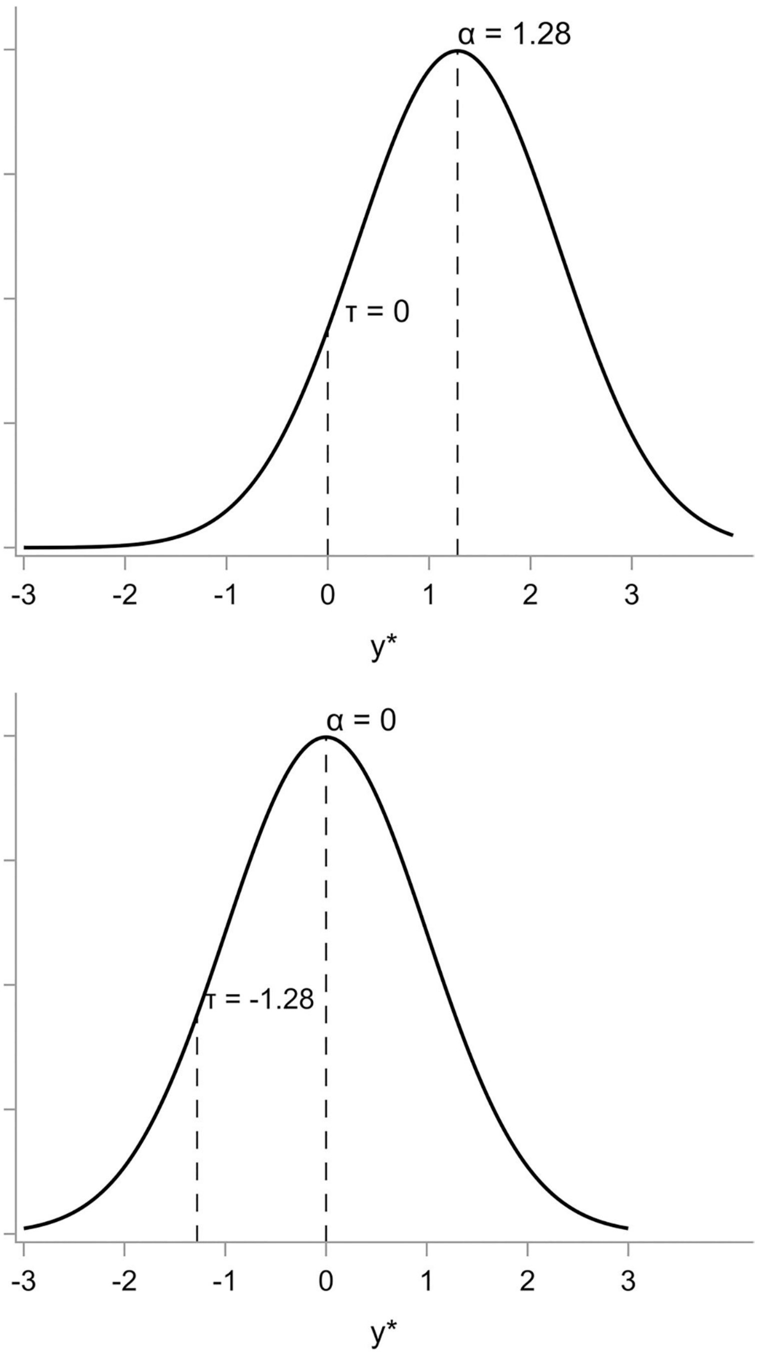

Fig. 8.

Comparison of the intercept (alpha) and the threshold (tau) for a binary outcome under two different identification strategies. The top panel constrains the threshold to 0 and estimates the intercept, the bottom panel constrains the intercept to 0 and estimates the threshold. The predicted probabilities are identical in either case

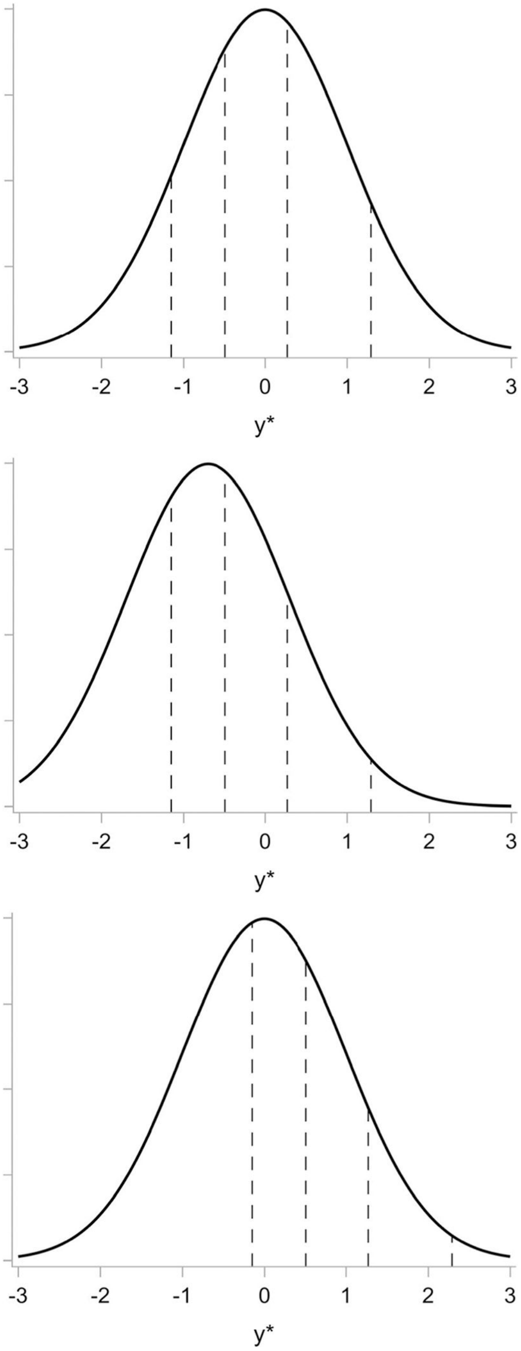

Fig. 9.

Graphical representation of multiple thresholds in a model with ordinal outcomes. In the top panel, the intercept is 0 and the thresholds are $\tau_1 = -1.16,\ \tau_2 = -0.50,\ \tau_3 = 0.26$, and $\tau_4 = 1.28$. The model allows between-person variance, so the middle panel shows the same thresholds for a person whose intercept is one standard deviation below average ($\text{mean} = -0.70$). The bottom panel shows the same information as the middle panel but parameterizes the change by keeping the intercept at 0 but shifting the threshold to the right rather than shifting the intercept to the left. The predicted probabilities in the middle and bottom panels are the same, they just present two different ways of looking at the model

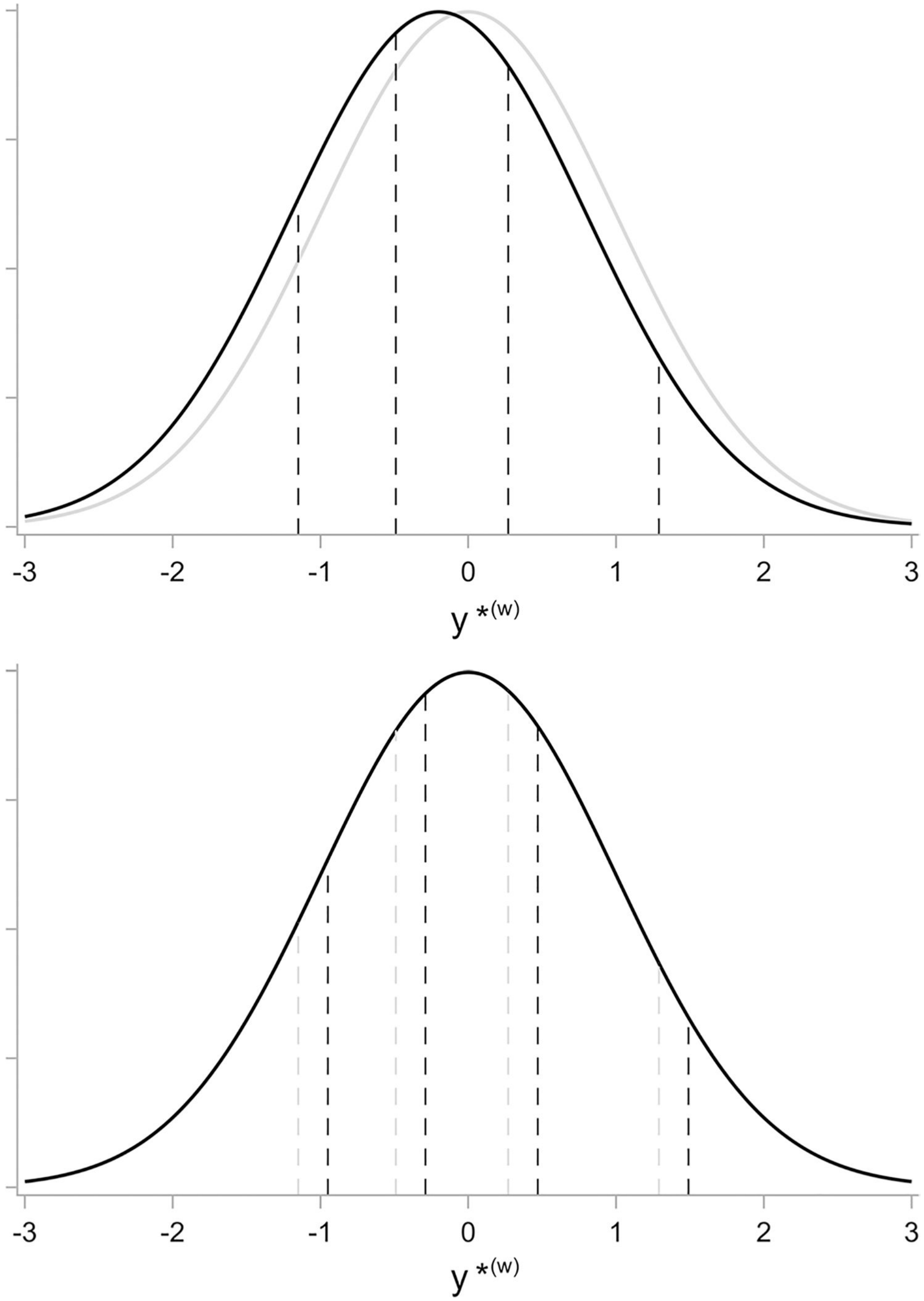

Fig. 10.

Graphical representation of within-person effects of a time-varying covariate. In the top panel, the thresholds are constant. The grey distribution represents the underlying normal within-person process when the covariate equals 0 and the black distribution represents the underlying normal within-person process when the covariate equals 1. In the bottom panel, the underlying normal within-person process has a constant mean of 0. The grey dashed lines represent the threshold when the covariate equals 0 and the black dashed lines represent the thresholds when the covariate equals 1. The predicted probabilities are equal in either case, these are just two different ways of interpreting the same effect

Table 1

Parameter estimates from multilevel autoregressive model for binary intervention adherence

Effect	**Notation**	**Est.**	**95% CrI**
Fixed effect parameters			
Threshold	τ_0	−1.28	[−1.67, −0.89]
Autocorrelation	γ_0	0.32	[0.10, 0.54]
Covariance parameters			
Intercept variance	σ_{00}	0.91	[0.36, 1.78]
Autocorrelation variance	σ_{11}	0.21	[0.11, 0.34]
Intercept, autocorrelation covariance	σ_{10}	−0.06	[−0.41, 0.23]

Note: The intercept is equal to the threshold with the sign reversed. The intercept is used for calculating predicted probabilities that the outcome equals 1; the threshold is used for calculating predicted probabilities that the outcome equals 0

Table 2

Parameter estimates from probit DSEM with binary intervention adherence as the outcome and binge eating avoidance as a latent-centered covariate

Effect	Notation	Est.	95% CrI
Fixed effect parameters			
Threshold	τ_0	0.25	[−1.37, 1.78]
Autocorrelation	γ_{10}	0.35	[0.15, 0.56]
BEA$^{(w)}$ effect	γ_{20}	0.06	[−0.07, 0.19]
BEA$^{(b)}$ effect	γ_{01}	0.23	[0.01, 0.44]
BEA latent mean	γ_{30}	7.22	[6.70, 7.72]
Variance parameters			
Intercept variance	σ_{00}	0.94	[0.37, 1.72]
Autocorrelation variance	σ_{11}	0.20	[0.10, 0.34]
BEA$^{(w)}$ effect variance	σ_{22}	0.08	[0.04, 0.14]
BEA$^{(b)}$ variance	σ_{33}	3.06	[1.91, 4.57]
BEA$^{(w)}$ variance	ω	3.13	[2.86, 3.41]
Covariance parameters			
Intercept, autocorrelation covariance	σ_{10}	0.00	[−0.32, 0.30]
Intercept, BEA$^{(w)}$ effect covariance	σ_{20}	0.06	[−0.12, 0.23]
Autocorrelation, BEA$^{(w)}$ effect covariance	σ_{21}	0.02	[−0.04, 0.08]

Note: The intercept is equal to the threshold with the sign reversed. The intercept is used for calculating predicted probabilities that the outcome equals 1; the threshold is used for calculating predicted probabilities that the outcome equals 0

Table 3

Predicted probabilities for intervention adherence for different latent person means of binge eating avoidance

Latent person mean $BEA_i^{(b)}$	**Predicted probability** $\Phi\left[\left(-225 + 0.225 \times BEA_i^{(b)}\right)/1.6\beta\right]$
0	0.44
1	0.49
2	0.55
3	0.60
4	0.65
5	0.70
6	0.75
7	0.79
8	0.83
9	0.86
10	0.89

Note: In the equation under the "Predicted Probability" column, −0.25 is fixed threshold estimate conditional on $BEA_i^{(b)}$ = 0, 0.225 is the between-person covariate effect of $BEA_i^{(b)}$, and 1.63 is the standard deviation of y*which is included to convert the quantity to scale of a standard normal distribution

Table 4

Parameter estimates from probit DSEM with five-category Likert response to “tempting food made it difficult for me to not binge eat today” as the outcome and binge eating avoidance as a latent-centered covariate

Effect	Notation	Est.	95% CrI
Fixed effect parameters			
Threshold 1	τ_1	−1.16	[−1.36, −0.94]
Threshold 2	τ_2	−0.50	[−0.69, −0.29]
Threshold 3	τ_3	0.26	[0.09, 0.47]
Threshold 4	τ_4	1.28	[1.11, 1.50]
Autocorrelation	γ_{10}	0.12	[0.00, 0.25]
BEA$^{(w)}$ effect	γ_{20}	−0.10	[−0.20, −0.01]
BEA$^{(b)}$ effect	γ_{01}	0.04	[−0.10, 0.17]
BEA latent mean	γ_{30}	−0.01	[−0.51, 0.49]
Variance parameters			
Intercept variance	σ_{00}	0.49	[0.26, 0.78]
Autocorrelation variance	σ_{11}	0.10	[0.05, 0.16]
BEA$^{(w)}$ effect variance	σ_{22}	0.06	[0.04, 0.10]
BEA$^{(b)}$ variance	σ_{33}	3.01	[1.90, 4.61]
BEA$^{(w)}$ variance	ω	3.11	[2.84, 3.39]
Covariance parameters			
Intercept, autocorrelation covariance	σ_{10}	−0.04	[−0.15, 0.06]
Intercept, BEA$^{(w)}$ effect covariance	σ_{20}	0.02	[−0.06, 0.10]
Autocorrelation, BEA$^{(w)}$ effect covariance	σ_{21}	0.00	[−0.04, 0.03]