Cross-Lagged Panel Modeling with Binary and Ordinal Outcomes

Bengt Muthén, Tihomir Asparouhov
& Katie Witkiewitz

Version 2 [*]

April 9, 2024

[*] We thank Noah Hastings who provided support with tables and figures.

Abstract

To date, cross-lagged panel modeling has been studied only for continuous outcomes. This paper presents methods that are suitable also when there are binary and ordinal outcomes. Modeling, testing, identification, and estimation are discussed. A two-part ordinal model is proposed for ordinal variables with strong floor effects often seen in applications. An example considers the interaction between stress and alcohol use in an alcohol treatment study. Extensions to multiple-group analysis and modeling in the presence of trends are discussed.

Keywords: panel data; random intercept; RI-CLPM; stress and drinking; alcohol treatment study

1 Introduction

The cross-lagged panel model (CLPM) and its random intercept counterpart RI-CLPM are popular models for investigating longitudinal relationships between two or more variables where the variables at each time point are regressed on themselves and each other at previous time points. For an overview and a discussion of the merits of CLPM and RI-CLPM, see, e.g., Hamaker (2023). To date, however, the literature covers only continuous variables. This paper presents methods that are suitable also when there are binary and ordinal variables.

The binary and ordinal case needs special considerations in terms of modeling and estimation. Maximum likelihood estimation is generally not feasible but Bayesian and weighted least squares methods can be used. The paper demonstrates analysis methods that work well in practice for both binary and ordinal variables as well as combinations of binary, ordinal, and continuous variables.

Section 2 discusses modeling and testing. Section 3 treats identification, and estimation matters for the case of a binary variable. Section 4 presents simulations for the binary univariate case as well as the binary bivariate RI-CLPM case. Section 5 discusses applications of analyses with a binary variable using alcohol data from a large randomized treatment study. Section 6 considers ordinal variables and presents a two-part ordinal model suitable for variables that have strong floor effects. Section 7 continues the alcohol treatment example using an ordinal alcohol risk variable. Section 8 presents extensions to the analysis of multiple groups as well as models that allow trends. Section 9 concludes. Mplus scripts for key analyses are given in the Supplementary Material.

2 Binary, univariate outcome: Modeling and testing concepts

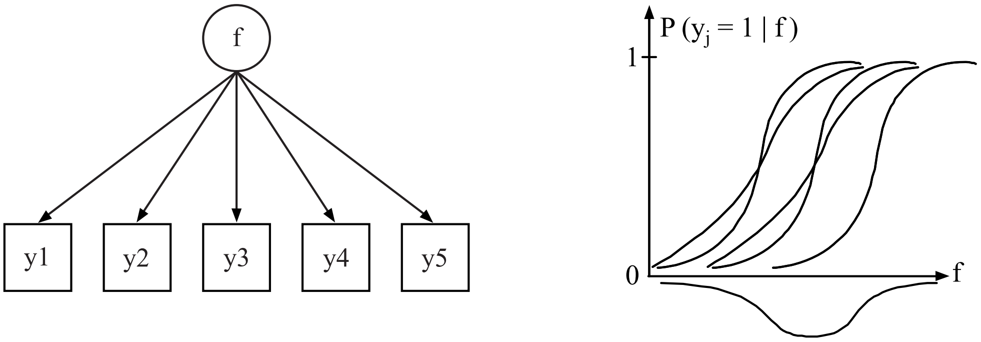
This section gives an introduction to modeling and testing concepts for binary and ordinal variables. Readers who are unfamiliar with categorical variable methods are recommended to study books such as Agresti (2012, 2018), Long (1997), and de Ayala (2022).[1] The case of a binary outcome is considered as a starting point. For this case it is instructive to first consider basics of Item Response Theory (IRT) and then relate that to factor analysis with binary variables. Following the description of these techniques, which are suitable for cross-sectional data, modeling of longitudinal data central to this paper is discussed. The section ends with a discussion of model testing.

2.1 IRT and factor analysis

The left part of Figure 1 shows a model with a latent variable f influencing five variables y1-y5. With continuous observed variables y, the arrows represent linear regressions with slopes referred to as factor loadings. For binary observed variables, the arrows represent non-linear regressions as shown in the right part of the figure. The regressions express the probability of $y_j = 1$ as opposed to 0 as a function of the value of the

[1] Because Mplus is used in this paper, Mplus Short Course Topic 2 posted at `https://www.statmodel.com/topic2.shtml` is also informative.

Figure 1: Model with one latent variable represented as a model diagram (a circle represents a latent variable and a square represents an observed variables) and as probability curves



latent variable. For example, the variables may represent incorrect/correct responses to five different math test items where the latent variable is referred to as ability or achievement. As the ability increases, the probability of answering the item correctly increases. For a given ability value, the five regression curves show that the right-most curve represents the most difficult item in that the probability of answering it correctly is the lowest. The curves also differ in how well they discriminate between ability values where steeper curves represent larger probability differences between lower and higher ability values. The parameterization of difficulty and discrimination is used in Item Response Theory (IRT). For an introduction to IRT, see, e.g., de Ayala (2022). IRT uses logistic or normal (probit) distribution functions to describe the curves of Figure 1,

$$P(y=1|f) = 1/(1+e^{-a(f-b)}), \tag{1}$$

$$P(y=1|f) = \Phi[a(f-b)], \tag{2}$$

where a and b are the discrimination and difficulty parameters for a certain variable and Φ is the standard normal distribution function used for probit. The variables are assumed to be independent conditioned on f. Here, f has mean 0 and variance 1. The logistic and probit curves are very similar.

Figure 2 shows an equivalent formulation of the IRT model which is akin to factor analysis using continuous outcomes. Factor analysis of binary and ordinal variables (see Christofferson, 1975; Muthén, 1978) considers a continuous latent response variable y* underlying each binary observed variable y, where y = 0 or 1 is determined by exceeding a threshold τ or not,

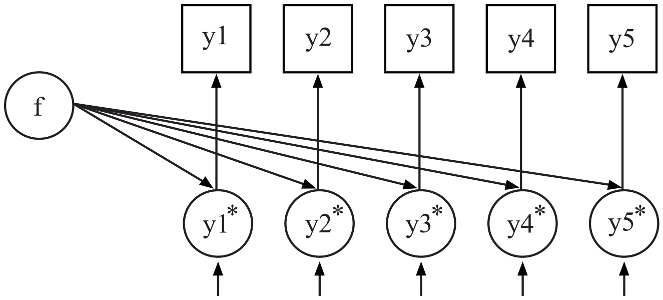
$$y = \begin{cases} 1, & \text{if } y^* > \tau \\ 0, & \text{if } y^* \leq \tau, \end{cases} \tag{3}$$

where the latent response variable y^* follows a linear regression on the factor f,

$$y^* = \lambda\, f + \epsilon, \tag{4}$$

where ϵ is a residual with mean zero and variance θ. The θ parameter is not available in IRT. Unless multiple groups or multiple timepoints are considered, θ is not identified

Figure 2: 1-factor model using a latent response variable representation



but is fixed at 1, or alternatively, the y* variance is fixed at 1.[2] The variables are assumed to be independent conditioned on f. The factor analysis parameterization uses thresholds and factor loadings. In Figure 2, the arrows between the y^* circles and the y boxes refer to the threshold relationship of (3) while the arrows between f and y^* refer to the factor loadings of (4). The short arrows at the bottom of the y^* circles refer to the ϵ residuals. Continuing the math test example, the y* value for a certain variable as induced by the factor may be just above the threshold or far above it, thereby capturing the idea that the skill needed to solve the item correctly could be measured in a finer gradation. The threshold formulation is also useful for the generalization to ordinal outcomes where a variable has more than one threshold.

IRT and binary factor analysis are equivalent models which can be seen as follows. Assuming a normally distributed residual ϵ, the model of (3) and (4) implies a probit regression of y on f as in (2),

$$\begin{aligned} P(y=1|f) &= P(y^* > \tau|f), && (5)\\ &= 1 - P(y^* \leq \tau|f), && (6)\\ &= 1 - \Phi[(\tau - \lambda f)/\sqrt{\theta}], && (7)\\ &= \Phi[(-\tau + \lambda f)/\sqrt{\theta}], && (8) \end{aligned}$$

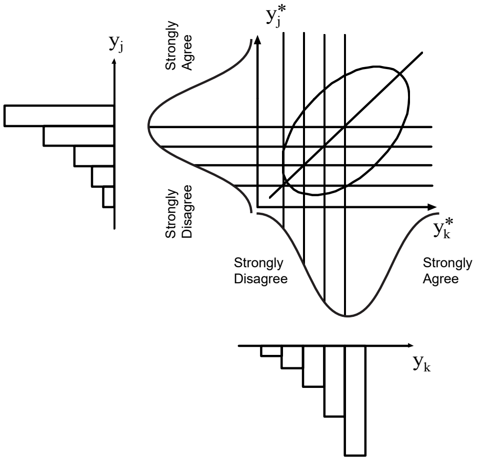
where λf is the mean of y^* conditioned on f and the last equality of $1 - \Phi[x] = \Phi[-x]$ is due to the symmetry of the distribution function. The translation between the two parameterizations of a, b and τ, λ is

$$\begin{aligned} a &= \lambda/\sqrt{\theta}, && (9)\\ b &= \tau/\lambda, && (10) \end{aligned}$$

where θ is fixed at 1. When both ϵ and f in (4) are normally distributed, y^* is normal because a sum of two normal distributions is normal. For the factor model, multivariate normality for the set of y* variables can be specified together with uncorrelated (or independent) residuals ϵ. Specifying a logistic instead of a normal distribution for the ϵ residual in (4) results in a logistic regression of y on f. Modeling with the logistic distribution, however, does not generalize to the models considered in this

[2] This correspond to the Theta versus Delta parameterizations in Mplus.

Figure 3: Ordinal model using a latent response variable representation



paper because there is not a multivariate version of the logistic distribution with the flexibility that the multivariate normal distribution offers. For technical details with further IRT and factor models, see Muthén and Asparouhov (2016) and Asparouhov and Muthén (2020a). The equivalence between the IRT formulation of (1) and (2) and the factor analysis formulation of (3) and (4) is well known (see, e.g., Muthén, 1983 with formal proofs in Takane & de Leeuw, 1987).

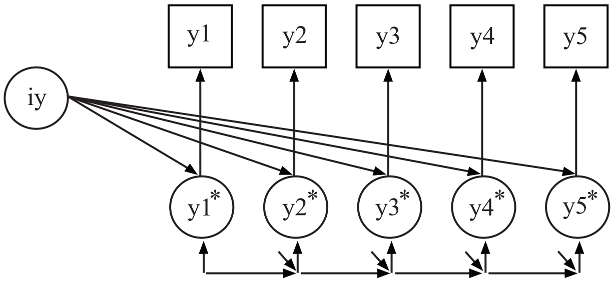
Assuming a normal distribution for the factor f together with multivariate normal residuals ϵ, not only univariate normal residuals ϵ, is equivalent to assuming a multivariate normal distribution for the y* variables. The correlations between the normal y* variables are referred to as tetrachoric when the observed variables are binary. The y* concept also covers the ordinal case of Figure 3. This shows an example of two 5-category ordinal variables represented by two y* variables, each of which has four thresholds. The thresholds are not equally spaced over the y* distribution, causing strong ceiling effects. There is a linear relationship between the y* variables as follows from normality which motivates the use of correlations among the y* variables (in the ordinal case, the correlations are referred to as polychoric and polyserial if one variable is continuous). The regular (Pearson) correlation between the y variables is quite different than the correlation between the y* variables.[3] For example, a factor model that holds for a set of y* correlations may not hold for the corresponding set of y correlations. The representation drawing on multivariate normal y* variables behind binary and ordinal variables will be referred to as a multivariate probit model.

2.2 Longitudinal modeling

So far, the observed variables in Figure 1 and Figure 2 have been viewed as different variables, that is, corresponding to a cross-sectional analysis. In that case, the focus is on the relationships between f and the y's. This paper, however, considers longitudinal

[3] See Mplus Short Course Topic 2 posted at https://www.statmodel.com/topic2.shtml, slides 133-139.

Figure 4: Random intercept model with auto-regressive residuals (iy represents the random intercept for y)



analysis where the same variable is measured at several time points. In the longitudinal case, the factor of Figure 2 corresponds to a random intercept where the factor loadings are all fixed at 1. The strength of covariance among the y* variables at different time points is constant and captured by the random intercept variance. With a trend, one or more random slope latent variables can be added.

In longitudinal modeling for continuous outcomes it is well-established that correlation among the residuals needs to be allowed for in order for the model to fit well (see, e.g., Chi & Reinsel, 1989). The random intercept alone cannot capture all the correlation among the y*'s and thereby not among the y's. In the earlier discussion of IRT and factor analysis, the model was specified for univariate outcomes with the added specification of independence of the outcomes conditioned on the latent variable. In the longitudinal modeling with auto-correlated residuals, this is not sufficient because the y outcomes are no longer independent conditioned on the latent variable f. Instead, the multivariate probit model specification is needed. This makes the modeling, testing and estimation more complex. Related longitudinal modeling was presented in Hedeker and Gibbons (1994), Muthén (1996), and Hedeker and Mermelstein (2000) but not with auto-regressions among the residuals which are key in the current setting. Asparouhov and Muthén (2020b) included auto-regressive residuals but focused on the different setting of intensive longitudinal data. The current paper is novel in that the modeling includes auto-regression of residuals applied to the framework of RI-CLPM.

Figure 4 shows a model with linear auto-regressions of lag 1 for the y* residuals. In this way, the y* variables are related to each other, that is, there is a correlation between them beyond what is predicted by their common dependence on the random intercept *iy*. This is different from Figure 2 which does not consider relationships among the y*'s. The statistical representation of Figure 4 is given in Section 3. Figure 4 is a categorical counterpart to the univariate (single process) part of a continuous-outcome RI-CLPM (Hamaker et al., 2015). The relationships among the residuals capture the dynamic within part of the model. Cross-lagged effects between two or more processes refer to the relationships between the residuals of these processes. The contribution of this paper is to show the applicability of the multivariate probit approach to the RI-CLPM for binary and ordinal outcomes.

2.3 Testing

The assumption of underlying y* normality of the multivariate probit model for Figure 4 should be tested. The model misspecification sensitivity to the normality assumption for the y^* variables has been investigated in a series of SEM articles. For example, Flora and Curran (2004), Rhemtulla et al. (2012), and Li (2016) found little sensitivity. Foldnes and Gronneberg (2022) criticized these investigations and found more sensitivity using new types of underlying non-normal distributions. It is unclear, however, how realistic these non-normal distributions are. Furthermore, they found that alternative approaches did not perform better, such as ignoring the categorical nature of the variables and treating them as continuous. It is clear, however, that the assumption of multivariate normality for the y* variables does not necessarily fit every data set and it is important to test the assumption. Following is a discussion of the testing topic. For the ordinal case, normality may be frequently rejected. Because of this, extensions of the model for the ordinal case to achieve good model fit are presented in Section 6.1.

Model fit assessment for categorical outcomes can be done in several different ways. One approach is analogous to what is used in structural equation modeling, where fit to covariances or correlations is considered. For the multivariate probit model, this amounts to fit of correlations among the latent response variables y* underlying each observed categorical y variable. This was studied in Muthén (1983, 1984) and Muthén et al. (1997) using chi-square testing based on a weighed least-squares estimator (WLSMV) for a multivariate probit model. The Muthén et al. (1997) WLSMV chi-square works well when the number of variables is not large and the sample size is not small which makes it particularly suitable for cross-lagged panel modeling. Considering the weighted least-squares chi-square test of fit, Foldnes and Gronneberg (2022, p. 562) found that it was inflated when normality does not hold, thereby providing a conservative testing approach. Foldnes and Gronneberg (2022) also studied a bootstrap testing procedure based on generating data from the estimated correlation matrix for the y* variables. Checking of model fit based on chi-square for a multivariate probit model is also possible with Bayesian estimation as discussed in Asparouhov and Muthén (2021a, b) using posterior predictive checking that produces a posterior predictive p-value (PPP). Using any fit statistic, the general posterior predictive checking approach is to compute the fit statistic for the observed data, generate a fit statistic distribution based on generated data from the estimated model, and find the proportion of cases where the latter is larger than the former. Based on the same overall chi-square as used with WLSMV, the Bayes approach, however, has low power for binary outcomes and is less powerful than the WLSMV chi-square test (Asparouhov & Muthén, 2021a). The Bayes approach is more powerful for ordinal variables.

A second approach considers the fit to the data in the form of response patterns, that is, a frequency table for all variables. A model may fit the y* correlations well but not the frequency table. Even a just-identified y* model that includes all possible correlations may not fit the frequency table because the assumption of underlying normality does not hold. With categorical variables, the model can be tested against data using the standard Pearson and likelihood-ratio chi-square frequency table tests.

Summing over the cells of the table, these two tests are expressed as:

$$Pearson: \ \sum_j (o_j - e_j)^2/e_j \tag{11}$$

$$Likelihood\ ratio: \ 2\sum_j o_j \ log(o_j/e_j) \tag{12}$$

Such testing was discussed in Muthén (1993). An example was given with rejection of underlying normality due misfit for only two out of 49 cells in the bivariate frequency table. It was conjectured that this was due to anomalous response behavior and that the normal y* model presented a smoothed version of the data.

In the multivariate case, there are, however, typically too many frequency table cells with many cells having estimated frequencies close to zero, invalidating the tests. For example, with 8 binary variables there are $2^8 = 256$ possible response patterns, where many patterns are often not observed (zero cells in the frequency table) leading to the two tests disagreeing strongly and becoming useless. A practical approach is to consider fit to the most frequent response patterns, e.g., the twenty most frequent. There are, however, alternative frequency table checks where the tables are collapsed into univariate and bivariate tables which ensures higher frequencies. In particular, bivariate frequency checking is a useful way to find model misspecification. Asparouhov and Muthén (2022) presents significance testing of standardized residuals for both response patterns and bivariate frequency tables.[4]

This paper will do model testing using a combination of chi-square testing using WLSMV estimation, Bayes PPP testing, and using standardized residuals for response patterns, uni- and bi-variate frequency tables. Not all of these test can be made in a given situation but depend on which estimators can be used for the different model types.

3 Binary univariate case: Identification and estimation

Consider the binary random intercept probit model shown in Figure 5 for T = 5. Using a standard statistical notation for an intercept, α is the same as the random intercept called iy in Figure 4. The model can be expressed as follows for individual i and time point t. For the y^*_t continuous latent response variable at time t with threshold parameter τ_t, $y^*_t > \tau_t$ implies $y_t = 1$ while otherwise $y_t = 0$. The model specifies the linear relations

$$y^*_{it} = \alpha_i + \epsilon_{it}, \tag{13}$$

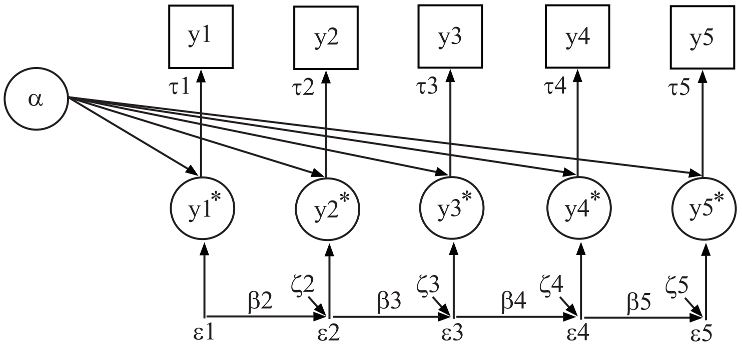
$$\epsilon_{it} = \beta_t \ \epsilon_{it-1} + \zeta_{it}; \ t = 2, \ldots, T, \tag{14}$$

$$\epsilon_{i1} = \zeta_{i1}, \tag{15}$$

where $\alpha_i \sim N(0, \psi)$ represents the random intercept, ϵ_{it} represents the residual for y^*_t, β_t represents the auto-regression, and $\zeta_t \sim N(0, \theta_t)$ are the residuals in the auto-regressions. This results in multivariate normal y^* variables, that is, it is a multivariate probit model.

[4] In Mplus, TECH10 for WLSMV and Bayes give standardized residuals for response patterns, uni- and bi-variate frequency tables, and Bayes PPP for Pearson fit to uni- and bi-variate tables.

Figure 5: Random intercept model with auto-regressive residuals



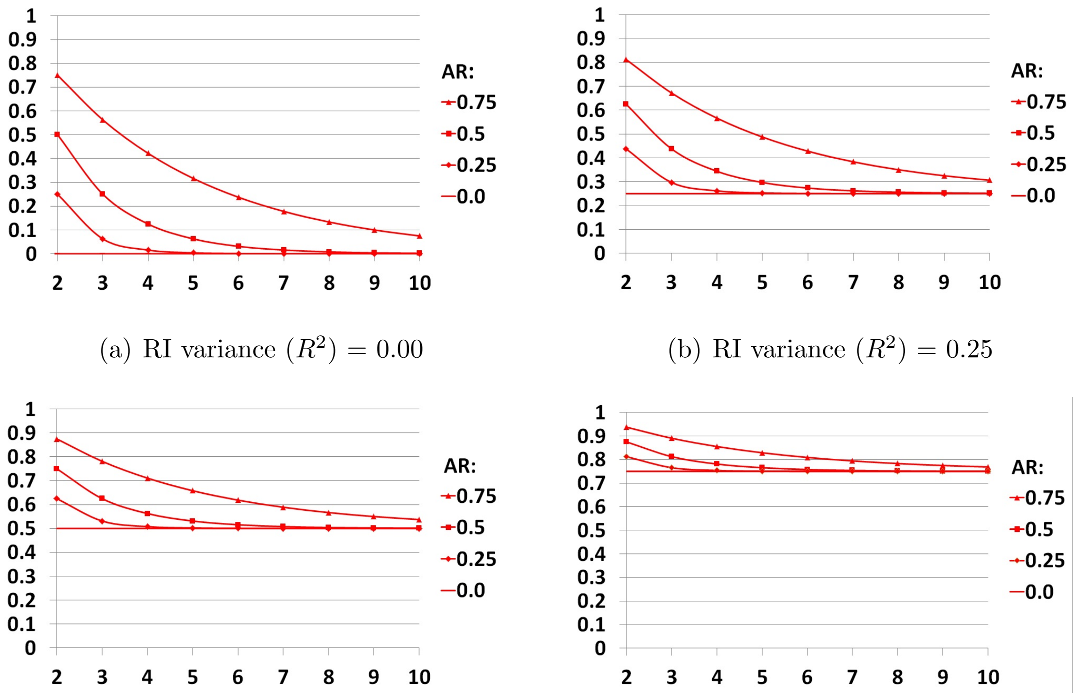
3.1 Identification

A major distinction between this model and the corresponding model for continuous outcomes is that the variances for the ζ residuals are not all identified. This can be portrayed as the loss of information due to not observing the y* variables directly but only in a discretized fashion. For a binary variable, the mean π and variance $\pi(1-\pi)$ are mathematically related so that the sample mean and variance do not provide information about two separate parameters as in the continuous case. A maximum of $T-1$ $V(\zeta_{it})$ variances can be identified as will be discussed below. Typically, however, it is difficult to estimate $T-1$ variances without incurring large standard errors for model parameters. Estimation faces an empirical identification issue where the number of identifiable variances depends on the data in terms of correlations across time, number of time points, and sample size. Fixing all residual variances at 1 is often reasonable as the default. This is also the standard in maximum-likelihood estimation of growth models for ordinal outcomes when there is no residual correlation (see, e.g., Hedeker & Gibbons, 1994). A first step to relax this default model would be to free the first residual variance which may be the largest, but in practice even this variance may obtain a large standard error with no improvement in model fit. This has been observed in the data sets encountered so far.

Identification issues for the binary random intercept model with auto-regressive residual are as follows. Consider first the impact of random intercept variance and auto-regression on the y* correlations over time. Figure 6 shows four panels which differ by the magnitude of the random intercept variance. Within each panel four curves are shown differing by the magnitude of auto-regression. The curves show the correlation between y* at $t=1$ and y* at $t=2,3,\ldots T$ for $T=10$. The curves are computed by the formula $\text{Corr}(y_1^*, y_t^*) = \psi + \beta^{t-1}(1-\psi)$ where ψ is the random intercept variance, β is the constant auto-regression among the residuals, and y* variances are all 1. Due to the unit y* variances, ψ is the same as the R-square of y* explained by the random intercept. The formula shows that as t increases, β^{t-1} decreases due to $|\beta| < 1$. This means that as t increases, the second term on the right-hand side of the formula decreases and the correlation decreases down towards the asymptote of ψ, the random intercept variance. The larger ψ is, the higher the asymptote. Also, the higher the auto-regression, the slower the decline in correlations over time. For instance, panel (c) of Figure 6 shows the case of random intercept variance 0.5 where with auto-regression of 0.25, the correlation between the first two time points is a little above 0.6 and declines to the asymptote of 0.5 at approximately $t = 4$.

While the threshold parameters are trivially identified in terms of the proportion y=1 for the outcomes, the key model parameters ψ, β_t, and residual variances θ_t need to be identified in terms of correlations among the y* variables. Figure 6 suggests how this identification is accomplished. This can be viewed as choosing the estimates of the ψ, β_t, θ_t parameters to fit the curve of the sample correlations at different time distances.

Figure 6: RI and AR1 impact on y* correlations across time (T=10)



(a) RI variance (R^2) = 0.00

(b) RI variance (R^2) = 0.25

(c) RI variance (R^2) = 0.50

(d) RI variance (R^2) = 0.75

Muthén and Asparouhov (2002), see also Muthén (1996), showed that growth modeling with categorical outcomes and no auto-correlated residuals can identify $T - 1$ residual variances in addition to the random effect variances (a simulation with correlated residuals was presented in Muthén, 1996). Random intercept modeling is a special case of such modeling. A growth model for binary outcomes needs $T \geq 4$ while the random intercept model needs $T \geq 3$. Figure 6 (c) shows that in the case of random intercept variance $\psi = 0.5$ and auto-regression 0.25, the auto-regression itself gives zero correlation contribution at t $\approx$ 4, that is, a time distance of at least 3. Three time points spaced at least 3 time points apart would therefore have no auto correlation, leading to the identification of ψ and $T - 1$ θ_t's based on Muthén and Asparouhov (2002).

The actual identification expressions can be presented as follows. Consider again the Figure 6 (c) case of T = 10 with an auto-regression of 0.25 so that the auto-regression itself gives zero contribution at t $\approx$ 4. The correlations over longer time distances are solely due to the random intercept variance ψ. This implies that ψ is known (identified), in this case as 0.5. Consider the correlation between y_a^* and y_b^* for timepoints a and b,

$$Corr\ (y_a^*, y_b^*) = \psi/(\sqrt{\psi + \theta_a}\sqrt{\psi + \theta_b}). \tag{16}$$

With ψ known (identified) and fixing the residual variance θ_1 to 1, (16) with $a = 1$ shows that correlations between time 1 and later timepoints identify the remaining $T - 1$ θ_t's when considering the correlations for longer time distances where the auto-regressive contributions are zero.

Consider again the example of the auto-regressive contribution to the correlation being zero for $t \geq 4$, that is, at a time distance of at least 3. With ψ known (identified) and $\theta_1 = 1$, the correlation between $t = 1$ and $t = 4$ identifies θ_4, the correlation between $t = 1$ and $t = 5$ identifies θ_5, etc. up to the correlation between $t = 1$ and $t = 10$ identifying θ_{10}. In all cases, the time distance is at least 3. It remains to identify θ_2 and θ_3 for which the time distance is less than 3. But knowing for instance θ_{10} as just stated, (16) shows that with $a = 2$, $b = 10$, the correlation between $t = 2$ and $t = 10$ identifies θ_2 and that with $a = 3$, $b = 10$, the correlation between $t = 3$ and $t = 10$ identifies θ_3. The T-1 auto-regression parameters of β are then identified from among the remaining correlations.

3.2 Estimation

Estimation of the univariate random intercept probit model may be carried out by maximum likelihood (ML), weighted least-squares (WLSMV), and Bayes. All three estimators are available in Mplus (Muthén & Muthén, 1998-2017). The ML estimator, however, needs to use numerical integration over the T+1 latent variables and is therefore not feasible for a typical number of time points due to the number of quadrature points increasing exponentially with T resulting in too time consuming computations and loss of numerical precision. WLSMV (Muthén et al., 1997) is a fast estimator not needing numerical integration and handling the multivariate probit model also for larger T. It can give information about the empirical identification status in that it presents the condition number of the estimated information matrix.[5] The WLSMV estimation uses a convenient residual specification where the residuals can be referred to directly as latent variables; see Asparouhov and Muthén (2023).[6] This means that the residuals can be regressed on each other as is needed for the auto regressions and for the cross-lagged regressions in the bivariate case.[7] WLSMV is, however, disadvantaged because it does not handle MAR missingness like the full-information ML and Bayes estimators. The Bayes estimator combines a practical approach with full-information estimation. The Bayes approach to be used in the application sections handles the multivariate probit model using an efficient algorithm (Asparouhov & Muthén, 2020b) together with the residual specification (see Asparouhov & Muthén, 2023). In some cases, non-symmetric confidence/credibility intervals are needed. With Bayes, this is

[5] In Mplus, this is the ratio of smallest to largest eigenvalue of the estimated information matrix.

[6] This is the hat notation in Mplus. The Theta parameterization is used.

[7] In contrast, the weighted least squares estimation in Muthén (1983, 1984, 1996) could estimate only correlations among residuals.

Table 1: Number of parameters, sample statistics, and degrees of freedom for the binary outcome univariate random effect probit model with fixed residual variances

T	# parameters	# sample statistics	DF
3	3+1+2=6	6	0
4	4+1+3=8	10	2
5	5+1+4=10	15	5
10	10+1+9=20	55	35

obtained automatically while with WLSMV, bootstrapping is needed. Because of the arbitrary scale of the y^* latent response variables, it is useful to present estimates in a standardized metric where the y^* variances are 1.

Table 1 shows the number of parameters τ_t, ψ, β_t , the number of sample statistics, and degrees of freedom for the univariate random effect probit model with fixed residual variances θ_t. Here, degrees of freedom refers to the number of restrictions imposed on the sample statistics of univariate proportions and correlations. This is the degrees of freedom of the chi-square test of fit for the weighted least-squares estimator (WLSMV). T = 3 is the minimum number of time points required for identifying the model. It should be noted, however, that T = 3 is a bare minimum for this type of analysis with categorical outcomes because this generally provides little information to distinguish between correlation due to the random intercept versus due to autocorrelation. More time points are strongly recommended.

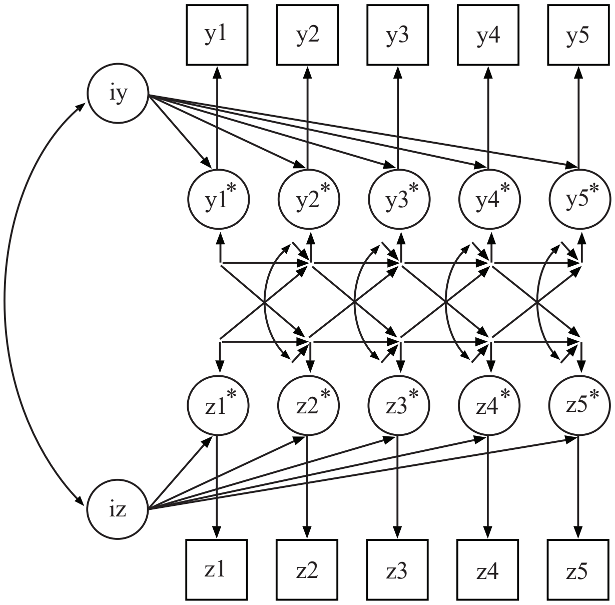
3.3 Binary bivariate case: RI-CLPM

The generalization of the binary univariate outcome model to the bivariate case is shown in Figure 7. This is the binary counterpart to RI-CLPM (Hamaker et al., 2015). The bivariate model offers no further identification complications beyond the univariate case. Each univariate part follows the identification rules just discussed. The cross-lagged parameters are identified in terms of the correlations between the y* and z* variables. Estimation can be performed by Bayes or WLSMV.

4 Binary outcome simulations: Univariate and bivariate cases

Simulations for the case of a univariate binary outcome are shown in Table 2 and Table 3 for the case of T = 5. Population values are based on analyses with suicidal ideation and substance abuse data (Ialongo, 2022). Data are generated with time-varying thresholds representing binary outcomes of low prevalence with $P(y_1 = 1)$ ranging from 0.12 to 0.20. The low-prevalence binary case represents lower-end categorical information, so that better performance can be expected for binary variables with more even split and

Figure 7: Bivariate RI-CLPM for categorical outcomes



for ordinal variables. Time-varying thresholds and time-varying auto-regressions are specified in the analyses. Residual variances are fixed at 1. For simplicity, there is no missing data. 500 replications are carried out using both WLSMV and Bayes estimation. With Bayes, 2,000 draws (iterations) are recorded, having skipped every 10th iteration for better standard error estimation due to lower autocorrelation between consecutive parameter values in the iterations. WLSMV uses first- and second-order information from univariate proportions and correlations whereas Bayes uses full information. With binary outcomes and T = 5, WLSMV uses information from 15 sample statistics (5 univariate proportions and 10 tetrachoric correlations). The $2^5 = 32$ response patterns represent the potential raw data that is used by the full-information estimation by Bayes. This means that Bayes uses about twice as many sample statistics as WLSMV which is expected to reduce the variability of the estimates. Many of the response pattern frequencies are, however, low which implies that WLSMV uses key parts of the available information so that the reduction in variability by Bayes may not be large.

WLSMV results in Table 2 show that at N = 500, somewhat biased parameter estimates and standard errors are obtained. In particular, the first two auto regressions are underestimated and the random intercept variance is overestimated with its standard error underestimated. These biases may result in inclusion of a random intercept when it is not needed. The model has 5 degrees of freedom. The WLS chi-square testing results are good with a mean of 4.86 and a 5% reject proportions of 0.042. For N = 1000, all results are acceptable. Bayes results in Table 3 show a similar picture. For N = 500, the Bayes standard errors perform better than WLSMV and are on the whole smaller as expected by a full-information estimator as compared to the limited-information estimation by WLSMV.

Table 2: Monte Carlo results for univariate binary outcome for T = 5 using WLSMV

		ESTIMATE		S. E.		95%	% Sig
	Population	Average	Std. Dev.	Average	M. S. E.	Cover	Coeff
			N = 500				
Z5ˆ ON							
Z4ˆ	0.122	0.0890	0.1762	0.1588	0.0320	0.924	0.118
Z6ˆ ON							
Z5ˆ	0.089	0.0754	0.1795	0.1715	0.0323	0.938	0.078
Z7ˆ ON							
Z6ˆ	0.166	0.1556	0.1940	0.1750	0.0377	0.936	0.162
Z8ˆ ON							
Z7ˆ	0.126	0.1245	0.1788	0.1723	0.0319	0.944	0.138
Thresholds							
Z4$1	1.282	1.3066	0.1311	0.1212	0.0177	0.932	1.000
Z5$1	1.438	1.4659	0.1341	0.1269	0.0187	0.936	1.000
Z6$1	1.663	1.7010	0.1522	0.1363	0.0245	0.926	1.000
Z7$1	1.786	1.8233	0.1515	0.1412	0.0243	0.930	1.000
Z8$1	1.879	1.9205	0.1510	0.1466	0.0245	0.946	1.000
Variances							
IZ	1.536	1.6280	0.3284	0.2934	0.1161	0.920	1.000
			N = 1000				
Z5ˆ ON							
Z4ˆ	0.122	0.1060	0.1136	0.1104	0.0131	0.942	0.166
Z6ˆ ON							
Z5ˆ	0.089	0.0834	0.1236	0.1214	0.0153	0.944	0.098
Z7ˆ ON							
Z6ˆ	0.166	0.1582	0.1256	0.1239	0.0158	0.954	0.264
Z8ˆ ON							
Z7ˆ	0.126	0.1271	0.1203	0.1219	0.0145	0.956	0.198
Thresholds							
Z4$1	1.282	1.2992	0.0846	0.0848	0.0074	0.960	1.000
Z5$1	1.438	1.4551	0.0872	0.0882	0.0079	0.956	1.000
Z6$1	1.663	1.6882	0.1012	0.0946	0.0108	0.922	1.000
Z7$1	1.786	1.8074	0.0969	0.0977	0.0098	0.948	1.000
Z8$1	1.879	1.9037	0.1006	0.1014	0.0107	0.960	1.000
Variances							
IZ	1.536	1.5953	0.2059	0.2018	0.0458	0.956	1.000

Table 3: Monte Carlo results for univariate binary outcome for T = 5 using Bayes

		ESTIMATE		S. E.		95%	% Sig
	Population	Average	Std. Dev.	Average	M. S. E.	Cover	Coeff
			N = 500				
Z5ˆ ON							
Z4ˆ	0.122	0.0967	0.1546	0.1575	0.0245	0.956	0.102
Z6ˆ ON							
Z5ˆ	0.089	0.0669	0.1571	0.1671	0.0251	0.960	0.062
Z7ˆ ON							
Z6ˆ	0.166	0.1571	0.1749	0.1717	0.0306	0.936	0.158
Z8ˆ ON							
Z7ˆ	0.126	0.1347	0.1760	0.1731	0.0310	0.938	0.156
Thresholds							
Z4$1	1.282	1.3044	0.1190	0.1195	0.0146	0.942	1.000
Z5$1	1.438	1.4800	0.1266	0.1259	0.0177	0.932	1.000
Z6$1	1.663	1.7088	0.1361	0.1345	0.0206	0.950	1.000
Z7$1	1.786	1.8283	0.1409	0.1393	0.0216	0.938	1.000
Z8$1	1.879	1.9298	0.1381	0.1457	0.0217	0.952	1.000
Variances							
IZ	1.536	1.6598	0.2836	0.2962	0.0956	0.940	1.000
			N = 1000				
Z5ˆ ON							
Z4ˆ	0.122	0.1072	0.1101	0.1091	0.0123	0.946	0.172
Z6ˆ ON							
Z5ˆ	0.089	0.0841	0.1130	0.1167	0.0128	0.954	0.098
Z7ˆ ON							
Z6ˆ	0.166	0.1585	0.1278	0.1206	0.0164	0.936	0.288
Z8ˆ ON							
Z7ˆ	0.126	0.1307	0.1271	0.1204	0.0161	0.928	0.216
Thresholds							
Z4$1	1.282	1.2940	0.0866	0.0833	0.0076	0.952	1.000
Z5$1	1.438	1.4598	0.0869	0.0869	0.0080	0.954	1.000
Z6$1	1.663	1.6874	0.0924	0.0928	0.0091	0.950	1.000
Z7$1	1.786	1.8067	0.0947	0.0960	0.0094	0.946	1.000
Z8$1	1.879	1.9037	0.0940	0.0998	0.0094	0.952	1.000
Variances							
IZ	1.536	1.5962	0.1844	0.1986	0.0375	0.944	1.000

Simulations for the bivariate binary outcome case of RI-CLPM with T=5 are shown in Table 4 for the WLSMV estimator and in Table 5 for the Bayes estimator. The population parameter values are again chosen from the Ialongo (2022) study with one variable having the same univariate parameters as those of the univariate simulation and the other with similar values. Time-varying thresholds and auto-regressions are again specified in the analyses. The cross-lagged effects are specified as time varying. For simplicity, there is no missing data. With Bayes, 5,000 draws (iterations) are recorded, having skipped every 10th iteration for better standard error estimation. 500 replications are carried out. With two binary outcomes and T = 5, WLSMV uses information from 55 sample statistics (10 univariate proportions and 45 tetrachoric correlations). The $2^{10} = 1024$ response patterns represent the potential raw data that is used by the full-information estimation by Bayes. This means that Bayes uses about 20 times as many sample statistics as WLSMV and this is expected to make important reduction in the variability of the estimates. The tables show results for only the new parameters of cross-lagged effects.

For WLSMV, N=500 is clearly insufficient as is seen in the parameter estimate bias and standard error bias. N=1000 shows an improvement and N=2000 shows acceptable results. With N=2000, the power to detect the cross-lagged effect of Z7ˆ ON Y6ˆ is estimated as 0.858 (see the last column). The model has 34 parameters and 21 degrees of freedom. The WLSMV chi-square testing of model fit is performing well with chi-square mean and 5% rejection proportions for N=1000/N=2000 of 20.3/20.5 and 0.05/0.04.

For Bayes, the results are acceptable already at N=500 and excellent at N=1000. For N=2000, the power to detect the Z7ˆ ON Y6ˆ effect is estimated as 0.944. As expected, Bayes has lower variability in the estimates than WLSMV. The advantage of the full-information Bayes estimator versus the limited-information WLSMV estimator is clear from these results.

5 Binary outcome example

Data from the COMBINE Study of Alcohol Use Disorder are used to illustrate the techniques for categorical outcomes. COMBINE is a 16-week, multisite randomized double-blind clinical trial comparing treatments of alcohol dependence (Anton et al., 2006). The sample size is 1,383. The measurement occasions to be considered here are: Baseline, week 1, week 2, week 4, week 6, week 8, week 10, week 12, week 16. There are also follow-up measurement occasions up to week 52. For this illustration, the T = 8 time points of the treatment are used, week 1 - week 16. The focus is on the relationship between perceived stress and alcohol use during the trial. There is a robust literature examining associations between alcohol and stress using preclinical models with non-human animals, human laboratory studies, and intensive longitudinal studies (see Armeli et al., 2000; Becker, 2017; Sinha, 2022), but few studies have examined bidirectional effects during treatment among individuals with alcohol use disorder. The stress variable is based on a 4-item, brief version of The Perceived Stress Scale with scores of 0 to 16 (McHugh et al., 2013). The analyses will use two different measures of alcohol use. Alcohol Risk is measured as a 5-category variable: Abstinence, low risk, medium risk, high risk, very high risk. These are WHO-defined drinking risk levels based on amount of alcohol consumed. The Alcohol Risk variable is also used to define

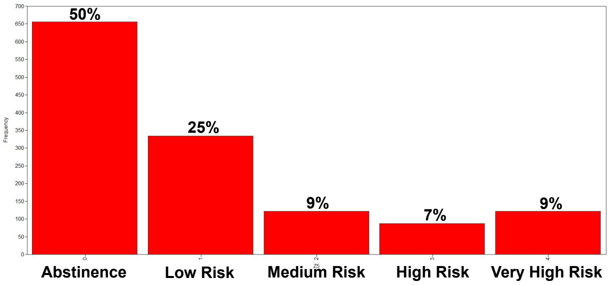
Table 4: Monte Carlo results for bivariate binary RI-CLPM, T = 5, WLSMV

		ESTIMATE		S. E.		95%	% Sig
	Population	Average	Std. Dev.	Average	M. S. E.	Cover	Coeff
N = 500							
Y7ˆ ON							
Z6ˆ	0.213	0.1987	1.1535	0.5677	1.3280	0.941	0.134
Z7ˆ ON							
Y6ˆ	0.375	0.4620	1.7628	0.7869	3.1086	0.899	0.312
N = 1000							
Y7ˆ ON							
Z6ˆ	0.213	0.2199	0.2962	0.1865	0.0876	0.918	0.258
Z7ˆ ON							
Y6ˆ	0.375	0.3882	0.2549	0.1876	0.0650	0.896	0.584
N = 2000							
Y7ˆ ON							
Z6ˆ	0.213	0.2116	0.1389	0.1239	0.0193	0.930	0.422
Z7ˆ ON							
Y6ˆ	0.375	0.3732	0.1383	0.1254	0.0191	0.924	0.858

Table 5: Monte Carlo results for bivariate binary RI-CLPM, T = 5, Bayes

		ESTIMATE		S. E.		95%	% Sig
	Population	Average	Std. Dev.	Average	M. S. E.	Cover	Coeff
N = 500							
Y7ˆ ON Z6ˆ	0.213	0.2409	0.2234	0.2629	0.0506	0.962	0.188
Z7ˆ ON Y6ˆ	0.375	0.3788	0.2398	0.2645	0.0574	0.956	0.394
N = 1000							
Y7ˆ ON Z6ˆ	0.213	0.2348	0.1645	0.1705	0.0275	0.952	0.338
Z7ˆ ON Y6ˆ	0.375	0.3803	0.1713	0.1731	0.0293	0.940	0.682
N = 2000							
Y7ˆ ON Z6ˆ	0.213	0.2244	0.1102	0.1130	0.0122	0.954	0.550
Z7ˆ ON Y6ˆ	0.375	0.3798	0.1140	0.1160	0.0130	0.950	0.944

Figure 8: Distribution of the alcohol risk variable at week 4



a binary variable of Abstinence versus not by combining the four highest categories. The analyses will treat the stress variable as continuous using linear regressions. This is not appropriate for Alcohol Risk which as shown in Figure 8 has a strong floor effect that would bias results from a linear model. The binary Abstinence variable is analyzed first, whereas analyses using ordinal models for Alcohol Risk are presented later.

5.1 CLPM and RI-CLPM analyses using the binary abstinence variable

This section presented results of the T = 8 analysis of the binary abstinence variable. 8% have missing data at all eight time points and are deleted, resulting in a sample size of 1,375. There is rather little attrition. At the last time point of week 16, 93% remain in the sample. The proportion non-abstinent varies between 0.45 and 0.51. The analyses will use both the limited information WLSMV estimator and the full-information Bayes estimator but large differences in results are not expected due to the low degree of missing data. With categorical data, the raw data can be represented by the response patterns observed in the sample. With binary outcomes and T = 8, there are $2^8 = 256$ possible patterns. In these data, 234 patterns have non-zero frequency, 125 patterns have frequency greater than 1, and 20 patterns have frequency greater than 10. The 234 observed response patterns constitute the full information in the raw data which is used by the Bayes estimator. In contrast, the WLSMV estimator uses only the 36 first- and second-order sample statistics corresponding to the proportions and correlations among the 8 variables.

The 20 most frequent response patterns are shown in Table 6. It is seen that 23% of the sample has the response pattern of all zeros, that is, individuals who are abstinent at all eight time points. The estimated frequencies for WLSMV and Bayes in Table 6 refer to the unrestricted binary probit model where no restrictions are placed on the thresholds or correlations among the continuous latent response variables. This tests if a probit model is suitable for the data in the first place before adding restrictions on the correlations. The unrestricted probit model has 36 parameters (8 thresholds and 28 correlations) whereas an unrestricted frequency table model has $2^8 - 1 = 255$ parameters. In other words, the unrestricted probit model is a very parsimonious representation of the data. Table 6 shows that this model fits reasonably well with no significant standardized residuals among the 20 most frequent patterns for WLSMV

Table 6: Response pattern frequencies for abstinence outcome

			Estimated		Stand'd Residual	
		Observed	Frequency		Z-score	
Pattern	Percentage	Frequency	WLSMV	BAYES	WLSMV	BAYES
00000000	22.69	312.00	291.16	289.78	1.28	1.37
11111111	19.85	273.00	291.67	291.80	-1.17	-1.18
11111110	2.76	38.00	26.23	25.47	1.86	1.98
00000001	2.47	34.00	31.02	33.42	0.49	0.10
01111111	1.60	27.00	28.62	28.88	-0.29	-0.33
00011111	1.60	22.00	15.89	14.00	1.26	1.66
00111111	1.38	19.00	16.52	18.91	0.54	0.02
11111011	1.09	15.00	12.11	10.59	0.72	1.10
10000000	1.09	15.00	12.23	12.11	0.68	0.72
11110000	1.02	14.00	11.98	11.35	0.52	0.68
00010000	1.02	14.00	12.11	16.52	0.48	-0.59
00000111	0.95	13.00	11.98	11.60	0.27	0.37
11111101	0.95	13.00	10.84	10.72	0.57	0.61
11101111	0.87	12.00	11.10	10.09	0.25	0.53
00000100	0.87	12.00	10.72	8.70	0.35	0.92
00001000	0.87	12.00	11.73	11.35	0.07	0.18
11000000	0.87	12.00	10.34	12.36	0.46	-0.10
11111100	0.87	12.00	9.58	10.21	0.67	0.49
11110111	0.87	12.00	11.85	11.98	0.04	0.01
11011111	0.80	11.00	8.20	8.57	0.81	0.70

and only 1 significant standardized residual for Bayes (z-score = 1.98).

Table 7 shows univariate analyses of the binary abstinence variable using the Bayes estimator.[8] The first model is the just mentioned unrestricted probit model. As described in Section 2, the Bayes estimator provides a posterior predictive p-value (PPP) where PPP$>$ 0.05 is often used as a descriptive measure of acceptable fit and PPP around 0.5 is considered excellent. The unrestricted probit model, model 1, gets a PPP of 0.520. PPP is, however, always around 0.5 for a model that is just-identified like model 1. Although Bayes uses more information than first- and second-order moments by using the further information in the raw data, the PPP model testing is based on chi-square which still concerns fit to the first- and second-order moments so the model is still just-identified. In other words, the H0 and H1 models are the same. WLSMV chi-square testing has zero degrees of freedom and can therefore also not be used to test the unrestricted model 1.

As mentioned in Section 2, testing the model against data can be done by checking the fit to the response patterns and the bivariate frequency tables. In the current

[8] Mplus scripts for key analyses are given in the Supplementary Material.

Table 7: Bayes results for univariate analysis of abstinence (N = 1375, T = 8)

	Model	# par's	PPP	# Sig. Residuals		Comment
				Resp Pattern	Bivar	
1.	Unrestricted	36	0.520	1*	0	Good fit
2.	AR1	15	0.082	2	4	Ok fit
3.	AR2	21	0.474	0	0	Good fit
4.	RI-AR1	16	0.189	2	4	Good fit
5.	RI-AR2	22	0.472	0	0	Good fit

* 3rd most frequent pattern with 38 observations and Z-score=1.98.

analyses, standardized residual fit for the 20 most frequent response patterns are used as one part in assessing overall model fit. In addition, fit to the bivariate frequency tables is considered. There are $8(8-1)/2 = 28$ bivariate frequency tables and since each table has 4 cells, there are 112 cells total available for testing of standardized residuals. Making the crude approximation of independent tests in the cells, a Type I error of 5% for the 112 test, or 6 tests, are expected to be significant when the model is correct. This number will be used as a threshold for a descriptive fit assessment. The unrestricted model of Table 7 shows that only 1 response pattern, the third most frequent pattern with frequency 38, has a significant misfit in terms of the standardized residual and the z-score is only 1.98. None of the bivariate standardized residuals show misfit. The overall assessment is that this model fits the data well which means that testing of restrictions on the correlations as in models 2 - 5 is appropriate.

Models 2 and 3 of Table 7 are models without the random intercept and using auto-regression with lags 1 and 2, respectively. From the improvement in fit, it is clear that a lag of 2 is motivated. Models 4 and 5 use a random intercept, where again there is a preference for using a lag of 2. Model 5 with lag 2 obtains a small variance of 0.089 for the random intercept with Bayesian credibility interval [0.001 0.375]. This indicates that there is not a large trait component for the tendency to report abstinence or not over the 8 weeks. Letting the first residual variance be freely estimated as discussed in Section 3 does not improve fit and gives a large standard error for the residual variance.

Turning to the analysis of primary interest, Table 8 shows results for bivariate analysis of the binary abstinence variable and the continuous stress variable. For these two variables, the Bayes PPP and the WLSMV chi-square refer to the full bivariate model, whereas the number of significant residuals refer to the binary abstinence variable only. The random intercept model for the bivariate case was shown in Figure 7, except that the continuous latent response variables are directly observed for the stress variable. The cross-lagged effects are lag 1 for all models. Time invariance is not imposed for any of the parameters. This model is referred to as RI-CLPM. The CLPM models 1, and 2 that do not have random intercepts fit poorly. The RI-CLPM model 3 with auto-regressions of lag 1 also fits poorly, whereas the RI-CLPM model 4 with auto-regressions of lag 2 fits well as assessed by both Bayes and WLSMV. Models 5 - 8 will be discussed in the next section.

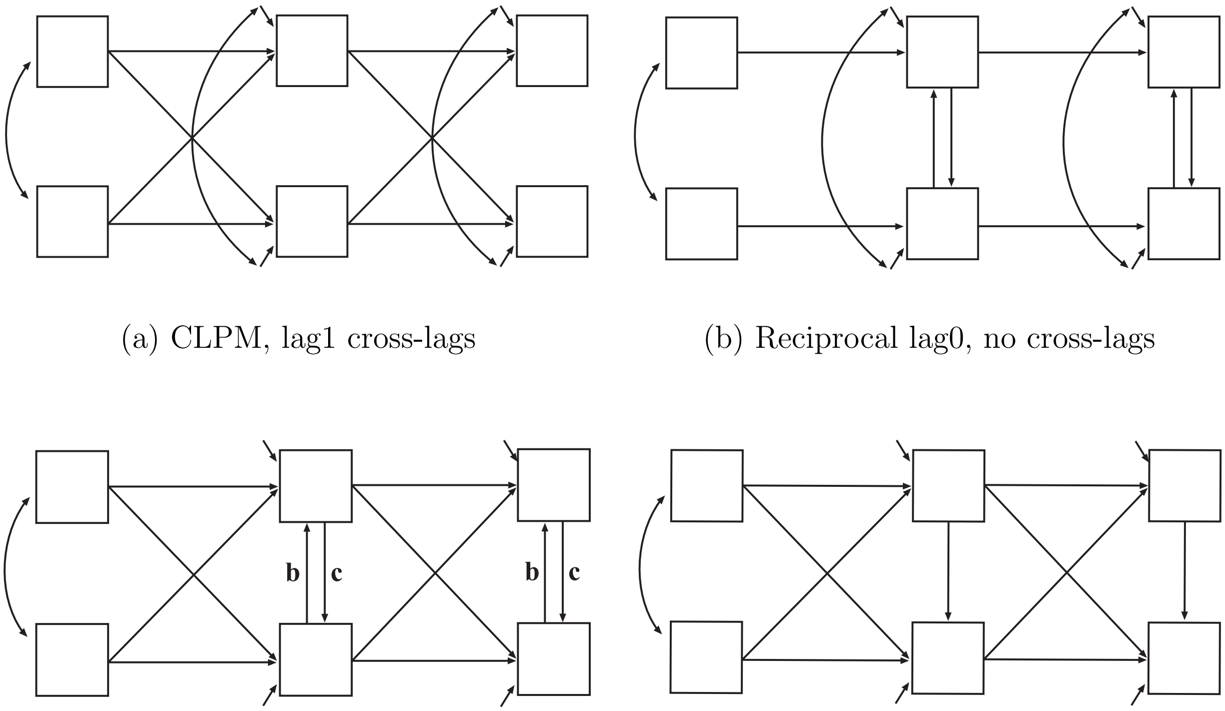
Table 8: Results for bivariate analysis of stress and abstinence (N = 1375, T = 8)

	Model	# par's	PPP/ χ^2	# Sig. Residuals		Comment
				Resp Pattern	Bivar	
1.	CLPM1, Bayes	60	0.000	1	19	Poor fit
2.	CLPM2, Bayes	72	0.000	0	0	Poor fit
3.	RI-CLPM1, Bayes	63	0.016	2	6	Poor fit
4.	RI-CLPM2, Bayes	75	0.283	0	0	Good fit
	RI-CLPM2, WLSMV	75	$\chi^2(69)$=83 (p=.1218)	0	0	Good fit
5.	RI-RCLPM, WLSMV	70	$\chi^2(74)$=84 (p=.1899)	0	0	Good fit
6.	RI-RLPM, WLSMV	63	$\chi^2(81)$=87 (p=.2999)	0	0	Good fit
7.	Single-direction lag 0, WLSMV	69	$\chi^2(75)$=84 (p=.2143)	0	0	Good fit
8.	Single-direction lag 0, WLSMV No cross-lagged effects	55	$\chi^2(89)$=87 (p=.5519)	0	0	Good fit

For model 4, the abstinence random intercept variance is now somewhat larger than in the univariate analysis, with Bayes estimate 0.191 and CI [0.081 0.413]. As mentioned earlier, it is useful to present estimates in a standardized metric given the arbitrary scale of the latent response variables. The random intercept variance estimate for abstinence translates to small R-square values for the latent response variables at the different time points, with values between 0.03 and 0.16. In contrast, the random intercept variance for the stress outcome gives high R-square values for the latent response variables, ranging from 0.53 to 0.56. For the residual auto regressions, the abstinence R-square values are in the range of 0.7 to 0.8 whereas the stress R-square values are lower, ranging from 0.10 to 0.23. The WLSMV estimates are similar. Freeing the first residual variance for abstinence as discussed in Section 3, does not lead to an estimate significantly different from the default of 1 for either estimator. The significance of the cross-lagged effects is found to be the same when having this residual variance fixed or free.

The interesting key finding of model 4 is that the cross-lagged estimates show no significant influence of stress on non-abstinence whereas all seven cross-lagged effects of non-abstinence on stress are significant with standardized values ranging from 0.17 to 0.28. These lagged effects suggest that failing to stay abstinent during the trial causes increased stress. The lag of 1 for the cross-lagged effects translates to 2 weeks, except for the second time point which is 1 week after the first and the eighth time point which is 4 weeks after the seventh.

Figure 9: Four equivalent panel models for T = 3



(a) CLPM, lag1 cross-lags

(b) Reciprocal lag0, no cross-lags

(c) Reciprocal lag0, lag1 cross-lags, no residual covariances

(d) Single-direction lag0, lag1 cross-lags, no residual covariances

5.2 Binary outcome: Contemporaneous and reciprocal modeling alternatives

Conclusions drawn from the CLPM and RI-CLPM models have been challenged in Muthén and Asparouhov (2023). They pointed out that there are several competing models that are equivalent or nearly equivalent in terms of model fit but have different interpretation. They argued for also examining models that allow contemporaneous (lag 0) effects instead of or in addition to cross-lagged effects. While this challenge was made in the context of continuous outcomes, the same principles hold also with categorical outcomes. Figure 9 displays four key models for T = 3 shown as equivalent in Muthén and Asparouhov (2023). For simplicity, no random intercepts are included. Model (a) is the regular CLPM, models (b) and (c) use reciprocal, lag 0 effects but differ in whether cross-lagged effects are included. Model (d) has a single-direction lag 0 effect and cross-lagged effects. Estimates of reciprocal effects in models (b) and (c) often find a significant lag 0 effect in only one direction, thereby giving support for model (d).

The reciprocal model (b) is referred to as RI-RLPM (random intercept reciprocal lagged panel model) and the reciprocal model (c) is referred to as RI-RCLPM (random intercept reciprocal cross-lagged panel model). Bayes estimation is not available in Mplus for the reciprocal models but they can be estimated using WLSMV. As pointed out in Muthén and Asparouhov (2023), for these models it is important to allow for non-symmetric confidence intervals which can be obtained in the WLSMV context using bootstrapping. The RI-RCLPM also needs to apply parameter constraints to obtain admissible parameter estimates. The analysis uses a restriction applied to the reciprocal effects held time invariant as described in Muthén and Asparouhov (2023) to avoid dual solutions and negative R-square (this is referred to as restriction a). The results are presented as model 5 in Table 8, showing that the model fits well and is more parsimonious than the RI-CLPM. Model 6 is the RI-RLPM which does not include cross-lagged effects (model type (b)) and uses time-invariant lag 0 effects. This model also fits well and is more parsimonious than model 5. Using bootstrapping, model 6 shows insignificant lag 0 effects of stress on non-abstinence but significant effects of non-abstinence on stress with standardized estimates ranging from 0.18 to 0.25 (the standardized effects vary despite invariant lag 0 effects due to varying latent response variable variances). Based on this finding, model 7 uses the single-direction, time-invariant lag 0 model (model type (d)) which also fits well. The model 7 lag 0 effect of non-abstinence on stress has standardized effects ranging from 0.22 to 0.29. The single-direction lag 0 model which instead uses the reverse effect of stress on non-abstinence fits the same and has a significant lag 0 effect but the standardized effects are much smaller, ranging from 0.06 to 0.08 (not shown). Model 7 does not have any significant cross-lagged effects and they are excluded in model 8 which is the most parsimonious of the eight models. The model 8 lag 0 effects of non-abstinence on stress have standardized effects ranging from 0.18 to 0.24. Freeing the first residual variance for the abstinence variable as discussed in Section 3 gave a larger standard error, did not result in an improvement in fit, and did not change the magnitude of the standardized lag 0 effect estimates.

The set of analyses in Table 8 indicate that there is an effect of non-abstinence on stress rather than the other way around. In line with the conclusions of Muthén and Asparouhov (2023), the time lag for the effect is, however, difficult to establish. The more traditional model 4 states that the effect has lag 1 which is mostly a time interval of two weeks in these data. In contrast, model 8 states that the effect is contemporaneous. Although model 8 is more parsimonious, both model 4 and model 8 fit the data well. There is not a strong statistical argument for choosing between the models. The models are not nested due to the differences of including residual covariances or not and including lag 0 effects or not. Therefore chi-square difference testing cannot be applied. It may be disappointing to not be able to determine the lag of the effect, but that is the nature of the design of data collection. More importantly, the direction of the effect has been determined.

6 Ordinal outcome

This section turns to the case of ordinal outcomes such as for the alcohol risk outcome in Figure 8. The unrestricted probit model is often rejected for ordinal outcomes when there is a strong floor effect. When applying the unrestricted probit model to the alcohol risk variable, five response patterns have significant standardized residuals with an especially strong misestimation of the most common response pattern of abstinence at all eight time points where the observed frequency 312 obtains the estimate 267. The total number of bivariate cells is $25 \times 8(8-1)/2 = 700$ of which 273 cells show significant standardized residuals which is far greater than the 35 suggested by the 5% threshold. It is clear that an alternative model is needed for this variable.

The bivariate probit model for two ordinal variables was shown in Figure 3. With C categories, the number of parameters in the model is C1-1+C2-1+1 (2 sets of thresholds + 1 polychoric correlation) = C1+C2-1 parameters. For C1=C2=5, this adds to 9 parameters. The unrestricted multinomial model for the two variables has C1*C2-1 parameters which for C=5 adds to 24 parameters. This is the model that is tested against in the bivariate frequency tests and where model misfit is often found. However, intermediate models are possible. This paper uses a model which has C1+C2 parameters, that is, adding one parameter to the unrestricted probit model for two variables. For C=5, it has 10 parameters. Despite adding only a single parameter, this model fits the data considerably better. The model will be referred to as the two-part ordinal model.

6.1 Two-part ordinal model

The two-part ordinal model is inspired by two-part regression modeling of semicontinuous outcomes proposed by Duan et al. (1983) and two-part growth modeling with semicontinuous outcomes in Olsen and Schafer (2001); see also two-part growth mixture modeling in Muthén (2001). For ordinal outcomes, the model draws on the two-part regression analysis with an ordinal outcome that was used in Muthén, Muthén and Asparouhov (2016). The model is suitable for outcomes that have a strong floor effect as seen for the alcohol risk variable. The idea of the model is shown in Figure 10 for the case of a 4-wave growth model with random intercept and random slope growth factors. A 5-category ordinal variable is split into an ordinal part p (positive categories) for individuals who are above the floor value and a binary part b defined by being at the floor value (b=0) or above it (b=1). The ordinal outcome is missing when the binary outcome is zero. A strength of the two-part model is that the two parts can have different relations to covariates as indicated by the x variable in the figure. For example, a treatment dummy variable can have different influence on the two parts. Each part uses a probit model with continuous latent response variables specified for the binary variable and the ordinal variable. For the ordinal part, there are C-2 thresholds and for the binary part there is one.

The two-part model of Figure 10 corresponds to a single outcome, where the ordinal and binary parts are correlated only via their random intercepts. The univariate two-part ordinal model with random intercepts can be expressed as follows for the binary part b and ordinal part p in terms of the corresponding latent response variables,

$$b^*_{it} = \alpha_{b_i} + \epsilon_{b_{it}}, \tag{17}$$

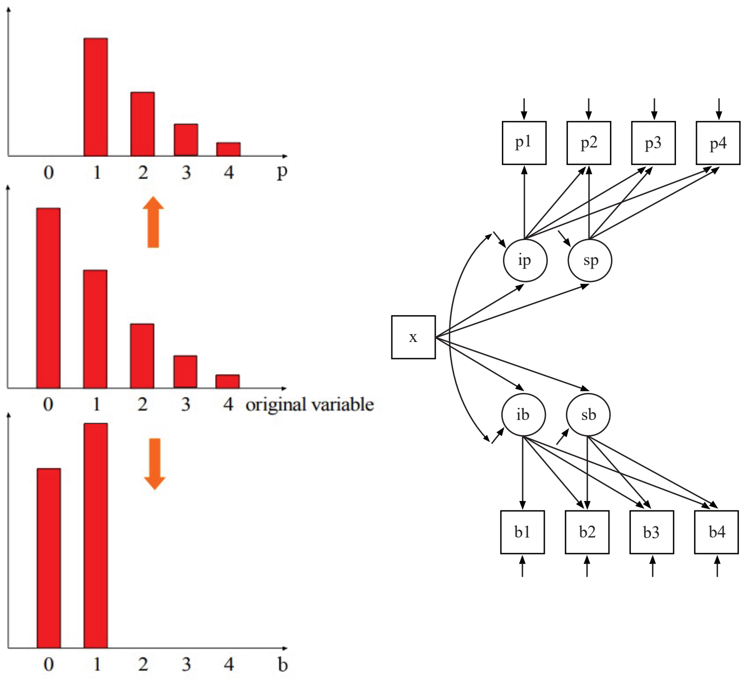
$$p^*_{it} = \alpha_{p_i} + \epsilon_{p_{it}}, \tag{18}$$

where $\alpha_{bi} \sim (N, 0, \psi_b)$ and $\alpha_{pi} \sim (N, 0, \psi_p)$ denote random intercepts that are correlated, and the normally distributed ϵ residuals typically have auto-regressions as before in Equation (14). Here, b^*_{it} has a single threshold τ_{bt} for each timepoint t while for an outcome with C categories, p^*_{it} has $C-2$ thresholds for each time point t, where for the ordinal variable p_{it} observed in category c,

$$p_{it} = c \Leftrightarrow \tau_{c-1,t} \leq p^*_{it} < \tau_{c,t}. \tag{19}$$

As before, however, a first step is to test the fit of the unrestricted multivariate normal probit model. For the unrestricted two-part ordinal model, there are no random

Figure 10: Two-part model



intercepts and correlations are instead considered between the latent response variables for the two parts themselves. The exception is correlation between the two parts at the same time point which are not well determined given that the ordinal part is never observed when the binary part is zero; these concurrent correlations are fixed at zero in the estimation of the unrestricted model.

For a bivariate RI-CLPM of ordinal outcomes, there can in principle be four model parts, an ordinal and a binary for each of the two outcomes. The four processes would be correlated via their random intercepts only. In the current application, a two-part model is considered for the alcohol outcome but is not needed for the continuous stress outcome. This model therefore has three correlated random intercepts.

The two-part ordinal model is estimated by Bayes which properly handles the missing data on the ordinal outcome when the binary outcome is zero. ML would also handle the missing data correctly, but as before, ML involves numerical integration with too many dimensions to be practical. Because WLSMV is a limited-information estimator, it cannot handle this special missing data situation properly.

7 Ordinal outcome example

The univariate analysis of the T = 8 alcohol risk outcomes involves eight ordinal and eight binary variables when applying the two-part ordinal model. Table 9 shows univariate analysis results using both the regular ordinal probit model and the two-part ordinal probit model. Model 1 is the unrestricted regular ordinal probit model. This model is not appropriate for the data in that it has strong indications of misfit

Table 9: Univariate analysis of 5-category alcohol risk using regular and two-part ordinal models (N=1375, T=8)

	Model	# par's	PPP/ χ^2	# Significant Residuals Resp Pattern*	Bivar	Comment
			Regular ordinal probit			
1.	Unrestricted	60	0.498	5 (312)	273 (39%)	Poor fit
2.	AR2	45	0.151	5 (312)	277	Poor fit
3.	RI-AR2	46	0.135	5 (312)	274	Poor fit
			Two-part ordinal probit			
4.	Unrestricted	144	0.472	1 (12)	29 (4%)	Good fit
5.	AR2	58	0.145	2 (46, 13)	106 (15%)	Poor fit
6.	RI-AR2	61	0.228	1 (12)	52 (7%)	OK fit

* Observed frequency in parentheses.

with a significant standardized residual for 5 response patterns including the most frequent response pattern of 312 with individuals being in the abstinence category at all time points. Model 1 also has 273 significant standardized residuals for the bivariate frequency tables which is 39% of all 700 bivariate cells.[9] The more restricted models 2 and 3 are consequently also misfitting. These results show the need for the two-part ordinal model.

Turning to the two-part ordinal probit models, model 4 is the unrestricted model and, unlike model 1, model 4 shows a good fit, making it possible to proceed with the more restrictive models 5 and 6. Model 5 is the lag 2 model without random intercepts for the two processes and it fits poorly. Adding the random intercepts of model 6, the three extra variance-covariance parameters produce a much better fit. Model 6 has sizeable variances for the two random intercepts. The random intercept R-squares range from 0.3 to 0.6 with slightly higher values for the binary part. In contrast, model 3 has ignorable random intercept variance and R-squares about 0.04.

The bivariate analysis of stress and alcohol risk with a two-part ordinal representation involves eight continuous, eight ordinal, and eight binary variables. The RI-AR2 model is used for both stress and alcohol risk. The three random intercepts are correlated and lag 1 cross-lagged effects are allowed among all three sets of variables. Concurrent residual correlations are allowed between the continuous variables on the one hand and the ordinal and binary variables on the other hand. Concurrent residual correlations are not included between the ordinal and binary variables due to lack of information since the ordinal variable is missing for a binary variable of zero.[10] This RI-CLPM has 137 parameters. It obtains a good fit with PPP = 0.181, 1 signifi-

[9] The bivariate frequency testing of the two-part ordinal model using TECH10 was introduced in Mplus in Version 8.8.

[10] Mplus uses the Gibbs random walk algorithm for this Bayesian analysis.

cant standardized residual for the response pattern with frequency 12, and 52 (7%) significant standardized residuals for the bivariate frequency tables.

The RI-CLPM with a two-part ordinal representation of alcohol risk has 5 out of 7 significant cross-lagged effects for the binary part influencing stress and 6 out of 7 significant cross-lagged effects for the ordinal part influencing stress. The standardized effects range from 0.08 to 0.22 for the binary part and 0.16 to 0.29 for the ordinal part. For the cross-lagged effects of stress influencing the two parts of alcohol risk, 1 out of 7 effects are significant for the binary part and zero for the ordinal part. The conclusion is that increased alcohol risk has a lagged positive effect on stress, with a larger effect for the ordinal part than the binary part. There is almost no evidence of lagged effects from stress to alcohol risk.

Reciprocal (lag 0) effect modeling using the two-part ordinal model can currently not be estimated using Bayes (WLSMV cannot handle two-part modeling as mentioned earlier). Single-direction lag 0 two-part ordinal modeling is, however, possible using Bayes. The lag0 effects are held time invariant. This model has 123 parameters and is therefore more parsimonious than the two-part ordinal RI-CLPM with 137 parameters. As expected, the model fits about the same in the two directions and has similar fit as the RI-CLPM. The effects from alcohol risk to stress are larger than from stress to alcohol risk for both the binary and ordinal parts. In standardized terms, the effects of binary risk on stress range from 0.18 to 0.22 and the effects from ordinal risk to stress range from 0.27 to 0.39. The effects from stress to binary risk range from 0.12 to 0.15 and the effects from stress to ordinal risk range from 0.14 to 0.19. These analyses do not give a clearcut conclusion of the direction of influence between the two outcomes but the effects are stronger in the alcohol to stress direction. As in the binary case, it is not possible to determine if the effects have lag 1 or lag 0.

8 Extensions

Multiple-group analysis makes it possible to study group differences in parameters of all the models discussed. Because group membership is typically time invariant, the influence of groups on the time-invariant random intercepts is of key interest, particularly group differences in their means. For instance, the zero means of the random intercepts in the two-part ordinal model of Equations (17) and (18) can instead be estimated while holding thresholds invariant across the groups. By fixing the random intercept means to zero for one group, the thresholds of that group are identified by that group's proportions, and this in turn identifies the random intercept means for the other groups through their observed proportions. Group differences in other parameters may also be studied such as cross-lagged or contemporaneous effects.

The COMBINE example has a special interest in group differences in the random intercept means of the alcohol risk variable. This double-blind randomized clinical trial of alcohol use disorder treatment has nine groups, one placebo group and eight groups with different combinations of medication and therapy. Each group consists of approximately 150 subjects.

Multiple-group two-part ordinal analysis of the nine groups adds the $(9 - 1)3 = 24$ parameters of the means of the three random intercepts to the previous section's bivariate two-part ordinal models for stress and alcohol risk. The random intercept means for the placebo group are fixed at zero as a comparison. The analyses use both the RI-CLPM version and the single-direction lag 0 models of the previous section. The models are estimated by Bayes. The RI-CLPM version uses 161 parameters and obtains PPP = 0.296 which is an improvement over the previous analysis with PPP = 0.181. The maximum number of significant bivariate cells for any of the nine groups is only 6 with a total of 23. The single-direction lag 0 model has 147 parameters with time-invariant lag 0 effects. As before, similar fit for lag 0 regression in both directions are obtained and the fits are similar to that of the RI-CLPM.

Because all three models fit well, the choice between the three models is not expected to matter in terms of estimating the treatment effects. The results show that neither the choice between cross-lagged versus lag 0 effect modeling, nor the choice of the direction of the lag 0 effect matters in terms of which groups have random intercept means significantly different from the zero value of the placebo group. Here, a negative mean represents a beneficial treatment outcome. No group has significant random intercept means for the stress outcome. Four groups have significant negative means for the random intercept of the binary part. Two of these four groups also have significant negative means for the random intercept of the ordinal part. The four groups with significant effects for abstinence are:

- Naltrexone
- Naltrexone + acamprosate
- Placebo + behavioral intervention
- Naltrexone + acamprosat + behavioral intervention

These treatments were also found to be the most effective in the analyses of the COMBINE study. The second and third treatment listed were the ones found to also decrease the risk of a higher degree of alcohol risk and can therefore be said to be the most successful treatments.

It is also possible to allow group differences in other model parameters such as the effect of alcohol on stress. The groups may be estimated with these parameters unconstrained and then tested for group invariance. A Wald chi-square test of invariance can be carried out based on Bayes estimates as described in Asparouhov and Muthén (2021b). Using the single-direction model with time-invariant lag 0 effect for both the binary and ordinal alcohol risk parts on stress, group invariance was rejected on the 5% level with $\chi^2(16) = 28.37$ (p = 0.0285).

A further extension is to add a growth model. For example, linear growth for the two-part ordinal model is shown in Figure 10 where random slopes *sp* and *sb* are added to the random intercepts. In the bivariate analysis of stress and alcohol risk using the two-part ordinal model for alcohol risk, there would be three growth models. Autoregressions among the three variables can be added. A multiple-group version of this model would allow the means of the random slopes to also vary across groups. In a model with no random slope, the random intercept refers to the level at all time points, in this case during the eight time points following the start of treatment. Including a random slope, a parameterization can be chosen so that the random intercept refers to the status at the last time point and treatment effects evaluated for those random intercept means. Analysis using this model did not change the above findings of which treatments had significant effects.

9 Conclusions

This paper presented modeling, testing, identification, and estimation for the case of binary and ordinal variables in cross-lagged panel modeling. Simulations showed that estimation with both Bayes and weighted least squares methods worked well given a sufficient sample size and a sufficient number of time points. A larger sample size and more time points are required than for continuous variables. A two-part ordinal model was proposed for ordinal variables with strong floor effects. Using a randomized study of alcohol treatment, the methods were used to examine the interaction between stress and alcohol use. Extensions to multiple-group analysis and modeling in the presence of trends were discussed.

The substantive results of the current study are consistent with preclinical data indicating that alcohol consumption increases subsequent stress, and that stress does not strongly predict lagged alcohol consumption. The time scale of weeks may be a limitation, given that some intensive longitudinal studies with heavy social drinkers have shown within day and day-to-day effects of stress on alcohol consumption (Armeli et al., 2000; Wemm et al., 2022). Future research should test the association between stress and alcohol use among individuals with alcohol use disorder using intensive longitudinal data collected during treatment.

Binary and ordinal variables also appear in panel studies as factor indicators. Random effects modeling of such a multiple-indicator case was studied in Muthén (1983, 1984). However, in the multiple-indicator case, the CLPM and RI-CLPM modeling still considers a continuous outcome, namely the factor, so that CLPM and RI-CLPM analysis can draw on the continuous-variable modeling of Mulder and Hamaker (2020).

Further methods for the joint longitudinal analysis of several categorical variables are discussed in Muthén and Asparouhov (2022). They include bivariate latent transition analysis and analysis with distal outcomes. With intensive longitudinal data where there are many time points spaced closely in time, dynamic structural equation modeling (DSEM; Asparouhov et al., 2018; Hamaker et al., 2023) is available for binary and ordinal outcomes. DSEM analyses are intended for data with $T \geq 20$ and are not suitable for the small number of time points in panel data that is considered here.

References

[1] Agresti, A. (2012). Categorical data analysis. Third edition. New York: John Wiley & Sons.

[2] Agresti, A. (2018). An introduction to categorical data analysis. Third edition. New York: John Wiley & Sons.

[3] Anton, R.F. and others (2006). Combined pharamacotherapies and behavioral interventions for alcohol dependence. The COMBINE study: A randomized controlled trial. Journal of the American Medical Association, 295, No. 17, 2003 - 2017.

[4] Armeli, S., Carney, M.A., Tennen, H., Affleck, G. & O'Neill, T.P. (2000). Stress and alcohol use: A daily process examination of the tressor-vulnerability model. Journal of Personality and Social Psychology.

[5] Asparouhov, T., Hamaker, E.L. & Muthén, B. (2018). Dynamic structural equation models. Structural Equation Modeling: A Multidisciplinary Journal, 25:3, 359-388, DOI: 10.1080/10705511.2017.1406803

[6] Asparouhov, T., & Muthén, B. (2020a). IRT in Mplus. Version 4. Technical report. www.statmodel.com.
`https://www.statmodel.com/download/MplusIRT.pdf`

[7] Asparouhov, T. & Muthén, B. (2020b). Comparison of models for the analysis of intensive longitudinal data. Structural Equation Modeling: A Multidisciplinary Journal, 27(2) 275-297, DOI: 10.1080/10705511.2019.1626733

[8] Asparouhov, T. & Muthén, B. (2021a). Bayesian analysis of latent variable models using Mplus. Version 5, September 18, 2021. `https://www.statmodel.com/download/BayesAdvantages18.pdf`

[9] Asparouhov, T. & Muthén, B. (2021b). Advances in Bayesian model fit evaluation for structural equation models, Structural Equation Modeling: A Multidisciplinary Journal, 28:1, 1-14, DOI: 10.1080/10705511.2020.1764360

[10] Asparouhov, T. & Muthén, B. (2022). Assessing model fit for SEM models with categorical variables via contingency tables. Technical Report. `https://www.statmodel.com/download/Tech10.pdf`

[11] Asparouhov, T. & Muthén, B. (2023). Residual structural equation models. Structural Equation Modeling: A Multidisciplinary Journal, 30, 1-31.
DOI: 10.1080/10705511.2022.2074422

[12] Becker, H.C. (2017). Influence of stress associated with chronic alcohol exposure on drinking. Neuropharmacology, Aug 1;122:115-126.
doi: 10.1016/j.neuropharm.2017.04.028. Epub 2017 Apr 19.

[13] Chi, E.M. & Reinsel, G.C. (1989). Model for longitudinal data with random effects and AR(1) errors. Journal of the American Statistical Association, 84, 406, 452-459.

[14] Christoffersson, A: Factor analysis of dichotomized variables. Psychometrika, 1975, 40, 5-32

[15] de Ayala, R.J. (2022). The Theory and Practice of Item Response Theory. Second edition. Medicine and Health Science books.

[16] Duan, N., Manning, W.G. Morris, C.N. & Newhouse, J.P. (1983). A comparison of alternative models for the demand for medical care. Journal of Business & Economic Statistics, 2, 283-289.

[17] Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychological Methods, 9(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

[18] Foldnes, N. & Gronneberg, S. (2022). The sensitivity of structural equation modeling with ordinal data to underlying non-normality and observed distributional forms. Psychological Methods, 27(4), 541-567.

[19] Hamaker, E. L., Kuiper, R. M., Grasman, R. P. (2015). A critique of the cross-lagged panel model. Psychological Methods, 20, 102-116.

[20] Hamaker, E.L., Asparouhov, T. & Muthén, B. (2023). Dynamic structural equation modeling as a combination of time series modeling, multilevel modeling, and structural equation modeling. Chapter 31, pp. 576-596 in: The Handbook of Structural Equation Modeling (2nd edition); Rick H. Hoyle (Ed.); Publisher: Guilford Press.

[21] Hamaker, E. L. (2023). The within-between dispute in cross-lagged panel research and how to move forward. Forthcoming in Psychological Methods.

[22] Hedeker, D., & Gibbons, R. D. (1994). A random effects ordinal regression model for multilevel analysis. Biometrics, 50, 933–944.

[23] Hedeker, D., & Mermelstein, R.J. (2000). Analysis of longitudinal substance use outcomes using ordinal random-effects regression models. Addiction, 95 (Supplement 3), S 381-394.

[24] Ialongo, N. (2022). Personal communication.

[25] Li, C.H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. Psychological Methods, 21(3), 369-87.

[26] Long, S. (1997). Regression models for categorical and limited dependent variables. Thousand Oaks: Sage.

[27] McHugh et al. (2013). Positive affect and stress reactivity in alcohol-dependent outpatients. J. Studies in Alcohol and Drugs.

[28] Mulder, J.D. & Hamaker, E.L. (2020). Three extensions of the Random Intercept Cross-Lagged Panel Model. Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2020.1784738

[29] Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

[30] Muthén, B. (1983). Latent variable structural equation modeling with categorical data. Journal of Econometrics, 22, 48-65.

[31] Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

[32] Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K.A. Bollen, & J.S. Long (Eds), Testing Structural Equation Models (pp. 205-243). Newbury Park, CA: Sage. http://www.statmodel.com/bmuthen/articles/Article_045.pdf

[33] Muthén, B. (1996). Growth modeling with binary responses. In A.V. Eye & C. Clogg (eds.). Categorical variables in developmental research: Methods of analysis (pp. 37-54). San Diego, CA: Academic Press. http://www.statmodel.com/bmuthen/articles/Article_064.pdf

[34] Muthén, B., du Toit, S.H.C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished technical report. http://www.statmodel.com/download/Article_075.pdf

[35] Muthén, B. (2001). Two-part growth mixture modeling. Unpublished report. http://www.statmodel.com/download/2PGMM.pdf

[36] Muthén, B. & Asparouhov, T. (2016). Multi-Dimensional, Multi-Level, and Multi-Timepoint Item Response Modeling. In van der Linden, W. J., Handbook of Item Response Theory. Volume One. Models, pp. 527-539. Boca Raton: CRC Press

[37] Muthén, B. & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Mplus Web Note No. 4. https://www.statmodel.com/download/webnotes/CatMGLong.pdf

[38] Muthén, B., Muthén, L, & Asparouhov, T. (2016). Regression and mediation analysis using Mplus. Los Angeles, CA: Muthén & Muthén.

[39] Muthén, L.K. & Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

[40] Muthén, B. & Asparouhov, T. (2022). Using Mplus to do cross-lagged modeling of panel data, part 2: Categorical variables. Mplus Web Talk No. 4, Part 2. http://www.statmodel.com/Webtalk4P2.shtml

[41] Muthén, B. & Asparouhov, T. (2023). Can cross-lagged panel modeling be relied on to establish cross-lagged effects? The case of contemporaneous and reciprocal effects. Forthcoming in Psychological Methods.

[42] Olsen, M.K. & Schafer, J.L. (2001). A two-part random effects model for semicontinuous longitudinal data. Journal of the American Statistical Association, 96, 730-745.

[43] Rhemtulla, M., Brosseau-Liard, P.E. & 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.

[44] Sinha, R. (2022). Alcohol's negative emotional side: The role of stress neurobiology in alcohol use disorder. Alcohol Research. Current Reviews. Volume 42, issue 127.

[45] Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. Psychometrika, 52(3), 393-408.

[46] Wemm, S. E., Tennen, H., Sinha, R., & Seo, D. (2022). Daily stress predicts later drinking initiation via craving in heavier social drinkers: A prospective in-field

daily diary study. Journal of Psychopathology and Clinical Science, 131(7), 780–792. https://doi.org/10.1037/abn0000771