Fudan Journal of the Humanities and Social Sciences
https://doi.org/10.1007/s40647-025-00447-2

ORIGINAL PAPER

Sample Size Planning and Power Analysis for Detecting Cross-lagged Effects in Longitudinal Studies with Ordinal Outcomes

Sijing Shao[1] · Ziqian Xu[2] · Wen Qu[3] · Ross Jacobucci[4]

Received: 28 September 2023 / Revised: 7 April 2025 / Accepted: 11 April 2025

Abstract

This paper investigates how multilevel models (MLMs) handle hierarchical and longitudinal data, such as repeated measures nested in individuals, which are common in social science research. Effective sample size planning is critical for MLMs, with power analysis serving to determine the necessary sample sizes. However, research on sample size planning for MLMs with ordinal outcomes is limited, despite its increasing popularity for applied researchers. Additionally, many studies examine the cross-lagged effects, which trace how changes in one variable at an earlier time influence another variable later. To address these issues, we conducted a simulation study to investigate how sample size, the autoregressive (AR) effect, and cross-lagged effects influence statistical power within a multilevel autoregressive framework. The results provide practical guidance for researchers designing longitudinal studies with ordinal outcomes. Furthermore, we developed **OrdPower**, an easy-to-use R package, which social science researchers can use to plan sample sizes for MLMs with ordinal outcomes, especially when cross-lagged effects are the primary focus.

Keywords Ordinal outcome · Multilevel autoregressive models · Sample planning · Power analysis · Cross-lagged effect

✉ Wen Qu
wqu@fudan.edu.cn

[1] Department of Psychology, Cornell University, Ithaca, USA

[2] Department of Psychology, University of Notre Dame, Notre Dame, USA

[3] Fudan Institute for Advanced Study in Social Sciences, Fudan University, 220 Handan Road, Shanghai 200433, China

[4] Center for Healthy Minds, University of Wisconsin-Madison, Madison, USA

Published online: 05 June 2025

1 Introduction

In social science research, particularly in psychological studies, tracking temporal changes is crucial for understanding the dynamics in process-oriented mechanisms. Intensive time sampling methods, such as ecological momentary assessments (EMA; Shiffman et al. 2008), offer a more detailed perspective on these changes compared to traditional cross-sectional approaches. As a result, the use of repeated measures designs has increased, introducing complex nested data structures that require advanced analytical techniques.

Multilevel models (MLMs; Grimm, Ram, & Estabrook, 2016; Raudenbush & Bryk 2002; Singer, Willett, Willett, & others, 2003) are widely used statistical methods for addressing the nested data structures inherent in repeated measures. In MLMs, the dependence of observations within individuals is accounted for by specifying person-level random effects. However, these models are not without their challenges. In longitudinal studies, residuals may exhibit temporal autocorrelation, which violates the independence assumption central to traditional MLMs. As a result, standard MLM approaches may fail to address these temporal correlations, potentially leading to biased standard error estimates and misleading inferences.

One approach to handle this issue is to specify the covariance structure (Fitzmaurice, Laird, & Ware, 2012; Verbeke 1997), which, when correctly applied, refines the standard error estimations (Wang et al. 2019). Another approach is to incorporate autoregressive effects to examine *inertia*, reflecting how a variable's current value influences its future values, thus accounting for temporal dependencies within the data (Hamaker & Grasman 2015).

Psychological researchers frequently examine the relationship among different variables over time (Hayashi, Yuan, & Bentler, 2024; Li & Hu 2024; Liao, Song, & Bard, 2024; Lu 2025; Wu, Ram, Marks, Streeper, & Conroy, 2024). Cross-lagged effects improve multilevel models by examining the directional relationships between variables across time points, which is essential for understanding causal dynamics in longitudinal data. However, conventional vector autoregressive (VAR) models, what are typically applied to single-subject time-series data, are less suited for psychological research, which often involves studying multiple individuals. To overcome this limitation, researchers have developed multilevel autoregressive models (MLM AR; Hamaker, 2015; Bolger et al., 2013) within the multilevel models (Bolger & Laurenceau 2013) framework. These models enable the simultaneous examination of both within-person dynamics and between-person differences, incorporating cross-lagged effects to provide a more nuanced understanding of intra-individual changes and inter-individual differences.

2 Sample Size Planning

As MLM is a popular framework handling longitudinal data, determining the sample size for each level during the study design phase is crucial. In nested data, sample sizes must be planned for at least two levels: (1) the number of clusters (i.e., participants) and (2) the number of assessments within each cluster (i.e., repeated measurements for each participant). Within the MLM framework (Bolger & Laurenceau 2013; Snijders & Bosker 2011), various methods have been developed to perform power analyses, though most primarily focus on cases with normally distributed error terms.

However, in many psychological studies, particularly those using ordinal scales like Likert-scale questions, the assumption of normally distributed error terms presents a challenge. This issue arises for several reasons. First, while responses on ordinal scales are numerically coded, they do not represent true continuous variables; the intervals between scale points are uneven, making the normality assumption inappropriate. Second, ordinal data often exhibit non-normal distributions, including ceiling or floor effects, which violate the assumption of normally distributed errors. Consequently, linear models fail to adequately account for these effects, which can lead to inaccurate predictions. Furthermore, in situations involving skewed data, which are common in clinical settings, assuming normality can lead to biased coefficient estimates (McKelvey & Zavoina 1975), thus distorting research conclusions (Winship & Mare 1984). Addressing these issues is crucial to ensure the accuracy and validity of psychological research, particularly in studies involving ordinal outcomes.

Although the importance of using ordinal distributions in MLMs (Bauer & Sterba 2011; Liu & Agresti 2005) is becoming more widely understood, few studies have investigated sample size planning for MLMs with correctly defined categorical outcomes. Although Kumle et al. (2021) provided guidance on power estimation for generalized linear mixed models, it does not specifically address ordinal outcomes. When the outcome is categorical, the assumption of normally distributed error terms generally does not hold. Moineddin et al. (2007) conducted simulations to examine the effects of varying sample sizes when the outcomes are binary, while Ali et al. (2016) compared estimation methods in power analyses for ordinal outcomes. These studies laid the groundwork for sample size planning in ordinal MLMs but did not extend to cross-lagged effects, which are particularly important in longitudinal studies. Therefore, it is crucial to develop an approach and provide practical guidelines for sample planning when estimating the cross-lagged effects is the primary research focus.

Additionally, in addition to assuming outcome variables are normally distributed, the random effects—which allow the autoregressive or cross-lagged parameters to vary across individuals—are typically assumed to follow a normal distribution. However, in clinical research, individual behaviors and patterns of change can deviate significantly from this assumption, leading to skewed random effect distributions. For example, in suicide research, a subset of individuals often exhibit extreme levels of suicide risk, resulting in skewed distributions of key

covariates. Such non-normal patterns can cause different subgroups to emerge, each with distinct predictor–outcome relationships, thereby producing a skewed distribution of random slopes. Thus, assuming that random effects follow a normal distribution may be unrealistic in these contexts. Ignoring this violation can potentially lead to biased parameter estimates and misestimated random effects variances. To our knowledge, the effects of violating the normality assumption for random effects on statistical power in multilevel models, especially in estimating cross-lagged effects, have not been thoroughly examined and remain unclear. This paper aims to address this research gap, by providing insights into sample size planning and power analysis when random effects deviate from normality in psychological research with cross-lagged effect.

Our contributions in this paper are threefold: (1) illustrate the generalized version of multilevel autoregressive models with ordinal outcomes; (2) conduct a simulation study to assess the sufficient sample sizes required to detect different levels of cross-lagged effects under various conditions in multilevel model with ordinal outcomes, while also evaluating the impact on model estimation; and (3) present an easy-to-use tool called **OrdPower** (Shao, Xu, & Jacobucci, 2023) in R for researchers who wish to perform sample size planning for ordinal outcomes under the multilevel framework.

3 Method

To investigate cross-lagged effects in longitudinal studies with ordinal outcomes, we use multilevel autoregressive models. First, we present the basic MLM AR and its extension for ordinal outcomes. Next, we introduce a generalized MLM AR that incorporates cross-lagged effects. This model serves as the main framework for our study. We also introduce a newly developed R package, designed to assist empirical researchers in conducting sample size planning for using this model.

3.1 Multilevel Autoregressive Models (MLM AR)

The outcome variable y_{it} for person i at time t is modeled at level 1:

$$y_{it} = \beta_{0i} + \beta_{1i,t-1} + \in_{it}, \tag{1}$$

where β_{0i} denotes random intercept at an individual level, and β_{1i} denotes random AR parameter, representing the inertia of the individual i. The β_{1i} ranges from -1 to 1 to ensure stationarity of the process (Hamilton 2020), with estimates typically falling between 0 and 0.6 in psychological studies (Wang et al. 2012).

An AR parameter closer to 0 indicates that past events have less impact on future ones, making it easier for the individual to return to their baseline level. For example, when an individual with low inertia (i.e., β_{1i} close to 0) might experience a sudden increase in suicidal thoughts, but would quickly return to their usual mental state (e.g.,

the expected value of y_{it}). In contrast, someone with higher inertia (a larger β_{1i}) would have more difficulty bouncing back and would remain affected longer.

For simplicity, we consider only the lag-1 AR parameter in this article. Higher lags can be easily included by extending Eq. 1 with terms such as $\beta_{ki}y_{i,t-k}$, where k denotes kth lag. Both β_{0i} and β_{1i} are allowed to vary across individuals through the following level 2 equations:

$$\beta_{0i} = \beta_{00} + \mu_{0i}, \tag{2a}$$

$$\beta_{1i} = \beta_{10} + \mu_{1i}, \tag{2b}$$

where β_{00} is the fixed average intercept across individuals, and μ_{0i} accounts for the variability in the random intercepts, with the standard deviation of μ_{0i} denoted by$\sigma_{\mu_{0i}}$. Similarly, the fixed AR parameter β_{10} represents the average inertia and the variability across people is accounted for by μ_{1i}, with its standard deviation denoted by $\sigma_{\mu_{1i}}$.

3.2 MLM AR with Ordinal Outcome

When the longitudinal outcome variable is ordinal, several approaches can be used, including cumulative model, sequential model, and adjacent category model (Bürkner & Vuorre 2019; Jacobucci et al. 2021). Among these models, the cumulative model is the most popular one, and it is the primary focus of this paper.

Cumulative Model: The cumulative model framework assumes that the observed ordinal outcome Y is categorized from an underlying latent continuous variable $\tilde{Y}$ and the categorization is based on some thresholds. When there are K categories in the outcome variable, $K-1$ thresholds are needed for categorization. For example, when the outcome variable has five levels of the Likert-scaled item, $\tilde{Y}$is partitioned based on four thresholds a_1, a_2, a_3, and a_4. The underlying latent variable $\tilde{Y}$is assumed to follow a normal distribution with cumulative distribution function F.

The probability of $\tilde{Y}$being categorized into the kth category can be modeled as:

$$P(Y = k) = F(a_k) - F\left(a_{k-1}\right) \tag{3}$$

with the categorization

$$y_{it} = 1 \; if \; \tilde{y}_{it} < a_1 \tag{4a}$$

$$y_{it} = 2 \; if \; a_1 \leq \tilde{y}_{it} < a_2 \tag{4b}$$

$$y_{it} = 3 \; if \; a_2 \leq \tilde{y}_{it} < a_3 \tag{4c}$$

$$y_{it} = 4 \; if \; a_3 \leq \tilde{y}_{it} < a_4 \tag{4d}$$

$$y_{it} = 5 \; if \; \tilde{y}_{it} \geq a_4 \tag{4e}$$

The underlying latent variable $\tilde{Y}$ *it* is standard normally distributed and modeled at level 1 in multilevel AR models:

$$\tilde{y}_{it} = \beta_{0i} + \beta_{1i,} \cdot \tilde{y}_{i,\,t-1} + \in_{it}, \tag{5}$$

where β_{0i} (random intercept) and β_{1i} (random AR parameter) vary across individuals, as specified in the level 2 Eqs. (2a) and (2b), discussed earlier in MLM AR section. Additionally, for a five-level Likert scale, four threshold parameters must be estimated alongside the model coefficients.

In the context of ordinal outcomes, the intraclass correlation coefficient (ICC) represents the proportion of variance at level 2 relative to the total variance (Hox, Moerbeek, & Van de Schoot, 2017). For guidance on sample planning in multilevel models with ordinal outcomes when the ICC effect is of interest, readers may refer to Ali et al. (2016).

3.3 Cross-lagged Effects

The MLM AR can be expanded to include VAR models by incorporating past lags of other variables. When the outcome variable is ordinal, cross-lagged effects can be incorporated into models (5) by including cross-lagged variables X_{kp}, where k denotes the number of cross-lagged variables, and p denotes the number of lags. In clinical studies, only lag-1 is typically considered, and for illustration purposes, we consider only one cross-lagged variable. The simplified VAR model in the multilevel framework can be specified as:

$$\tilde{y}_{it} = \beta_{0i} + \beta_{1i}\tilde{y}_{i,t-1} + \beta_{2i}x_{i,t-1} + \epsilon_{it}, \tag{6}$$

where β_{2i} is the cross-lagged coefficient and allows to vary across people by:

$$\beta_{2i} = \beta_{20} + \mu_{2i}, \tag{7}$$

where β_{20} is the fixed cross-lagged effect that is the same for all individuals, and the variability of the cross-lagged effect across people is accounted for by μ_{2i} with its standard deviation denoted by $\sigma_{\mu_{2i}}$. The random effects of the intercept and AR(1) are modeled in level 2 (2a) (2b). The path diagram of the MLM AR with cross-lagged effect model is in Fig. 1.

3.4 R Package

An R package called **OrdPower** was created for empirical researchers who plan to perform sample planning for longitudinal ordinal data, with a focus on testing the cross-lag effect. The package can be installed via GitHub. The package will also be available in CRAN in the near future. An example is in Appendix.

```
library( devtools)
install_github ("shaosijing/OrdPower")
library(OrdPower)
```

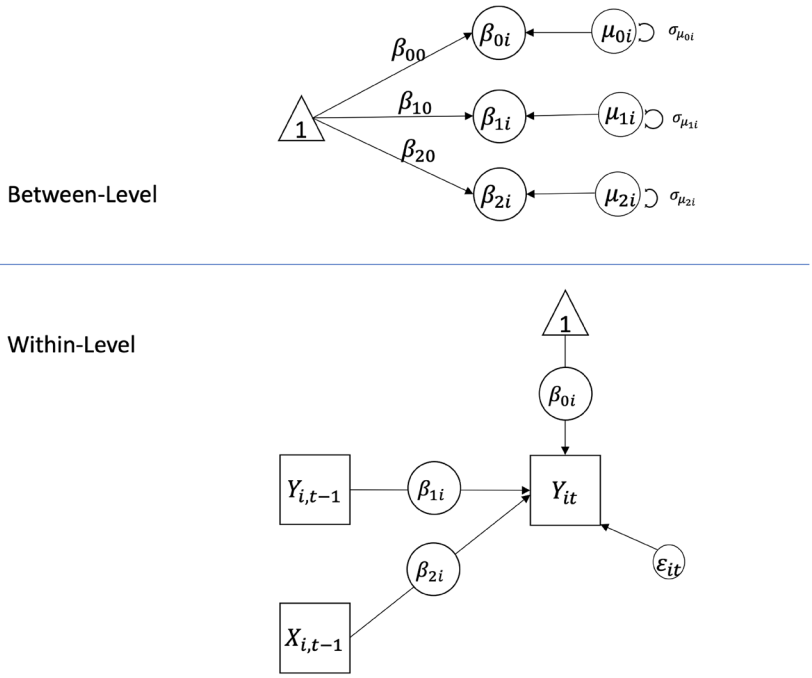

Fig. 1 Path diagram of multilevel AR model with cross-lagged effect. Note: For simplicity, we only include the observed ordinal Y_{it}, which are generated from the latent continuous $\tilde{Y}_{it}$

The function ord_power within the **OrdPower** package can be used to obtain power values given expected parameters. Applied researchers can check the power with different combinations of the number of participants and assessments. Arguments used in the function ord_power include the following.

- n. The number of categories for the ordinal outcome variable.
- numSample. The number of participants.
- numAssess. The number of assessments.
- thresh_CON. The marginal distribution with options of 1=bell-shaped, or 2=right-skewed.
- autoreg_coeff. The fixed AR(1) effect.
- crosslag_coeff. The fixed cross-lag effect.
- gamma_00_sd. The standard deviation of the random intercept.
- gamma_01_sd. The standard deviation of the random AR(1) effect
- gamma_02_sd. The standard deviation of the random cross-lag effect.
- corr. Whether the random effects are correlated or not.

Table 1 Simulation Conditions

Variables	Labels	Levels
number of participants	N	30, 50, 80
number of assessments	T	28, 42, 100
marginal distribution	MargDist	(1) bell-shaped, or (2) right-skewed
AR(1) fixed effect	010	low (0.2), medium (0.5), high (0.8)
cross-lagged fixed effect	020	0, 0.1, 0.2, 0.3
random intercept heterogeneity	$\sigma\mu_{0i}$	1,2,3,4
AR(1) heterogeneity	$\sigma\mu_{1i}$	0, 0.1, 0.3, 0.5
cross-lagged heterogeneity	$\sigma\mu_{2i}$	0, 0.1, 0.3, 0.5
cross-lagged normality violation	CLViol	(1) cross-lagged normality violation variable level block (2)medium violation with moderately right-skewed (3) high violation with highly right-skewed (1) AR(1) normality violation block
AR(1) normality violation	ARViol	(2)medium violation with moderately right-skewed (3) high violation with highly right-skewed
correlation	corr	(1) random effects are correlated, or (2) not correlated

3.5 Simulation Study

This simulation study focuses primarily on sample size planning[1] for detecting cross-lagged effects in longitudinal MLM AR with ordinal outcomes. Additionally, we examine the influence of several other parameters to provide a comprehensive understanding of factors affecting statistical power in these models.

3.6 Simulation design

Our simulation design examines the effects of the number of participants (*N*) and the number of assessments (*T*) on the power to detect cross-lagged effects. Beyond these factors, we investigate nine additional parameters critical to model performance. All simulation conditions, along with their corresponding labels, are summarized in Table 1.

The selection of condition levels was informed by empirical research on high-risk behaviors in intensive longitudinal data collection, such as in studies by Jacobucci et al. (2023) and Ammerman et al. (2022), ensuring that our simulation reflects real-world scenarios. The following subsections provide specific details regarding these 11 conditions.

[1] In the simulation study, *sample size* refers to both the *number of participants* and the *number of assessments* within the MLM framework.

Sample Sizes: We considered three levels of participants ($N = 30, 50, 80$) and three levels of measurement assessments ($T = 28, 42, 100$)[2], reflecting conditions typically encountered in longitudinal psychological studies.

Marginal Distribution Shapes: We manipulated the thresholds for the ordinal outcomes (i.e., five categories in this study) to simulate two different marginal distribution shapes:

- **Bell-shaped distribution**: Thresholds set at $-1.5, -1, 1, 1.5$, creating higher probabilities for central categories.
- **Skewed distribution**: Thresholds set at 0, 1, 2, 3, where the lower categories have higher probabilities.

These distribution shapes are reflective of common clinical and psychological scales, where outcomes may not always be symmetrically distributed. These adjustments enabled us to evaluate how different marginal distributions affect model performance.

AR(1) Fixed Effect: The autoregressive coefficients were examined at three levels: 0.2, 0.5, and 0.8, representing low, medium, and high levels of individual inertia, respectively.

Cross-Lagged Fixed Effects: Cross-lagged coefficients were examined at four levels: 0 (no effect), 0.1, 0.2, and 0.3. In conditions where the cross-lagged coefficient was set to 0, the corresponding random effect for cross-lagged terms was not considered either. This allowed us to examine the model's ability to detect cross-lagged effects.

Heterogeneity of effects: The level of heterogeneity among individuals is manipulated by varying the standard deviations of the random AR(1) and cross-lagged effects,

$\sigma_{\mu_{1i}}$ and $\sigma_{\mu_{2i}}$, across four levels: 0, 0.1, 0.3, and 0.5. The standard deviation for the random intercept, $\sigma_{\mu_{0i}}$, has four levels: 1, 2, 3, and 4.

Normality Violations of AR(1) and Cross-Lagged Effects: To evaluate the robustness of statistical models under realistic data conditions, we introduced three levels of normality violations in the AR(1) and cross-lagged effect distributions. These were modeled using distinct levels of skewness:

- No or minimal violation: The distribution is approximately normal, with skewness set to 0.
- Moderate violation: A moderately right-skewed distribution with skewness set to 0.7.
- High violation: A highly right-skewed distribution with skewness set to 1.5.

Skewed distributions in moderate and high levels are especially relevant in clinical settings, where many participants might not experience an effect, leading to zero-inflation and right-skewed data. These varied skewness levels allowed us

[2] Because clinical assessments typically use weeks as the time unit, 28 days correspond to 4 weeks, while 42 days correspond to 6 weeks.

to test the robustness of statistical models under different and realistic assumption violations commonly encountered in clinical and psychological research. The R package *covsim* (Grønneberg et al. 2022) was used to simulate these conditions, which allowed us to assess the impact of assumption violations on the performance of the models.

Correlations: Correlations among the random effects μ_{0i} with μ_{1i}, μ_{1i} with μ_{2i}, and μ_{0i} with μ_{2i} are tested under two conditions: either set at −0.5, 0.5, and −0.5, or assumed to be uncorrelated.

In each condition, we conduct five hundred data simulations and analyses. The R package Ordinal (Christensen & Christensen 2015) is used for data analyses in R (R Core Team 2021).

4 Metrics

Power, type I error, relative bias (RB), bias, and root mean square error (RMSE) are the key metrics for assessing model performance. Power evaluates the model's ability to detect true effects, while type I error measures the rate of false positives. RB and bias indicate the accuracy of the effect estimates, with RB focusing on the proportional error, and bias showing the absolute difference between estimated and true values. RMSE captures the overall magnitude of error in the estimates. Detailed explanations of these metrics are provided below.

Power: When the cross-lagged effect exists (in 92.3% of the conditions), power is defined as the percentage of detecting the effect. In this study, it is calculated as shown in Eq. 8, where n represents the number of times the cross-lagged effect is detected out of 500 replications. A power value greater than 0.8 is considered high (Cohen 1992).

$$\text{power} = \frac{n}{500} \tag{8}$$

Type I Error: Type I error rate reflects the probability of falsely detecting a non-existent effect. In 7.7% of the conditions where the cross-lagged effect is absent, the number of times it is incorrectly identified as significant is denoted by n^*. A type I error close to 0.05 is considered acceptable. The type I error is calculated as:

$$\text{type } I \text{ error} = \frac{n*}{500} \tag{9}$$

Relative Bias: When the cross-lagged fixed effect exists (i.e., $\beta_{20} \neq 0$) the accuracy of estimating the effect is obtained through the *RB* as shown in Eq. 10. Researchers frequently consider an RB of less than 5–10% to be acceptable. In Eq. 10, $\widehat{\beta}_{20}$ is the average of estimated β_{20} from 500 replications in each condition, and β_{20} is the true parameter value

$$RB = \frac{\hat{\beta}_{20} - \beta_{20}}{\beta_{20}} \tag{10}$$

Bias: When the cross-lagged fixed effect does not exist (i.e., β_{20}= 0), its accuracy is obtained by *bias* as shown in Eq. 11. Smaller *RB* or *bias* indicates more accurate parameter estimate of β_{20}.

$$bias = \hat{\beta}_{20} - \beta_{20}. \tag{11}$$

Root Mean Square Error: RMSE for each condition is obtained as shown in Eq. 12, where the index r denotes the rth replication out of the 500 replications. The RMSE focuses on the magnitude of error, taking into account the total number of replications, with lower values indicating that estimates are closer to the true values (Hastie, Tibshirani, Friedman, & Friedman, 2009). While there are no universally accepted criteria for what constitutes a "good" or "acceptable" RMSE in multilevel models with ordinal outcomes, it often depends on statistical guidelines and specific clinical contexts. The RMSE is calculated by

$$RMSE = \sqrt{\frac{\sum_{r=1}^{500} \left(\hat{\beta}_{20}^{r} - \beta_{20}\right)^2}{500}}. \tag{12}$$

Although both RMSE and relative bias assess the difference between estimated and true values, we include both metrics to accommodate varying preferences among researchers in social sciences and psychology.

Coverage Rate: The coverage rate of the 95% confidence intervals serves as an indicator of how accurately the standard errors reflect the true variability of the estimated cross-lagged fixed effect. Specifically, the coverage rate is the proportion of simulation replications in which the true parameter value falls within the constructed 95% confidence interval.

$$\text{Coverage Rate} = P(\hat{\beta}_{20} - Z_{\frac{\alpha}{2}} \times SE_{\hat{\beta}_{20}} \leq \beta_{20,\text{true}} \leq \hat{\beta}_{20} + Z_{\frac{\alpha}{2}} \times SE_{\hat{\beta}_{20}}) \tag{13}$$

Acceptable coverage rates typically range between 92.5% and 97.5%, ensuring that the confidence intervals neither underestimate nor overestimate the precision of the parameter estimates (Bradley 1978).

Model Selection: Within each replication, three models are fitted to the simulated dataset:

1. MLM AR(1) with random intercept only,
2. MLM AR(1) with random intercept and random AR(1) effect,
3. MLM AR(1) with random intercept, random AR(1), and random cross-lagged effects.

The model with random intercept and cross-lagged effects, but without random AR(1) effects, was excluded because it is rarely used in empirical studies. The

estimated parameters from the corresponding correct model are used to calculate the evaluation metrics.

For model selection, we evaluate the performance of two commonly used criteria: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (Burnham & Anderson 2004). Both criteria aim to balance model fit with model complexity. To assess their ability to identify the correct models, we calculate the percentage of times each criterion selects the correct model.

In spite of sufficient sample size planning, we also conducted an analysis of variance (ANOVA) with partial eta-squared (η_p^2) to determine how factors such as sample size, AR(1) effects, and cross-lagged effects influence model selection based on AIC and BIC. These criteria help identify the best-fitting model by balancing accuracy and complexity. The η_p^2 metric quantifies how much each factor contributes to model selection while accounting for the effects of other variables, allowing us to isolate the unique impact of each factor. Effect sizes were interpreted using Cohen's guidelines,

where $\eta_p^2 > 0.14$ indicates a large effect, and $\eta_p^2 > 0.06$ is considered medium (Cohen 2013). We also examined how the presence of a cross-lagged effect RB and statistical power, and how its absence affects bias and type I error rates. Additionally, we assessed RMSE in both scenarios, with and without the cross-lagged effect.

5 Results

This section presents the results of our simulations, focusing on how various factors influence sample size planning, power analysis, and the statistical accuracy of parameter estimates in multilevel autoregressive (MLM AR) models with ordinal outcomes and especially with cross-lagged effect. The findings provide insights into the dynamics of model selection and the implications of these factors for effectively detecting cross-lagged effects.

Table 2 ANOVA table on power

	Df	Sum Sq	Mean Sq	*F* value	*p* value	η_p^2
N	2	628.608	314.304	11,260.155	< 0.001	0.129
T	2	525.511	262.755	9413.384	< 0.001	0.110
MargDist	1	355.745	355.745	12,744.808	< 0.001	0.078
β_{10}	2	1566.647	783.324	28,063.093	< 0.001	0.270
β_{20}	2	3759.017	1879.509	67,334.658	< 0.001	0.470
CLViol	2	0.316	0.158	5.654	0.004	< 0.001
$\sigma\mu_{0i}$	3	1294.047	431.349	15,453.367	< 0.001	0.234
$\sigma\mu_{1i}$	3	675.983	225.328	8072.512	< 0.001	0.007
$\sigma\mu_{2i}$	3	4975.754	1658.585	59,419.909	< 0.001	0.540
ARViol	2	8.132	4.066	145.661	< 0.001	0.002
corr	1	15.384	15.384	551.135	< 0.001	0.004
Residuals	151,608	4231.826	0.028			

Table 3 Average power of the normality violation of cross-lagged effect (CLViol; 1 for normal, 2 for moderately right-skewed, 3 for highly right-skewed) and heterogeneity of the cross-lagged effect estimates $\sigma\mu_{2i}$ as a function of the number of participants (*N*), number of assessments

N	*T*	β20	*CLV* iol			σμ2i			
			1	2	3	0	0.1	0.3	0.5
30	28	0.1	0.262	0.259	0.252	0.491	0.224	0.143	0.095
		0.2	0.58	0.576	0.57	0.941	0.549	0.415	0.271
		0.3	0.73	0.729	0.728	**0.995**	0.706	0.629	0.496
30	42	0.1	0.346	0.34	0.333	0.66	0.309	0.174	0.107
		0.2	0.64	0.637	0.635	**0.983**	0.647	0.491	0.311
		0.3	0.776	0.775	0.779	**0.999**	0.774	0.7	0.558
30	100	0.1	0.49	0.483	0.476	**0.936**	0.511	0.226	0.127
		0.2	0.718	0.718	0.719	**0.999**	0.791	0.619	0.381
		0.3	**0.842**	**0.843**	**0.849**	**1**	**0.868**	**0.804**	0.661
50	28	0.1	0.379	0.372	0.365	0.693	0.345	0.21	0.134
		0.2	0.678	0.671	0.673	**0.988**	0.658	0.548	0.396
		0.3	**0.801**	**0.801**	0.799	**1**	0.781	0.723	0.632
50	42	0.1	0.466	0.455	0.452	**0.849**	0.45	0.254	0.149
		0.2	0.729	0.726	0.726	**0.997**	0.738	0.631	0.452
		0.3	**0.841**	**0.84**	**0.841**	**1**	**0.834**	0.778	0.697
50	100	0.2	0.798	0.799	0.566	**0.986**	0.656	0.341	0.179
		0.1	0.574	0.569	**0.806**	**1**	**0.847**	0.749	0.551
		0.3	**0.9**	**0.898**	**0.901**	**1**	**0.914**	**0.861**	0.795
80	28	0.1	0.489	0.482	0.477	**0.859**	0.462	0.302	0.184
		0.2	0.751	0.748	0.747	**0.998**	0.732	0.652	0.529
		0.3	**0.853**	**0.851**	**0.85**	**1**	**0.846**	0.785	0.724

Table 3 (continued)

N	*T*	$\beta 20$	*CLV* iol			σμ2i			
			1	2	3	0	0.1	0.3	0.5
80	42	0.1	0.565	0.56	0.554	**0.952**	0.578	0.367	0.214
		0.2	0.798	0.795	0.797	**1**	**0.801**	0.723	0.597
		0.3	**0.886**	**0.884**	**0.883**	**1**	**0.888**	**0.831**	0.779
80	100	0.1	0.65	0.646	0.645	**0.997**	0.75	0.481	0.263
		0.2	**0.861**	**0.863**	**0.87**	**1**	**0.886**	**0.825**	0.709
		0.3	**0.933**	**0.931**	**0.932**	**1**	**0.948**	**0.897**	**0.864**

Table 4 Average power of the normality violation of AR(1) effect (ARViol; 1 for normal, 2 for moderately right-skewed, 3 for highly right-skewed), heterogeneity of the AR(1) effect estimates $\sigma_{\mu_{1i}}$, and AR(1) fixed effect (β_{10}) as a function of the number of participants (N), number of assessments (T), and cross-lagged effect aggregated across other conditions

N	T	β_{20}	ARViol			$\sigma\mu_{1i}$				β_{10}		
			1	**2**	**3**	**0**	**0.1**	**0.3**	**0.5**	**0.2**	**0.5**	**0.8**
30	28	0.1	0.254	0.258	0.261	0.414	0.26	0.248	0.228	0.313	0.26	0.2
		0.2	0.568	0.573	0.585	**0.872**	0.569	0.555	0.53	0.698	0.588	0.438
		0.3	0.722	0.727	0.739	**0.981**	0.721	0.712	0.693	**0.864**	0.753	0.568
30	42	0.1	0.335	0.34	0.344	0.561	0.341	0.325	0.3	0.41	0.345	0.263
		0.2	0.63	0.636	0.645	**0.946**	0.629	0.616	0.592	0.763	0.657	0.49
		0.3	0.769	0.774	0.786	**0.996**	0.769	0.764	0.744	**0.903**	**0.806**	0.618
30	100	0.1	0.481	0.483	0.485	**0.853**	0.475	0.455	0.434	0.565	0.497	0.385
		0.2	0.711	0.716	0.728	**0.997**	0.715	0.701	0.675	**0.827**	0.748	0.577
		0.3	**0.839**	**0.843**	**0.853**	**1**	**0.846**	**0.836**	**0.817**	**0.942**	**0.881**	0.709
50	28	0.1	0.366	0.371	0.379	0.595	0.375	0.357	0.331	0.458	0.378	0.28
		0.2	0.663	0.672	0.687	**0.961**	0.666	0.654	0.632	**0.82**	0.693	0.508
		0.3	0.792	0.798	**0.811**	**0.998**	0.789	0.788	0.777	**0.935**	**0.829**	0.637
50	42	0.1	0.453	0.455	0.465	0.75	0.456	0.439	0.409	0.556	0.467	0.348
		0.2	0.718	0.724	0.739	**0.989**	0.722	0.71	0.687	**0.866**	0.756	0.557
		0.3	**0.833**	**0.838**	**0.851**	**1**	**0.833**	**0.83**	**0.82**	**0.961**	**0.875**	0.684
50	100	0.1	0.564	0.568	0.578	**0.952**	0.56	0.544	0.518	0.668	0.591	0.449
		0.2	0.793	0.799	**0.811**	**1**	0.798	0.792	0.767	**0.913**	**0.84**	0.649
		0.3	**0.894**	**0.898**	**0.907**	**1**	**0.899**	**0.895**	**0.882**	**0.983**	**0.937**	0.777
80	28	0.1	0.475	0.478	0.494	0.76	0.481	0.464	0.436	0.597	0.49	0.36
		0.2	0.739	0.745	0.762	**0.993**	0.739	0.731	0.716	**0.899**	0.775	0.572
		0.3	**0.845**	**0.849**	**0.86**	1	**0.838**	**0.839**	**0.84**	**0.97**	**0.88**	0.703
80	42	0.1	0.552	0.555	0.571	**0.884**	0.554	0.538	0.508	0.683	0.575	0.419
		0.2	0.787	0.793	**0.811**	**0.999**	0.791	0.785	0.766	**0.936**	**0.83**	0.623
		0.3	**0.877**	**0.882**	**0.894**	**1**	**0.875**	**0.877**	**0.873**	**0.986**	**0.917**	0.748
80	100	0.1	0.639	0.644	0.658	**0.989**	0.639	0.626	0.597	0.754	0.679	0.507
		0.2	**0.856**	**0.861**	**0.876**	**1**	**0.864**	**0.86**	**0.839**	**0.966**	**0.911**	0.717
		0.3	**0.927**	**0.93**	**0.939**	**1**	**0.928**	**0.929**	**0.924**	**0.996**	**0.968**	**0.833**

Power values greater than 0.8 are indicated in bold

5.1 When Cross-Lagged Effect Exists ($\beta_{20} \neq 0$)

Power analysis and sufficient sample size: Table 2 summarizes how various factors affect the power when a cross-lagged effect is present. The heterogeneity of the cross-lagged effect ($\sigma_{\mu_{2i}}$) and the cross-lagged fixed effect (β_{20}) had the most significant impact on power, with effect sizes of $\eta_p^2 = 0.540$ and $\eta_p^2 = 0.470$, respectively. Among the two sample size variables, the number of participants (N; $\eta_p^2 = 0.129$) had

Table 5 Average power of marginal distribution of the outcome (MargDist; 1 for normal, 2 for right-skewed), heterogeneity of intercept ($\sigma_{\mu_{0i}}$), and correlations among random effects (corr; 1 for random effects are correlated, 2 for uncorrelated) as a function of the number of participants (N), number of assessments (T), and cross-lagged effect (β_{20}), aggregated across other conditions

N	T	β_{20}	*MargDist*		$\sigma\mu_{0i}$				*corr*	
			1	2	1	2	3	4	1	2
30	28	0.1	0.238	0.277	0.348	0.292	0.224	0.168	0.262	0.254
		0.2	0.525	0.625	0.713	0.646	0.532	0.413	0.589	0.562
		0.3	0.667	0.79	**0.864**	**0.814**	0.696	0.545	0.745	0.713
30	42	0.1	0.315	0.364	0.441	0.379	0.304	0.236	0.343	0.336
		0.2	0.584	0.689	0.755	0.704	0.609	0.483	0.65	0.625
		0.3	0.717	**0.834**	**0.89**	**0.852**	0.757	0.608	0.791	0.762
30	100	0.1	0.442	0.521	0.556	0.523	0.466	0.39	0.487	0.479
		0.2	0.659	0.774	**0.803**	0.776	0.706	0.592	0.73	0.707
		0.3	0.788	**0.899**	**0.925**	**0.902**	**0.838**	0.718	**0.856**	**0.834**
50	28	0.1	0.341	0.403	0.493	0.421	0.33	0.246	0.377	0.367
		0.2	0.615	0.731	**0.811**	0.761	0.639	0.485	0.688	0.66
		0.3	0.738	**0.862**	**0.923**	**0.889**	0.786	0.606	**0.815**	0.786
50	42	0.1	0.422	0.493	0.567	0.509	0.426	0.33	0.464	0.451
		0.2	0.666	0.787	**0.845**	**0.804**	0.708	0.553	0.741	0.713
		0.3	0.782	**0.898**	**0.94**	**0.916**	**0.836**	0.672	**0.854**	**0.827**
50	100	0.1	0.518	0.619	0.639	0.613	0.56	0.469	0.576	0.563
		0.2	0.738	**0.861**	**0.882**	**0.859**	0.795	0.669	**0.812**	0.79
		0.3	**0.849**	**0.948**	**0.965**	**0.948**	**0.903**	0.783	**0.91**	**0.889**
80	28	0.1	0.44	0.524	0.612	0.548	0.444	0.326	0.489	0.476
		0.2	0.685	**0.812**	**0.881**	**0.845**	0.727	0.543	0.762	0.736
		0.3	0.795	**0.907**	**0.953**	**0.932**	**0.852**	0.669	**0.861**	**0.842**
80	42	0.1	0.512	0.606	0.667	0.623	0.537	0.412	0.568	0.551
		0.2	0.735	**0.858**	**0.906**	**0.877**	0.792	0.613	**0.808**	0.786
		0.3	**0.833**	**0.935**	**0.967**	**0.95**	**0.891**	0.729	**0.893**	**0.875**
80	100	0.1	0.588	0.703	0.717	0.702	0.64	0.531	0.652	0.642
		0.2	**0.806**	**0.921**	**0.937**	**0.92**	**0.868**	0.733	**0.873**	**0.856**
		0.3	**0.892**	**0.971**	**0.985**	**0.974**	**0.94**	**0.831**	**0.94**	**0.925**

Power values greater than 0.8 are indicated in bold

a stronger influence on power than the number of assessments (T; $\eta_p^2 = 0.110$), both being moderate but significant contributors to power. Tables 3, 4, and 5 display the power for detecting cross-lagged effects under different groups of factors, illustrating how sufficient sample sizes are determined across various conditions.

Table 3 presents the power to detect the cross-lagged effect of its related factors as a function of sample size (N), the number of assessments (T), and the expected magnitude of the cross-lagged effect (β_{20}). The results indicate that normality violations of the cross-lagged effect (CLViol) do not significantly influence the detection

Table 6 ANOVA results on RMSEs

	Df	Sum Sq	Mean Sq	*F* value	*p* value	η_p^2
N	2	39.565	19.782	25,818.308	< 0.001	0.236
T	2	23.297	11.649	15,202.750	< 0.001	0.154
MargDist	1	12.099	12.099	15,791.255	< 0.001	0.086
β_{10}	2	48.442	24.221	31,611.588	< 0.001	0.275
β_{20}	3	31.372	10.457	13,648.035	< 0.001	0.004
CLViol	2	0.002	0.001	1.525	0.218	< 0.001
$\sigma\mu_{0i}$	3	54.200	18.067	23,579.419	< 0.001	0.297
$\sigma\mu_{1i}$	3	15.254	5.085	6636.174	< 0.001	0.006
$\sigma\mu_{2i}$	3	166.699	55.566	72,520.700	< 0.001	0.566
ARViol	2	0.358	0.179	233.447	< 0.001	0.003
corr	1	0.394	0.394	514.718	< 0.001	0.003
Residuals	167,131	128.058	0.001			

of the cross-lagged effect, regardless of the distribution shape (whether symmetrical, moderately skewed, or highly skewed). As expected in MLM AR, increasing the sample size—whether by the number of participants or assessments—leads to a corresponding increase in power. Larger cross-lagged effects are more likely to be detected. However, when the expected cross-lagged effect (β_{20}) is small (e.g., 0.1), even with a large sample size ($N = 80$) and many assessments ($T = 100$), the power does not reach the 0.8 threshold. Conversely, when the expected cross-lagged effect is larger ($\beta_{20} = 0.3$), the power approaches 0.8 with a relatively small sample size ($N = 30$) and moderate assessments ($T = 42$).

Table 7 ANOVA results on RBs

	Df	Sum Sq	Mean Sq	*F* value	*p* value	η_p^2
N	2	0.043	0.022	1.110	0.33	< 0.001
T	2	25.512	12.756	654.669	< 0.001	0.009
MargDist	1	37.653	37.653	1932.456	< 0.001	0.013
β_{10}	2	303.023	151.512	7776.019	< 0.001	0.093
β_{20}	2	35.481	17.741	910.502	< 0.001	0.012
CLViol	2	2.068	1.034	53.060	< 0.001	0.001
$\sigma\mu_{0i}$	3	606.919	202.306	10,382.957	< 0.001	0.170
$\sigma\mu_{1i}$	3	81.512	27.171	1394.477	< 0.001	0.005
$\sigma\mu_{2i}$	3	407.137	135.712	6965.149	< 0.001	0.121
ARViol	2	8.622	4.311	221.263	< 0.001	0.003
corr	1	72.927	72.927	3742.836	< 0.001	0.024
Residuals	151,608	2954.001	0.019			

Table 8 95% CI coverage rates for the estimates of the cross-lagged effect (β_{20}) by the number of participants (N) and number of assessments (T) (outcome distribution: bell-shaped)

Bell-shaped	3*N*			N *T*		
β_{20}	30	50	80	28	42	100
0.1	0.936	0.933	0.931	0.936	0.936	0.934
0.2	0.935	0.930	0.927	0.935	0.933	0.931
0.3	0.932	0.927	0.924	0.931	0.930	0.930

Table 9 95% CI coverage rates for the estimates of the cross-lagged effect (β_{20}) by the number of participants (N) and number of assessments (T) (outcome distribution: skewed)

3Skewed	*N*			*T*		
β_{20} β_{20}	330	350	880	328	342	3100
0.1	0.94	0.939	0.939	0.94	0.94	0.939
0.2	0.94	0.939	0.939	0.939	0.94	0.939
0.3	0.94	0.94	0.941	0.939	0.94	0.941

The heterogeneity of the cross-lagged effect shows a different trend. As the heterogeneity increases, the power to detect the cross-lagged effect decreases. When there is no heterogeneity ($\sigma_{\mu_{2i}} = 0$), the power remains high, even with smaller sample sizes ($N=30$, $T=28$, $\beta_{20} = 0.2$), and increases to nearly 1 with larger sample sizes. However, when heterogeneity is large ($\sigma_{\mu_{2i}} = 0.5$), a much larger sample size ($N=80$) and a greater number of assessments ($T=100$) are required to achieve a power above 0.8.

Table 4 presents the power results of cross-lagged effect with AR(1)-related factors. The impact of normality violations in the AR(1) effect is minimal, with skewed distributions slightly improving power, though the overall effect is negligible. This aligns with the small effect sizes shown in Table 2, indicating subtle influences of distribution shape on power. Regarding the heterogeneity of the AR(1) effect ($\sigma_{\mu_{1i}}$), the pattern is similar to that observed with $\sigma_{\mu_{2i}}$ (heterogeneity of the cross-lagged effect).

When there is less variation in AR(1) effects across individuals (i.e., more consistency), the power to detect the effect is higher.

For the AR(1) fixed effect (β_{10}), the observed pattern indicates that power increases as both sample size and the magnitude of the expected cross-lagged effect grow. When the expected cross-lagged effect is 0.3, and the AR(1) fixed effect is small (about 0.2), it is possible to achieve power over 80% power with a small sample size ($N=30$) and a limited number of assessments ($T=28$), illustrating the efficiency of data collection under such conditions. However, as the AR(1) fixed effect increases to medium (0.5) or large (0.8) values, the required number of participants and the number of assessments rise significantly. For example, with an AR(1) fixed effect of 0.5 and $N=30$ participants, the power to detect a small cross-lagged effect

Table 10 Bias, RMSE, and type I error for each condition when there is no cross-lagged effect

	N			T			MargDist		β_{10}			β_{20}		CLViol	
level	30	50	80	28	42	100	1	2	0.2	0.5	0.8	0	1	2	3
bias	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
RMSE	0.042	0.032	0.025	0.043	0.034	0.022	0.034	0.033	0.031	0.032	0.035	0.033	0.033	0.033	0.033
Type I Error Rate	0.052	0.052	0.053	0.054	0.052	0.051	0.053	0.052	0.052	0.053	0.053	0.053	0.052	0.053	0.052
AIC_correct_prop	0.676	0.722	0.756	0.646	0.705	0.802	0.710	0.726	0.741	0.717	0.696	0.718	0.719	0.717	0.718
BIC_correct_prop	0.569	0.640	0.688	0.549	0.618	0.730	0.629	0.636	0.645	0.631	0.621	0.632	0.632	0.633	0.632

	$\sigma_{\mu_{0i}}$				$\sigma_{\mu_{1i}}$				$\sigma_{\mu_{2i}}$		ARViol			corr	
level	1	2	3	4	0	0.1	0.3	0.5	0	1	2	3	1	2
bias	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
RMSE	0.029	0.031	0.034	0.038	0.039	0.030	0.031	0.032	0.033	0.033	0.033	0.033	0.033	0.033
Type I Error Rate	0.051	0.053	0.053	0.054	0.058	0.051	0.050	0.051	0.053	0.052	0.053	0.053	0.053	0.052
AIC_correct_prop	0.813	0.761	0.696	0.601	0.808	0.321	0.836	0.908	0.718	0.723	0.721	0.710	0.713	0.723
BIC_correct_prop	0.739	0.663	0.601	0.527	0.930	0.036	0.663	0.901	0.632	0.641	0.639	0.617	0.605	0.660

Table 11 ANOVA table on type I error

	Df	Sum Sq	Mean Sq	*F* value	*p* value	η_p^2
N	2	0.002	0.001	1.017	0.362	< 0.001
T	2	0.020	0.010	9.303	< 0.001	0.001
MargDist	1	< 0.001	< 0.001	0.018	0.893	< 0.001
β_{10}	2	0.001	< 0.001	0.324	0.723	< 0.001
CLViol	2	0.007	0.004	3.395	0.034	< 0.001
$\sigma\mu_{0i}$	3	0.016	0.005	5.046	0.002	0.001
$\sigma\mu_{1i}$	3	0.167	0.056	52.272	< 0.001	0.010
ARViol	2	0.001	0.001	0.542	0.582	< 0.001
corr	1	0.001	0.001	0.594	0.441	< 0.001
Residuals	15,505	16.532	0.001			

Table 12 ANOVA results on bias

	Df	Sum Sq	Mean Sq	*F* value	*p* value	η_p^2
N	2	< 0.001	0	3.423	0.033	< 0.001
T	2	< 0.001	0	6.114	0.002	0.001
MargDist	1	< 0.001	0	0.514	0.473	< 0.001
β_{10}	2	< 0.001	0	0.335	0.716	< 0.001
CLViol	2	< 0.001	0	0.386	0.680	< 0.001
$\sigma\mu_{0i}$	3	< 0.001	0	1.753	0.154	< 0.001
$\sigma\mu_{1i}$	3	< 0.001	0	0.200	0.896	< 0.001
ARViol	2	< 0.001	0	0.117	0.890	< 0.001
corr	1	< 0.001	0	2.733	0.098	< 0.001
Residuals	15,505	0.431	0			

(0.1) remains below 50%, even with $T = 100$. Increasing the sample size to $N = 80$ participants pushes the power closer to 80%. In cases where the AR(1) fixed effect is high (0.8), substantial increases in sample sizes may still to produce adequate power. Specifically, with $N = 80$ participants and $T = 100$ assessments, power for detecting a cross-lagged effect of 0.1 only reaches 50%, underscoring the difficulties of achieving sufficient statistical power under these conditions.

Table 5 presents the power results as a function of the marginal distribution of the outcome, heterogeneity of the intercept ($\sigma_{\mu_{20}}$), and the correlations among random effects. The marginal distribution of the outcome is simulated with two conditions: 1 for normal distribution and 2 for right-skewed. Similar to the effects of normality violations in the AR(1) model, the right-skewed condition generally results in higher power compared to the normal distribution under various sample sizes and

Table 13 Relative bias, RMSE, and power for each condition when there is a nonzero cross-lagged effect

	N			T			MargDist		β_{10}			β_{20}			CLViol		
level	30	50	80	28	42	100	1	2	0.2	0.5	0.8	0.1	0.2	0.3	1	2	3
RB	-0.068	-0.069	-0.069	-0.083	-0.071	-0.051	-0.084	-0.053	-0.018	-0.061	-0.127	-0.089	-0.064	-0.053	-0.064	-0.068	-0.073
RMSE	0.101	0.077	0.061	0.094	0.082	0.064	0.089	0.071	0.060	0.075	0.105	0.078	0.080	0.082	0.080	0.080	0.080
AIC_correct_prop	0.648	0.714	0.761	0.627	0.690	0.806	0.691	0.724	0.742	0.715	0.666	0.707	0.708	0.708	0.708	0.709	0.706
BIC_correct_prop	0.440	0.533	0.603	0.410	0.511	0.656	0.505	0.546	0.569	0.535	0.473	0.527	0.526	0.524	0.527	0.527	0.523

	$\sigma_{\mu_{0i}}$				$\sigma_{\mu_{1i}}$				$\sigma_{\mu_{2i}}$				ARViol			corr	
level	1	2	3	4	0	0.1	0.3	0.5	0	0.1	0.3	0.5	1	2	3	1	2
RB	-0.016	-0.021	-0.063	-0.173	0.004	-0.062	-0.075	-0.087	0.003	-0.053	-0.111	-0.137	-0.077	-0.070	-0.059	-0.090	-0.047
RMSE	0.060	0.067	0.081	0.111	0.039	0.083	0.083	0.085	0.034	0.077	0.101	0.123	0.082	0.080	0.078	0.078	0.082
AIC_correct_prop	0.767	0.735	0.690	0.637	0.817	0.574	0.747	0.774	0.723	0.250	0.869	0.983	0.710	0.710	0.703	0.712	0.704
BIC_correct_prop	0.603	0.554	0.500	0.445	0.937	0.327	0.536	0.611	0.637	0.010	0.534	0.885	0.530	0.529	0.519	0.518	0.533

cross-lagged effect conditions. For instance, with a small sample size of $N=30$ and $T=42$, the power reaches 0.834. Regarding the heterogeneity of the intercept ($\sigma_{\mu_{2i}}$), this factor has a substantial impact on power, with an effect size (η_p^2) of 0.234. Specifically, when the variability is smaller, the power of the model increases. The correlation among random effects has a minimal impact on power. Additionally, the difference between the correlated and uncorrelated conditions is relatively small.

Statistical accuracy of estimates: As shown in Table 6, variability in the cross-lagged random effect ($\sigma_{\mu_{2i}}$) has the strongest influence on RMSE (η_p^2 of 0.566). This finding suggests that greater heterogeneity in the cross-lagged effect makes parameter estimation more challenging. Variability in the random intercept ($\sigma_{\mu_{0i}}$) and the magnitude of the AR(1) fixed effect (β_{20}) also substantially affect RMSE, indicating that both initial variability and the strength of autoregressive influence shape estimation quality. The number of participants and the number of assessments also have relative large impact on RMSEs.

For relative bias (RB) analyses (Table 7), variability in the intercept ($\sigma_{\mu_{0i}}$) has a relatively large impact ($\eta_p^2= 0.170$). In addition, the variability in the random cross-lagged effect ($\sigma_{\mu_{2i}}$) and AR(1) fixed effect (β_{10}) moderately impact the RBs with partial eta-squared of 0.121 and 0.093, respectively. Except for these factors, most factors still produce acceptable bias levels. In contrast, the number of participants does not significantly affect RB ($p>0.05$).

Tables 8 and 9 present the coverage rates for the fixed cross-lagged effect (β_{20}) under bell-shaped and skewed outcome distributions, respectively. In both conditions, most fixed effect estimates achieve acceptable coverage rates (92.5% to 97.5%). Notably, although the estimation method assumes normality, the coverage rates are more stable under the skewed condition—a somewhat counterintuitive but noteworthy finding. This suggests that, despite the model's underlying normality assumption, certain data characteristics in the skewed distribution may mitigate coverage issues, resulting in more robust estimates.

5.2 When Cross-Lagged Effect is Absent ($\beta_{20}=0$ and $\sigma_{\mu_{2i}}=0$)

While our primary focus is on scenarios where a cross-lagged effect exists, it is also important to assess the model's behavior when no such effect is present. Examining this scenario provides insights into the model's baseline performance, informing decisions about study design and sample size planning even if the hypothesized cross-lagged relationship does not exist.

Type I error rate and Bias: Without a cross-lagged fixed effect, both type I error rates and bias remain consistently low across most conditions (see Table 10).

Although the number of assessments (T), variability in the random intercept ($\sigma_{\mu_{0i}}$), and AR(1) random effect ($\sigma_{\mu_{1i}}$) show significant differences in type I error rate (Table 11), their effect sizes are small ($\eta_p^2<0.01$). Similarly, while the number of assessments (T) significantly affects *bias*, the effect size is minimal

(Table 12). Other factors exhibit negligible impacts on bias and Type I error rates, suggesting the model's robustness in the absence of a cross-lagged effect.

5.3 Model Selection

Having examined scenarios both with and without a cross-lagged effect, we now turn to the performance of model selection criteria. Beyond measures of bias, error rates, and power, the ability to identify the correct model is crucial for guiding effective study design and analysis strategies.

As shown in Tables 13 and 10, the AIC generally demonstrates greater accuracy than the BIC in selecting the correct model across various conditions with or without cross-lagged effect. Notably, when dealing with right-skewed marginal distributions ($MargDist = 2$), AIC correctly identifies the true model in above 70% of cases, compared to around 60% for BIC. This finding suggests that, particularly under skewed distributional conditions, AIC may be the more reliable choice for guiding model selection in MLM AR studies with or without cross-lagged effects.

6 Conclusion and Discussion

This study provides practical guidelines for sample size planning in multilevel autoregressive (MLM AR) models with cross-lagged effects and ordinal outcomes. Such models are increasingly common in fields including psychology, public health, and education. By examining how sample sizes (both participants and assessments), effect sizes (autoregressive and cross-lagged), and random effect distributions influence statistical power and parameter estimation, we aim to help researchers make informed methodological decisions.

Our findings highlight the challenge that arises when a large autoregressive (AR(1)) effect coincides with a small cross-lagged effect. In such cases, the AR(1) dynamics dominate the outcome, leaving insufficient variance for the cross-lagged predictor. Even substantial increases in the number of participants (N) and assessments (T) may not yield sufficient power. These results emphasize the importance of realistic expectations, based on existing literature, pilot data, or expert knowledge, before investing substantial resources. For instance, a moderate cross-lagged effect and a relatively small AR(1) effect may be adequately detected with about $N = 30$ and $T = 42$ assessments. In contrast, weaker cross-lagged effects combined with strong AR(1) processes may require much larger samples (e.g., $N = 80$, $T = 100$) without a strong guarantee of sufficient power. Researchers should carefully consider theoretical considerations, plausible effect sizes, and practical constraints when designing longitudinal studies aimed at detecting cross-lagged relationships.

In addition to these considerations, our results show that heterogeneity in the random cross-lagged parameter and strong autoregressive influences reduce power. However, when there is no cross-lagged effect, well-specified models maintain low Type I error rates and bias. Increasing the number of assessments, in particular,

reduces the relative bias in estimating the cross-lagged parameter, suggesting that more frequent measurements can improve the estimation accuracy, even if simply increasing the number of participants does not.

Contrary to concerns that non-normal random effect distributions would substantially degrade power or parameter accuracy, we found only modest effects, especially when compared to the influence of parameters' magnitude and variability. However, large variability in intercepts or cross-lagged effects reduces power, emphasizing the importance of making sound assumptions about underlying parameter distributions. Furthermore, model selection criteria play a crucial role in these contexts. Our findings suggest that the Akaike information criterion (AIC) may outperform the Bayesian information criterion (BIC) when outcome distributions are skewed, offering guidance for researchers dealing with complex longitudinal data structures.

The **OrdPower** (Shao et al. 2023) R package, developed as part of this work, provides a user-friendly tool for researchers performing sample size planning for longitudinal ordinal data, particularly when the cross-lagged effect is of interest. By simplifying these calculations and improving the accuracy of power estimates, **OrdPower** can help design more informative studies.

The study has several limitations. First, we did not account for attrition in our calculations. In many ecological momentary assessment (EMA) studies, participant burden increases with the number of assessments, potentially leading to missing data.

The simple multiplication methods outlined in the R package are insufficient because different missing data mechanisms may require distinct approaches. Second, our simulations considered relatively simple models, only involving a single lagged version of the outcome and a single covariate. In practice, modern EMA studies often include multiple variables and more complex designs, which may affect generalizability. Finally, our approach follows a nomothetic framework, focusing on group-level dynamics. While idiographic approaches, such as time-series models (e.g., VAR), can capture individual-level processes more flexibly, they introduce inferential challenges due to unknown parameter distributions. Furthermore, although recent evidence supports the use of single-item measures in EMA studies (Song et al. 2023), multiple-item measures generally yield more reliable assessments of underlying constructs.

Future research could address these limitations by incorporating attrition models into sample size planning and considering varying time lags across individuals.

These extensions would better reflect real-world conditions and improve our understanding of the factors influencing power in longitudinal MLM AR with ordinal outcomes. By broadening the scope to more diverse modeling scenarios, researchers can develop study designs that more accurately reflect the complexities of their data and research questions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40647-025-00447-2.

Author contribution Sijing (S.J.) Shao served as lead for conceptualization, formal analysis, methodology, validation, writing, review and editing. Ziqian Xu contributed equally to formal analysis, software, and visualization. Wen Qu contributed equally to formal analysis, validation, visualization, writing, review and editing. Ross Jacobucci contributed supervision.

Funding This research received no specific grant from any funding agency.

Code availability The R code can be installed via GitHub by install_github (shaosijing / OrdPower).

Data availability All datasets used in this study were produced entirely through Monte-Carlo simulation; no empirical or proprietary data were involved.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ali, S., A. Ali, S.A. Khan, and S. Hussain. 2016. Sufficient sample size and power in multilevel ordinal logistic regression models. *Computational and Mathematical Methods in Medicine* 2016 (1): 7329158.

Ammerman, B.A., and K.C. Law. 2022. Using intensive time sampling methods to capture daily suicidal ideation: A systematic review. *Journal of Ajective Disorders* 299:108–117.

Bauer, D.J., and S.K. Sterba. 2011. Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods* 16 (4): 373.

Bolger, N., and J.-P. Laurenceau. 2013. *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford press.

Bradley, J.V. 1978. Robustness? *British Journal of Mathematical and Statistical Psychology* 31 (2): 144–152.

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

Burnham, K.P., and D.R. Anderson. 2004. Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods & Research* 33 (2): 261–304.

Christensen, R. H. B., & Christensen, M. R. H. B. (2015). Package 'ordinal'. *Stand*, *19* (2016).

Cohen, J. 1992. Statistical power analysis. *Current Directions in Psychological Science* 1 (3): 98–101.

Cohen, J. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

Fitzmaurice, G.M., N.M. Laird, and J.H. Ware. 2012. *Applied longitudinal analysis*. Wiley.

Grimm, K.J., N. Ram, and R. Estabrook. 2016. *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications.

Grønneberg, S., N. Foldnes, and K.M. Marcoulides. 2022. An R package for simulating non-normal data for structural equation models using copulas. *Journal of Statistical Software* 102 (3): 1–45. https://doi.org/10.18637/jss.v102.i03.

Hamaker, E.L., and R.P. Grasman. 2015. To center or not to center? investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology* 5:1492.

Hamilton, J.D. 2020. *Time series analysis*. Princeton University Press.

Hastie, T., R. Tibshirani, J.H. Friedman, and J.H. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Hayashi, K., Yuan, K.-H., & Bentler, P. M. (2024). On the relationship between factor loadings and component loadings when latent traits and specificities are treated as latent factors. *Fudan Journal of the Humanities and Social Sciences*, 1–15.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Jacobucci, R., B.A. Ammerman, and X. Li. 2021. Using ordinal regression for advancing the understanding of distinct suicide outcomes. *Suicide and Life-Threatening Behavior* 51 (1): 65–75. https://doi.org/10.1111/sltb.12669.

Jacobucci, R., K. McClure, and B.A. Ammerman. 2023. Comparing the role of perceived burdensomeness and thwarted belongingness in prospectively predicting active suicidal ideation. *Suicide and Life-Threatening Behavior* 53 (2): 198–206.

Kumle, L., M.L.-H. Võ, and D. Draschkow. 2021. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods* 53 (6): 2528–2543.

Li, M., & Hu, Y. (2024). A hybrid method: Resolving the impact of variable ordering in bayesian network structure learning. *Fudan Journal of the Humanities and Social Sciences*, 1–17.

Liao, X., Song, H., & Bard, D. E. (2024). Predicting successful treatment completion using baseline case characteristics through machine learning and ensemble modeling: A two-step approach. *Chinese Political Science Review*, 1–30.

Liu, I., and A. Agresti. 2005. The analysis of ordered categorical data: An overview and a survey of recent developments. *TEST* 14:1–73.

Lu, Z. 2025. Choosing among pca, fa, lca, lpa, and lda. *Fudan Journal of the Humanities and Social Sciences* 18 (1): 45–78.

McKelvey, R.D., and W. Zavoina. 1975. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology* 4 (1): 103–120.

Moineddin, R., F.I. Matheson, and R.H. Glazier. 2007. A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology* 7 (1): 1–10.

R Core Team. (2021). *R: A language and environment for statistical computing* (manual). Vienna, Austria. Retrieved from https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage

Shao, S., Xu, Z., & Jacobucci, R. (2023). *R package for sample size planning and power calculation of multilevel models with ordinal variables.* Retrieved from https://github.com/shaosijing/OrdPower

Shiffman, S., A.A. Stone, and M.R. Hufford. 2008. Ecological momentary assessment. *Annual Review of Clinical Psychology* 4 (1): 1–32.

Singer, J.D., J.B. Willett, and J.B. Willett. 2003. *& others*. Applied longitudinal data analysis: Modeling change and event occurrence. Oxford University Press.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. sage.

Song, J., E. Howe, J.R. Oltmanns, and A.J. Fisher. 2023. Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment* 30 (5): 1662–1671.

Verbeke, G. 1997. Linear mixed models for longitudinal data. In *Linear mixed models in practice*, 63–153. Springer.

Wang, L., E. Hamaker, and C. Bergeman. 2012. Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods* 17 (4): 567.

Wang, L., M. Yang, and X. Liu. 2019. The impact of over-simplifying the between-subject covariance structure on inferences of fixed effects in modeling nested data. *Structural Equation Modeling: A Multidisciplinary Journal* 26 (1): 1–11.

Winship, C., and R.D. Mare. 1984. Regression models with ordinal variables. *American Sociological Review* 49:512–525.

Wu, J., N. Ram, J. Marks, N.M. Streeper, and D.E. Conroy. 2024. Small data approaches to link faster time scale engagement dynamics with slower time scale outcomes in biobehavioral interventions. *Chinese Political Science Review*. https://doi.org/10.1007/s41111-024-00255-1.