Research Report

Machine Learning Techniques for Baseline Covariate Adjustment in Structural Equation Modeling

Mano (E.D.) van Holten

Student no. 7776632

Program:

Methods and Statistics for the Biomedical Behavioural and Social Sciences (MSBBSS)

Supervisors:

Dr. Jeroen D. Mulder

Prof. Dr. Ellen L. Hamaker

Target Journal:

Advances in Methods and Practices in Psychological Science

Wordcount: 2497

FETC approved: 24-0286

1 Introduction

Researchers in behavioral and psychological research are often interested in how two variables influence one another over time (Usami et al., 2019). Suppose a researcher is interested in the causal effect of a football player's self-confidence (X) on later performance (Y), and vice versa. Figure 1 displays an assumed causal structure between repeated measures of X_t and Y_t. The effect of X_{t-1} on Y_t (and vice versa) is called the cross-lagged parameter. A key threat to such an observational study are confounders: common causes of both self-confidence and performance. For example, older, more experienced players may better understand their own abilities, which increases their confidence and also leads to stronger match performance. Age and other covariates, which do not change over the course of a study, are referred to as baseline confounders (C). Other confounders may vary over time, such as a player sustaining an injury. For this research report, we focus exclusively on baseline confounders, which lead to biased estimates of the cross-lagged parameters when not, or inappropriately, controlled for.

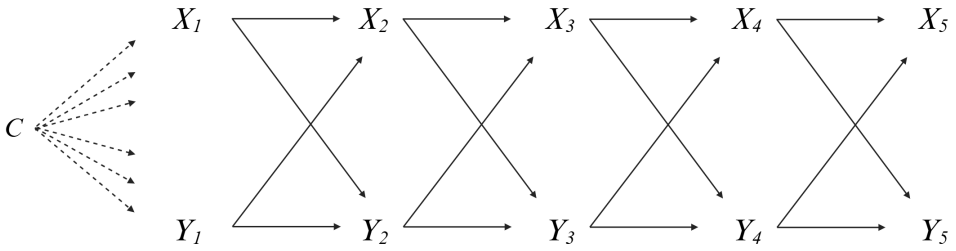

Figure 1: The data-generating mechanism, with a set of common causes (C).

A commonly used model in psychology for analyzing repeated-measures data is the Cross-Lagged Panel Model (CLPM), along its with extensions such as the Random Intercepts (RI) CLPM (Hamaker et al., 2015), and reciprocal versions of the Dynamic Panel Model (DPM) (Dishop and DeShon, 2022). The RI-CLPM separates stable between-person differences (traits) from within-person dynamics. Although it was not designed for confounder control, the RI-CLPM can act as a workaround for baseline confounding. Baseline confounders are not modeled explicitly, but are

absorbed into latent variables that capture stable interpersonal differences (e.g., age). A similar approach is the DPM, which originates in the econometric literature and has been adapted for estimation within the Structural Equation Modeling (SEM) framework (Allison et al., 2017). A key statistical assumption needed to enable control by latent variables is that the effects of baseline confounders do not change over time (Murayama and Gfrörer, 2024). However, baseline confounders, while time-invariant themselves, may have effects that vary over time. In the example study, a rule change by the national governing body of football may encourage high-intensity play. This shift benefits younger, fitter players, reversing the effect of age on performance, changing it from positive before the change, to negative afterward. The first goal of this research report is to show that the RI-CLPM and DPM produce biased estimates of the cross-lagged parameters when baseline covariates exert time-varying effects.

When this assumption is violated, researchers can attempt to measure all baseline confounders and explicitly model the effect of age separately at each measurement occasion. This relaxes the assumption of time-invariant effects and avoids relying solely on latent variables. In our example, we assume that all relevant baseline confounders have been observed, such that age can be included as a predictor of performance in the longitudinal model. However, this introduces another challenge: the researcher must specify how age affects performance. In practice, social scientists often assume this effect is linear. Unfortunately, like many other relationships in the social sciences (e.g., Hayes et al., 2007), age does not affect football performance in a linear fashion. Age interacts with other variables that also affect performance. For instance, its effect may depend on a player's position: physical fitness is likely more critical for a midfielder than for a goalkeeper. These interaction effects are not adequately controlled for by adding age and playing position only as linear covariates (Mize, 2019). The second goal of this research report is to examine potential bias that arises from explicitly modeling baseline confounders when their effects are incorrectly assumed to be linear.

While some longitudinal models allow for relaxation of the linearity assumption (e.g., Robins et al., 2000), they have generally not been implemented within the SEM framework (Bollen and Brand, 2010). A recent approach, baseline covariate adjustment (BCA) SEM (Irmer et al., 2025), decouples

control of time-invariant confounders from the estimation of the cross-lagged effects. BCA scores — residuals from regressing variables of interest on baseline covariates — replace observed variables in the SEM. A major strength of this approach is that the relationship between the variables of interest and the confounders can be flexibly modeled using machine learning (ML) techniques, such that the form of the adjustment function does not need to be assumed but can be learned from the data. This should in principle reduce model misspecification, and lead to more robust causal inference. However, ML methods haven't yet been implemented for BCA SEM, and they also come with their own challenges. ML models use more parameters than linear models to accommodate complex functional forms, so each parameter is estimated from less information. Even if a ML approach is unbiased on average, in practice researchers are more likely to obtain estimates that deviate substantially from the true cross-lagged effect (see Shmueli, 2010). Despite these concerns, machine learning methods have been successfully applied to longitudinal data (Sheetal et al., 2023), demonstrating their potential to reduce model misspecification bias. The third goal of this research report is to provide a proof-of-concept implementation of the BCA SEM approach with ML techniques.

This research report is structured around two simulations. In the first we investigate the performance of the RI-CLPM and DPM when baseline confounders change their influence over time (goal 1). In the second we investigate the impact of interactions effects when the effects of baseline confounders are wrongly assumed linear (goal 2), while also offering a proof-of-concept ML approach (goal 3).

2 Methods

For our simulations, we use the causal structure shown in Figure 1. As before, the reader may imagine X to be self-confidence, and Y to be football performance. We will generate three baseline confounders, which can be though of as age, playing position, or any other time-invariant common cause. We chose to simulate five measurement occasions for X_t and Y_t, as this seems to be enough for all included models to reliably converge (Orth et al., 2021). We opt for a large sample size of 1000 for this proof of concept, and leave issues regarding sample size for the discussion. We first present the parameter values for the causal process of interest, followed by the confounder effects.

2.1 Causal parameter values

The causal parameters are the cross-lagged and autoregressive effects, while the data-generating model additionally includes variances and covariances of X_t and Y_t, as well as residual variances and covariances. The lagged effect of X on Y is set to 0.1. By contrast, performance has no effect on self-confidence, with the corresponding cross-lagged parameter fixed at 0. In the thesis, this design will be used to examine how model misspecification can lead to type 1 errors. For this research report, we will focus only on the lagged effect of X on Y. The autoregressive effects for both variables are set to 0.2. The data are generated such that all variables, including confounders, have a variance of 1, allowing all effects to be interpreted as standardized. This is achieved by first computing the variance induced by the confounders at the current wave and by the predictors from the previous wave, after which sufficient error variance is added to ensure that each variable has variance one. Additionally, some variables affect X and Y at the same measurement occasion only. For example, a football match might be played in nice weather, simultaneously boosting players' confidence (or morale) and performance. After accounting for variance due to the confounders and lagged predictors, this remaining association between X_t and Y_t is introduced through a residual correlation, increasing the observed correlation between X_t and Y_t by 0.1.

2.2 Baseline confounder effects

We set the effects of the confounders such that they account for 15 percent of the variance in the outcomes, a small-to-medium effect. We consider two scenarios for the effects of the baseline confounders. A constant scenario, where these confounders affect all measurement waves equally, and a stepwise scenario. In the stepwise scenario, our hypothetical football rule change is introduced, causing the effects of the baseline confounders to change from the third measurement occasion onward. We emulate this change by multiplying the coefficients by a constant such that the variance explained (R^2) by the confounders from that occasion forward increases to 40 percent.

All effects in simulation study 1 are generated using a linear model. Simulation study 2 extends simulations study 1 by adding interaction terms among the three confounders. Specifically, three two-way interactions and one three-way interaction are constructed and included as additional variables, resulting in seven variables in total affecting X and Y. The interaction terms are standardized so that their variance equals 1 and are made independent of their corresponding linear components. This ensures that the variance explained by the confounders can be decomposed into linear effects and non-linear interaction effects. We fix this proportion at 0.5, meaning that non-linear interaction effects explain 7.5 percent of the outcome variance at the first measurement occasion. This proportion is kept constant across all measurement occasions, including in the stepwise scenario. Each scenario is simulated for 1000 replications.

2.3 Models

For simulation study 1, we consider four modeling approaches. The first is a CLPM without adjustment for confounding effects, representing a worst-case scenario in which the researcher wrongly assumes there are no common causes of the outcomes. The second approach uses the latent variable models discussed in the introduction — the RI-CLPM (Hamaker et al., 2015) and the DPM (Dishop and DeShon, 2022) — to estimate the cross-lagged parameters. The third approach explicitly adjusts for baseline confounders by including them directly as predictors in the CLPM. This

model corresponds exactly to the data-generating process and is therefore referred to as the true model in simulation study 1. The fourth approach is BCA SEM, which first obtains residuals by linearly regressing X and Y on the baseline confounders and then fits a CLPM to these residuals (Irmer et al., 2025). In simulation study 1, this approach is equivalent to the explicit adjustment approach. Simulation study 2 extends this fourth modeling approach, replacing the linear regression model by Extreme Gradient Boosting (XGB) (Chen, 2016). XGB is a machine-learning method that builds many decision trees sequentially, where each new tree focuses on correcting the errors of the previous ones. It was selected as a proof-of-concept method because it performs well with sample sizes of 1,000. The XGB model is tuned via cross-validation (see Hastie et al., 2009) in the first simulation replication, which is used for both tuning and evaluation. The resulting hyper-parameter settings are then reused in all subsequent replications to reduce computational burden. For simulation study 2, no true model is included. Additional machine learning methods will be implemented and compared in the thesis.

2.4 Outcome measures

Model performance is evaluated for the cross-lagged effect of X_{t-1} on Y_t using relative bias and the Root Mean Square Error (RMSE) (see Morris et al., 2019). Relative bias is calculated as:

$$\text{RB}_r = \frac{\theta_r - \theta_{\text{true}}}{\theta_{\text{true}}},$$

where θ_r denotes the estimated parameter from a given method obtained in replication r and $\theta_{\text{true}} = 0.1$ is the true population parameter value. Relative bias is then averaged across replications. RMSE is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\theta_r - \theta_{\text{true}})^2},$$

where R denotes the number of replications. We report Monte Carlo standard errors for both

performance measures.

2.5 Software

All simulations were conducted in R 4.5.2 (R Core Team, 2025). Data were generated using `mvtnorm` 1.3.3 (Genz and Bretz, 2009). The CLPM and its extentions were estimated in `lavaan` 0.6.20 (Rosseel, 2012), and XGB was implemented with `xgboost` 1.7.11.1 (Chen et al., 2025). Figures were created with `ggplot2` (part of the `tidyverse` 2.0.0; Wickham et al., 2019), using `viridis` 0.6.5 for color palettes (Garnier et al., 2024), `ggh4x` 0.3.1 for facet axis handling (van den Brand, 2025), and `patchwork` 1.3.2 for plot grids (Pedersen, 2025).

3 Results

Figure 2 presents the results of simulation study 1. In the constant scenario, all models yield unbiased estimates, except for the unadjusted CLPM. The RMSE plots indicate that bias is the primary source of error in this scenario, with explicit adjustment performing slightly better than the latent variable approaches. With 1,000 replications, the Monte Carlo standard errors are very small for all models, and will not be discussed further.

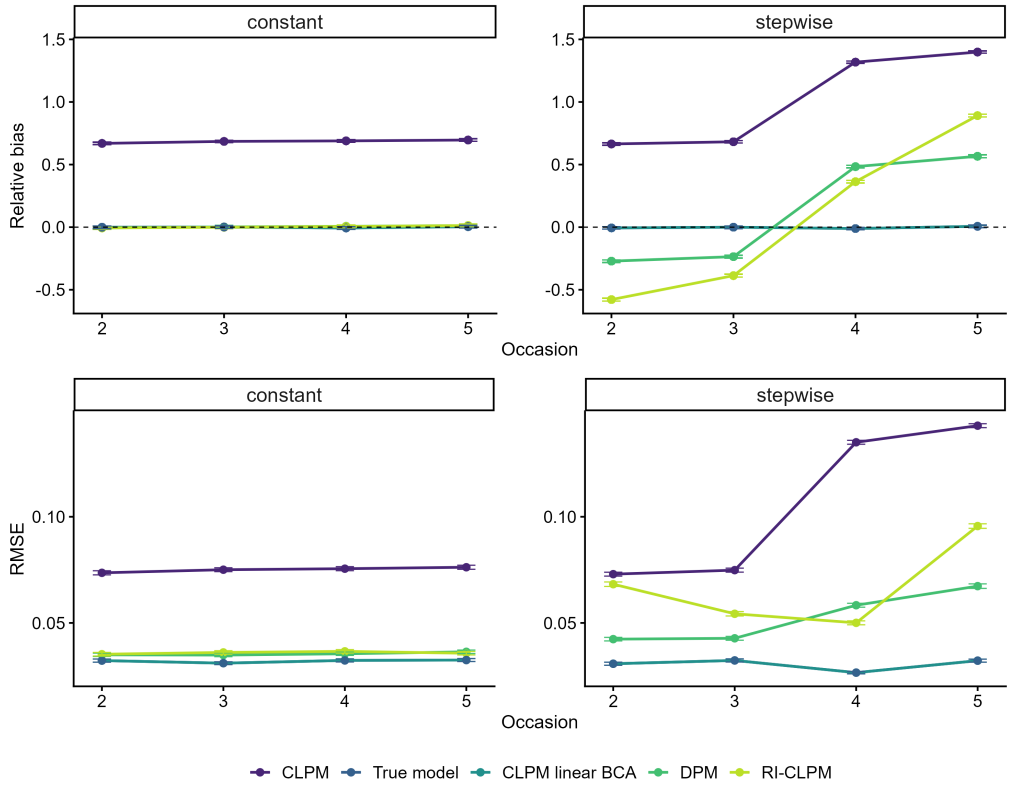

Figure 2: Simulation study 1 results – relative bias and RMSE over occasions with Monte Carlo standard errors.

As anticipated, the latent variable approaches show bias that differs across measurement occasions in the stepwise scenario. The RI-CLPM and DPM display distinct bias patterns, indicating different

ways of handling the stepwise change: the RI-CLPM spreads the bias gradually over time, whereas the DPM shows fairly stable bias before and after the step. The true model and the linear BCA CLPM remain unbiased, while the unadjusted model exhibits the largest relative bias. In the stepwise scenario, variability contributes more to overall error than in the constant scenario. The unadjusted model is more stable than the RI-CLPM and DPM, which narrows the differences between these models in terms of RMSE. The remaining RMSE difference between the RI-CLPM and DPM is largely due to bias. The explicit adjustment methods achieve the lowest RMSE overall. Figure 3 presents the results of simulation study 2.

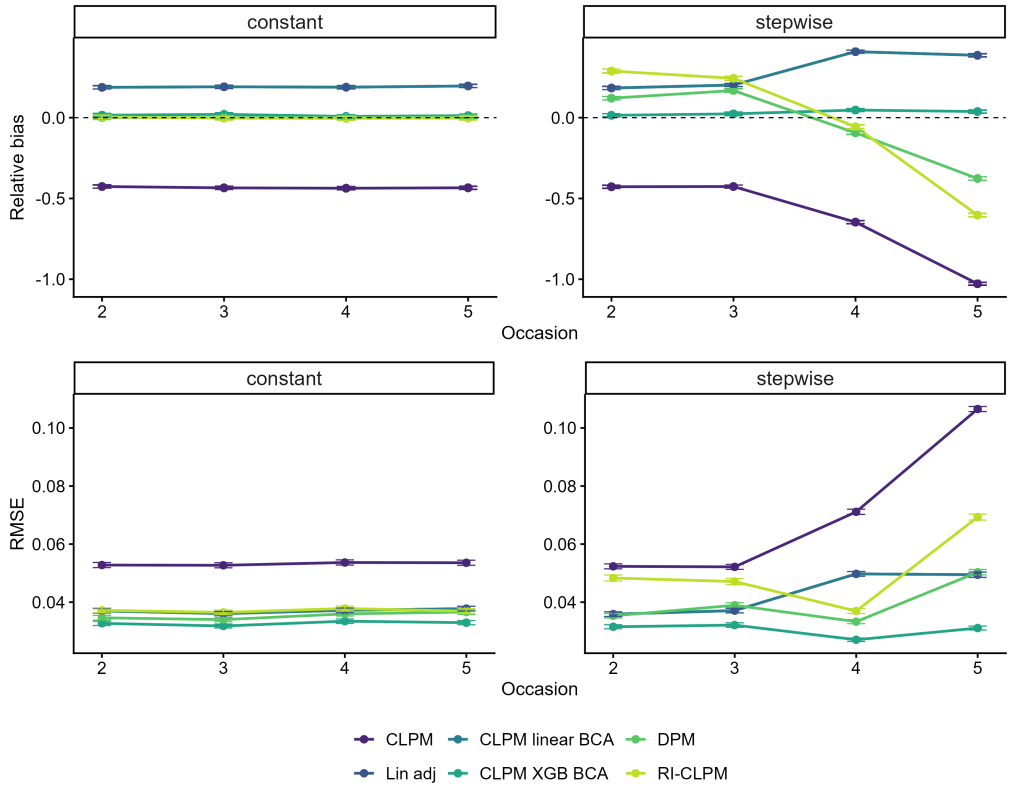

Figure 3: Simulation study 2 results – relative bias and RMSE over occasions with Monte Carlo standard errors.

In the constant scenario, models that explicitly but linearly adjust for confounders exhibit about

20 percent relative bias, reflecting their failure to capture interaction effects. Nevertheless, because they account for linear confounding, their bias is lower than that of the unadjusted CLPM. Latent variable approaches produce unbiased estimates, indicating that they remain unbiased when between-person differences are stable. The CLPM fitted to XGB-generated residuals is also unbiased, showing that the machine-learning method readily captures linear effects. Notably, the XGB BCA CLPM achieves the lowest RMSE, suggesting greater efficiency despite its added complexity. The DPM performs slightly better than the remaining methods, whereas the RI-CLPM, although unbiased, has an RMSE comparable to the linear adjustment models.

In the stepwise scenario, the latent variable approaches behave similarly to the case without non-linear data generation. As expected, the linear adjustment methods capture only linear effects and show increasing bias as the confounders explain more variance. The extreme gradient boosting approach performs best overall, although some residual bias remains, suggesting room for further tuning. In terms of RMSE, the RI-CLPM now shows clearly higher error, indicating greater variance than both the linear adjustment methods and the DPM. Although the XGB approach is closer to the other methods in this scenario, its lower bias still yields the lowest RMSE. Notably, XGB shows higher variability here than the linear methods, in contrast to the constant scenario.

4 Discussion

In this research report, two simulation studies were performed to show that (a) latent-variable approaches to baseline confounder control perform well only when confounder effects are stable over time, (b) misspecifying the functional form of covariate adjustment leads to biased cross-lagged estimates, and (c) machine-learning-based adjustment offers a promising route to relax linearity assumptions within the SEM framework. However, our design has some limitations. Orthogonalizing variables with respect to their linear components allows a clean separation of linear and non-linear variance, but it may also reduce the ability of linear adjustment methods to capture part of the non-linear confounding. Furthermore, a sample size of 1,000 likely favors higher-capacity methods,

as complex models can better exploit the additional information than simple linear approaches. However, such large samples are uncommon in the social sciences. At smaller sample sizes (e.g., 104 on average; (Fraley and Vazire, 2014)), other machine-learning approaches may be more data-efficient than the current XGB implementation. Promising alternatives include penalized regression methods such as the lasso (Tibshirani, 1996) and recent small-sample transformer models for tabular data such as TabPFN (Hollmann et al., 2022). Finally, our current implementation of ML does not yield standard errors, crucial for applied researchers to obtain confidence intervals and p-values needed for hypothesis testing. Baseline covariate adjustment methods underestimate standard errors when not properly corrected, because the SEM does not account for the uncertainty introduced by the residualization step (Freckleton, 2002). Addressing this limitation, together with the aforementioned issues will be the topic of inquiry over the coming months, to hopefully be included in the thesis.

5 Acknowledgments

The figures in this research report are strongly inspired on figures in forthcoming work by Vink et al. (2025).

References

Allison, P. D., Williams, R., and Moral-Benito, E. (2017). Maximum likelihood for cross-lagged panel models with fixed effects. *Socius*, 3:2378023117710578.

Bollen, K. A. and Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social forces*, 89(1):1–34.

Chen, T. (2016). Xgboost: A scalable tree boosting system. *Cornell University*.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano,

I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., and Yuan, J. (2025). *xgboost: Extreme Gradient Boosting*. R package version 1.7.11.1.

Dishop, C. R. and DeShon, R. P. (2022). A tutorial on bollen and brand's approach to modeling dynamics while attending to dynamic panel bias. *Psychological methods*, 27(6):1089.

Fraley, R. C. and Vazire, S. (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one*, 9(10):e109019.

Freckleton, R. P. (2002). On the misuse of residuals in ecology: regression of residuals vs. multiple regression. *Journal of Animal Ecology*, pages 542–545.

Garnier, S., Ross, N., Rudis, R., Camargo, A. P., Sciaini, M., and Scherer, C. (2024). *viridis: Colorblind-Friendly Color Maps for R*. R package version 0.6.5.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.

Hamaker, E. L., Kuiper, R. M., and Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological methods*, 20(1):102.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Hayes, A. M., Laurenceau, J.-P., Feldman, G., Strauss, J. L., and Cardaciotto, L. (2007). Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. *Clinical psychology review*, 27(6):715–723.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. (2022). Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*.

Irmer, J. P., Mulder, J. D., Vink, P. A., and Hamaker, E. L. (2025). Control for (non)linear effects of baseline confounders in causal cross-lagged panel research using baseline covariate adjusted sem. In preparation.

Mize, T. D. (2019). Best practices for estimating, interpreting, and presenting nonlinear interaction effects. *Sociological Science*, 6:81–117.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102.

Murayama, K. and Gfrörer, T. (2024). Thinking clearly about time-invariant confounders in cross-lagged panel models: A guide for choosing a statistical model from a causal inference perspective. *Psychological Methods*.

Orth, U., Clark, D. A., Donnellan, M. B., and Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of personality and social psychology*, 120(4):1013.

Pedersen, T. L. (2025). *patchwork: The Composer of Plots*. R package version 1.3.2.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Sheetal, A., Jiang, Z., and Di Milia, L. (2023). Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers. *Applied Psychology*, 72(3):1339–1364.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, pages 289–310.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Usami, S., Murayama, K., and Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological methods*, 24(5):637.

van den Brand, T. (2025). *ggh4x: Hacks for 'ggplot2'*. R package version 0.3.1.

Vink, P. A., Mulder, J. D., Irmer, J. P., and Hamaker, E. L. (2025). Methods to account for unobserved baseline confounders with time-invariant and time-varying effects in cross-lagged panel research. Manuscript in preparation.

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.