How to Evaluate Causal Dominance Hypotheses in Multilevel Vector Autoregressive Models

RESEARCH REPORT

Martijn Leonardus Gerardus van Dam (6806759)[1]

m.l.g.vandam@uu.nl

Supervisor: Dr. Rebecca M. Kuiper[1]

[1]Utrecht University

December 22nd, 2024

FETC-approved: 24-1988

Word Count: 2,489

Introduction

Background and rationale

Intensive longitudinal data (ILD) is becoming increasingly popular in social science, as it allows researchers to study processes of theoretical constructs in more detail than with cross-sectional data (McNeish and Hamaker 2020). ILD differs from longitudinal data in that it consists of more repeated measures, with relatively small time intervals between measurements (McNeish and Hamaker 2020). In social sciences, the dynamics of such processes are often of interest. Models that allow for the modeling of dynamics are Vector Autoregressive models (VAR; (Lütkepohl 2013)). In VAR models, variables are predicted using their own lagged values, which quantify the autoregressive effects (Chatfield and Haipeng 2019). Additionally, variables are regressed on lagged values of other variables to estimate cross-lagged effects, capturing how phenomena influence each other over time (Granger 1969; Chatfield and Haipeng 2019). Typically, regression coefficients are estimated using a single lag (lag-1), simplifying the model to a VAR(1) structure (Chatfield and Haipeng 2019). One can extend VAR(1) models to multilevel models, resulting in parameter estimates at the sample and person levels (Schuurman et al. 2016). Multilevel analysis aligns with the person-specific paradigm in social sciences, which states that information on processes measured at the sample level can hardly ever be applied to individuals (Molenaar and Campbell 2009). Thus, dynamic effects found at the sample level may not be true for every individual in the population.

Often, researchers have expectations about their data, which can be formalized in informative hypotheses, that is, hypotheses with (in)equality constraints on parameters (Hoijtink 2011). In VAR(1) models, these expectations mainly concern the strength of two cross-lagged parameters. Here, researchers expect one cross-lagged relationship to be stronger in absolute size over the other cross-lagged relationship (Schuurman et al. 2016; Sukpan and Kuiper 2024), known as causal dominance (Granger 1969). One way to evaluate informative hypotheses is with information criteria. (Kuiper, Hoijtink, and Silvapulle 2011) proposed the Generalized Order Restricted Information Criterion (GORIC), an extension of the Akaike's Information Criterion (AIC; (Akaike 1974)), which applies to linear normal models. The approximated GORIC (GORICA; (Altinisik et al. 2021)) extends this ability to a general class of linear models, allowing for informative hypothesis evaluation in various statistical models.

No empirical method is currently assessed for evaluating causal dominance in multilevel VAR(1) models. As a consequence, researchers are not able to formally evaluate such theories.

Given the adequate performance of the GORICA in various statistical models (see e.g., (Sukpan and Kuiper 2024; Altinisik et al. 2021)), it would be interesting to assess whether the GORICA applies to informative hypothesis evaluation in multilevel VAR(1) models. Hence, to provide researchers with such a tool, this study evaluates the properties of the GORICA in the models mentioned above. The main research question is: *What is the performance of the GORICA in multilevel VAR(1) models under varying conditions for evaluation of causal dominance?* Akin to previous studies (Altinisik et al. 2021; Sukpan and Kuiper 2024), a simulation study in which certain conditions, namely sample size, measurement occasions, and parameter values vary, is employed to answer the research question. To demonstrate the possible applications of the GORICA in multilevel VAR(1) models, hypotheses will be evaluated for both the sample and person levels.

The remainder of the article is organized as follows. First, multilevel VAR(1) is explained in more detail. Second, informative hypothesis evaluation using the GORICA is explained. Third, the merits of the simulation study are explained. Fourth, the results of the pilot simulations will be interpreted. Finally, the implications of the results and future research directions are discussed.

Multilevel VAR(1)

At its core, a vector autoregressive (VAR(1)) model predicts variable measurements based on preceding values at lag 1, capturing both autoregressive and cross-lagged effects (Chatfield and Haipeng 2019). The multilevel extension of this model estimates parameters at two levels: the sample level (fixed effects) and the individual level as random effects. To account for variability within and between individuals, the model uses deviations from person-specific means to estimate lagged effects at the within-person level. These deviations are assumed to be explained by preceding deviations of the same variable (autoregressive effects), other variables (cross-lagged effects), and residual variance. At the between-person level, means and parameter variability are modeled, assuming a normal distribution for the means. Fixed effects are assumed to follow a specific distribution from which person-specific parameters are sampled. Finally, random effects are often correlated to capture relationships between effects (Schuurman et al. 2016).

To fit the dynamic within-person processes while accounting for between-person variability, multilevel VAR(1) models are estimated using Bayesian techniques. The advantages of Bayesian techniques over maximum likelihood estimation are that Bayesian techniques are adaptive to varying model features (Schuurman et al. 2016; Asparahouv, Hamaker, and Muthén 2018) and that the complete model can be fit simultaneously without having to model each person separately. The details of Bayesian analysis in VAR(1) models nor in general are discussed here (for VAR(1), see, e.g., (Schuurman et al. 2016; Asparahouv, Hamaker, and Muthén 2018); for Bayesian in general, see, e.g., (Van de Schoot et al. 2013; Lambert 2018)).

GORICA

Akin to the AIC, the GORICA identifies the hypothesis closest to the true data-generating process (Altinisik et al. 2021). GORICA values are calculated for a set of $M + 1$ informative hypotheses, where M indicates the number of informative hypotheses in a set to which one additional safeguard hypothesis is added. A safeguard hypothesis includes the restrictions on parameters not captured in the informative hypothesis, or is an unconstrained hypothesis (Kuiper and Hoijtink 2013). Maximum likelihood estimates (MLE)[1] and the covariance matrix of the parameters in the hypothesis are required to calculate the GORICA values. For hypotheses on the comparison of regression coefficients, standardized parameter estimates are required to make a fair comparison (Schuurman et al. 2016). The GORICA for an informative hypothesis H_m is calculated as:

$$GORICA_m = -2L(\tilde{\theta}^m|\hat{\theta}, \hat{\Sigma}_{\hat{\theta}}) + 2PT_m(\theta),$$

where all θs indicate the parameters under evaluation and Σ the covariance matrix. The first part of the equation is referred to as the misfit. It is calculated as -2 times the MLEs of the parameter $\tilde{\theta}$ under the restrictions in hypothesis H_m, conditional on the MLEs of the unrestricted parameters $\hat{\theta}$ and their covariance matrix $\hat{\Sigma}_{\hat{\theta}}$. The second part is the penalty term, or complexity, reflecting the expected distinct parameters in the hypothesis (Altinisik et al. 2021). The hypothesis with the smallest GORICA value in a set of hypotheses is considered the best hypothesis out of that set.

[1] The GORICA is derived for the MLEs, but the model is estimated using Bayesian techniques. Bayesian methods do not always yield the exact MLEs but often yield estimates (very) close to the MLEs (Cosineau and Hélie, 2013). Moreover, the GORICA is based on the assumption of the likelihood being asymptotically normally distributed. Hence, this study evaluates whether the estimates obtained with Bayesian techniques also suffice.

GORICA values can only indicate which hypothesis has more support by the data but not how much more support. To enhance the interpretability of the GORICA, weights can be calculated that quantify the relative support in the data for one hypothesis over all other hypotheses in a set:

$$w_m = \frac{\exp\{-\frac{1}{2} GORICA_m\}}{\sum_{m'=1}^{M+1} \exp\{-\frac{1}{2} GORICA_{m'}\}},$$

where m is the hypothesis index, and $M+1$ denotes the total number of informative hypotheses plus one safeguard hypothesis. The weights range from 0 to 1, and a ratio of two weights quantifies how much more support there is in the data for one hypothesis over another out of the set. For instance, if H_m and H_{m+1} have weights of 0.8 and 0.2, respectively, then H_m is a better hypothesis than H_{m+1} as $0.8 > 0.2$. Notably, there is $0.8/0.2 = 4$ times more support for H_m than H_{m+1}.

Methods

Simulation

This study aimed to assess the performance of the GORICA in multilevel VAR(1) models to evaluate causal dominance hypotheses. To assess its performance, a simulation study in `R` (R Core Team 2021, version 4.3.1) was conducted where the number of participants, measurement occasions, parameter values of the cross-lagged coefficients, and hypotheses to evaluate were varied. Data were generated for a bivariate multilevel VAR(1) model. A trivariate model will be generated in future work by extending the bivariate model to include a third variable with autoregressive and cross-lagged structures. Specifications of the data and the data generation process for the bivariate multilevel VAR(1) model were based on the study of (Schuurman et al. 2016). Two variables, Y_1 and Y_2, were simulated with fixed autoregressive effects set to .4 for both variables and means of 3 and 2, respectively. The cross-lagged effects took on the following values: $\{\theta_{12}, \theta_{21}\} \in \{0.20, 0.10;\ 0.20, 0.15;\ 0.15, 0.15\}$, where the first pair was equal to the specification in (Schuurman et al. 2016). Additional parameter sets were selected to generate data where the parameter values were equal, and where the difference between the values was smaller than in the first set. To introduce individual-level variability, the autoregressive and cross-lagged parameters were modeled with a variance of .01. The variances of Y_1 and Y_2 were fixed at .25. These settings ensured that the generated data reflected meaningful dynamic processes while maintaining consistency with (Schuurman et al. 2016). Covariances of random effects were fixed at zero to decrease the computational time of model estimation. The number of participants (N) and measurement occasions (T) were varied to explore conditions commonly encountered in practice and to ensure robust parameter estimation. Sample sizes (N) took on the values {50, 75, 100, 150, 200}, reflecting findings by (McNeish 2019), who noted that $N < 50$ can produce biased estimates in similar models. Measurement occasions (T) took on the values {25, 50, 75, 100} to capture both relatively small and large designs. This choice was informed by recommendations from (Schultzberg and Muthén 2018) for unbiased estimation in multilevel AR(1) models and the range of values typically used in empirical studies (McNeish 2019). These conditions allowed for evaluating the GORICA's performance across various practical scenarios. An overview of the conditions is shown in Table 1.

Table 1: Simulation Conditions for the Bivariate Multilevel VAR(1) Model.

N	T	$\theta_{12},\ \theta_{21}$	Sets of hypotheses
50, 75, 100	25, 50	0.20, 0.10;	Set 1: $H_1 : \theta_{12} > \theta_{21}$ vs. H_{1c} : not H_1
150, 200	75, 100	0.20, 0.15;	Set 2: H_1 vs. $H_0 : \theta_{12} = \theta_{21}$ vs. H_{1c}
		0.15, 0.15	Set 3: $H_{a1} : -0.05 < \theta_{12} - \theta_{21} < 0.05$ vs. H_{a1c} : not H_{a1}
			Set 4: $H_{a2} : -0.01 < \theta_{12} - \theta_{21} < 0.01$ vs. H_{a2c} : not H_{a2}

Note. Each condition consists of sample size (N), measurement occasions (T), and pair of parameters ($\theta_{12},\ \theta_{21}$), where θ_{12} refers to the cross-lagged effect of Y_2 on Y_1 and θ_{21} of Y_1 on Y_2. The four sets of hypotheses are evaluated in each condition. On the sample level, H_1, H_{a1}, and H_{ha2} are true for the first two pairs of parameters and H_0 for the last parameter pair, while H_{a1} and H_{a2} are more parsimonious than H_0.

For the pilot simulation, 15 datasets were generated for the conditions with $N = 50, T = 25$ and $N = 75, T = 50$, across all parameter pairs. These datasets were analyzed using a multilevel VAR(1) model with the `rjags` package (Plummer 2024) and the `coda` package (Plummer et al. 2006) in R. The hypotheses were evaluated with the `goric` function of the `restriktor` package (Vanbrabant and Rosseel 2024). The specifications of the priors are shown in Table 2. The pilot simulation was conducted to provide preliminary insights into the GORICA's performance and to inform decisions on the appropriate number of datasets required for the full simulation study, where between 100 and 500 datasets will be generated per condition.

After obtaining the estimates for the parameters of interest, four sets of hypotheses, shown in Table 1 for the bivariate model, are evaluated to obtain the GORICA weights. The hypotheses are evaluated at the sample level and for each individual within a sample. Conclusions regarding the performance of the GORICA are measured by the rate at which the 'true' hypothesis (THR), that is, the hypothesis in accordance with the data-generating mechanism of a condition, is selected. A hypothesis is deemed correctly selected when it achieves the largest GORICA weight out of the set. Additionally, the ratio of GORICA weights is assessed to gain insight into the relative support for the hypothesis. In the pilot simulations, plots are created to visualize the variability of the GORICA weights across conditions. In the full simulation study, additional THR plots and tables with the median, 5^{th}, and 95^{th} percentiles of the weight ratio will be reported to provide a more detailed evaluation. Conditions with large N and/or T are expected to yield larger GORICA weights, weight ratios, and less fluctuation of the weights than conditions with small N and/or T. In the conditions where the parameters are equal, the THR is expected to converge to .50 when H_1 is evaluated against its complement, and H_{a1} and H_{a2} (see Table 1) are expected to yield the largest weights in the equal parameter conditions. The performances of the GORICA are assessed for both the sample-level and person-level estimates. Note that, due to the multilevel nature of the model, person-specific parameters are allowed to be incongruent with the true hypothesis. In such cases, the hypothesis that is true at the sample level may not be true.

Table 2: Prior Specifications for the Bivariate Multilevel VAR(1) Model

Parameter	Prior Specification
Fixed effects ($\theta_k\mu$; autoregressive and cross-lagged)	*Normal*(0, 1000)
Variances of random effects ($\theta_k\sigma$)	*Uniform*(0, 10)
Random autoregressive effects	*Normal*($\theta_k\mu$, $\theta_k\sigma$)
Random cross-lagged effects	*Normal*($\theta_k\mu$, $\theta_k\sigma$)
Residual variances (within-person, diagonal elements)	*Uniform*(0, 10)
Residual covariances (within-person, off-diagonal elements)	*Uniform*(-1, 1)

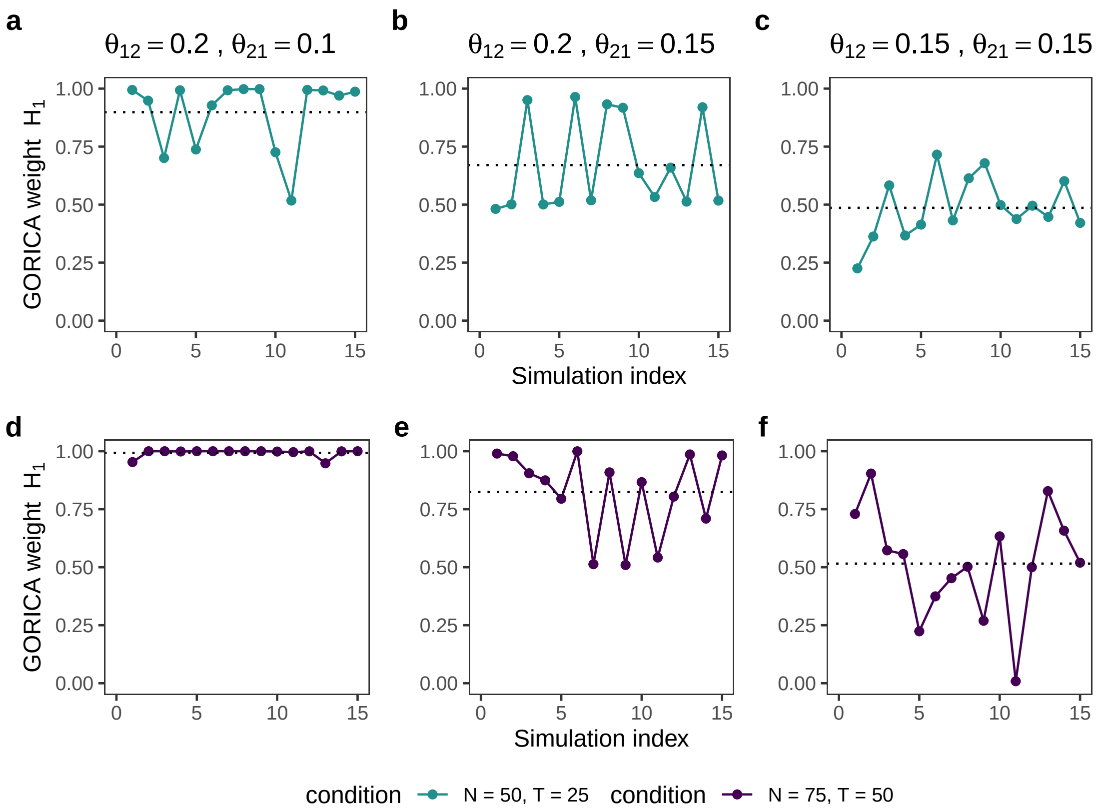
Note. The prior specifications in this paper are mainly based on Schuurman et al., 2016, and were weak or noninformative to aid the results primarily by the data. The variance of the fixed effect was decreased to 1000 in this study, as it was scaled to the data. Here, $\theta_k\mu$ and $\theta_k\sigma$ refer to the value of the fixed effect and variance of random effect for each regression parameter, respectively, as denoted by the subscript k. The random autoregressive and cross-lagged effects are the person-specific parameter estimates, sampled with fixed effects as hyperparameters.

Results

The pilot simulations examined the GORICA weights for $H_1 : \theta_{12} > \theta_{21}$ against its complement $H_c : \theta_{12} < \theta_{21}$ across six conditions defined by varying cross-lagged parameters ($\theta_{12},\ \theta_{21}$) and sample size/measurement occasions ($N = 50,\ T = 25$ or $N = 75,\ T = 50$). The results, shown in Figure 1, indicate that the GORICA weights are generally higher for H_1 in conditions where $\theta_{12} > \theta_{21}$. For $\theta_{12} = 0.2,\ \theta_{21} = 0.1$ (plots a and d), H_1 is consistently favored, with GORICA weights close to 1 across all simulations. This pattern is observed for both $N = 50,\ T = 25$ and $N = 75,\ T = 50$. For $\theta_{12} = 0.2, \theta_{21} = 0.15$ (plots b and e), the GORICA weights show greater variability, particularly in the smaller sample condition ($N = 50,\ T = 25$), where some weights favor H_c. Larger sample sizes ($N = 75,\ T = 50$) lead to more consistent support for H_1, with GORICA weights often exceeding 0.75. For $\theta_{12} = \theta_{21} = 0.15$ (plots c and f), the GORICA weights fluctuate around 0.5, reflecting equal support for H_1 and H_c. This is expected since the data-generating mechanism does not favor either hypothesis. The dashed lines in the plots represent the mean GORICA weight for each condition, further illustrating the increased stability and support for H_1 in larger sample size and measurement occasion conditions. On the contrary, in plot f, the weights show a large variability, even larger than in plot c. This is contrary to expectation, as one would expect the weights in the condition shown in plot c to show greater variability. The fluctuations may resulted from random sampling variability, as only 15 data sets were generated per condition.

Therefore, the variation in plot f is expected to be smaller than in plot c when additional data sets are created. Overall, these findings suggest that a larger sample size and number of measurement occasions contribute to the GORICA weights converging to the expected values of 1 and .5.

Figure 1: Results of pilot simulations for all parameter pairs and two different conditions of sample size and measurement occasions for the sample-level estimates.



Note. For all conditions, the GORICA weights of $H_1 : \theta_{12} > \theta_{21}$ are shown, which is evaluated against its complement $H_c : \theta_{12} < \theta_{21}$. The dashed lines indicate the average weights in the conditions.

Discussion

This pilot study assessed the performance of the GORICA in evaluating causal dominance hypotheses in multilevel VAR(1) models under varying sample sizes, measurement occasions, cross-lagged parameter specifications, and hypotheses. The results demonstrated that the GORICA reliably identified the true hypotheses when the data-generating mechanism supported it, with improving stability and precision in conditions with larger sample sizes and measurement occasions. As expected, when parameters were equal, GORICA weights appropriately fluctuated around 0.5, indicating no preference for either hypothesis. Additionally, when the difference between the parameters was larger, the GORICA weights showed less variability in the condition with a larger sample size and more measurement occasions.

These findings highlight the GORICA's potential as a tool for evaluating informative hypotheses in multilevel VAR(1) models, which aligns with prior research demonstrating the utility of the GORICA in various statistical models (see e.g., (Altinisik et al. 2021); (Sukpan and Kuiper 2024)). However, some unexpected variability in weights, particularly in the large sample condition with a pair of equal parameter values, suggests additional simulations are needed to confirm these results and ensure the robustness of the GORICA weights. Additionally, future research should vary small sample sizes with large numbers of measurement occasions and vice versa to gain insight into the contribution of the two conditions.

A potential limitation of the GORICA is the reliance on Bayesian estimates instead of maximum likelihood estimates (MLEs), which the GORICA assumes. While Bayesian estimation can produce values close to MLEs (Cousineau and Hélie 2013), it remains unclear for which datasets this was the case. This reliance introduces the possibility of bias in the GORICA values and weights. Additionally, the inflexibility of maximum likelihood estimation in multilevel VAR(1) models makes cross-validation of the results challenging. However, if the priors of the parameters are uninformative enough, the posterior parameter estimates should converge to MLEs, potentially minimizing the bias in the results. If the GORICA continues to perform well under these conditions, it provides evidence that, with uninformative priors, it can be a suitable method for evaluating causal dominance in multilevel VAR(1) models. Future research should investigate the extent and impact of potential bias in GORICA values and weights.

In conclusion, the pilot results suggest that the GORICA shows potential as a method for evaluating causal dominance in multilevel VAR(1) models. While initial findings are encouraging, the small number of datasets and conditions limit generalizability. Future research should include larger-scale simulations to assess the consistency of GORICA weights under diverse conditions and further investigate potential biases associated with Bayesian estimation. Despite the limitations, this study represents a meaningful step towards adapting the GORICA in multilevel VAR(1) models.

References

Akaike, H. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19 (6): 716–23. https://doi.org/10.1109/TAC.1974.1100705.

Altinisik, Y., C. J. Van Lissa, H. Hoijtink, A. J. Oldehinkel, and R. M. Kuiper. 2021. "Evaluation of Inequality Constrained Hypotheses Using a Generalization of the AIC." *Psychological Methods* 26 (5): 599–621. https://doi.org/10.1037/met0000406.

Asparahouv, T., E. L. Hamaker, and B. Muthén. 2018. "Dynamic Structural Equation Models." *Structural Equation Modeling: A Multidisciplinary Journal* 25 (3): 359–88. https://doi.org/10.1080/10705511.2017.1406803.

Chatfield, C., and X. Haipeng. 2019. *The Analysis of Time Series: An Introduction with r*. Boca Raton, Florida: CRC Press.

Cousineau, D., and S. Hélie. 2013. "Improving Maximum Likelihood Estimation Using Prior Probabilities: A Tutorial on Maximum a Posteriori Estimation and an Examination of the Weibull Distribution." *Tutorials in Quantitative Methods for Psychology* 9 (2): 61–71. https://doi.org/10.20982/tqmp.09.2.p061.

Granger, C. W. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica: Journal of the Econometric Society*, 424–38.

Hoijtink, H. 2011. *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. 1st ed. Chapman; Hall/CRC. https://doi.org/10.1201/b11158.

Kuiper, R. M., and H. Hoijtink. 2013. "A Fortran90 Program for the Generalized Order-Restricted Information Criterion." *Journal of Statistical Software* 54 (8). https://doi.org/10.18637/jss.v054.i08.

Kuiper, R. M., H. Hoijtink, and M. J. Silvapulle. 2011. "An Akaike-type information criterion for model selection under inequality constraints." *Biometrika* 98 (2): 495–501. https://doi.org/10.1093/biomet/asr002.

Lambert, B. 2018. *A Student's Guide to Bayesian Statistics*. London, UK: SAGE Publications Limited.

Lütkepohl, H. 2013. "Vector Autoregressive Models." In *Handbook of Research Methods and Applications in Empirical Macroeconomics*, edited by N. Hashimzade and M. A. Thornton, 139–64. Edward Elgar Publishing.

McNeish, D. 2019. "Two-Level Dynamic Structural Equation Models with Small Samples." *Structural Equation Modeling* 26 (6): 948–66. https://doi.org/10.1080/10705511.2019.1578657.

McNeish, D., and E. L. Hamaker. 2020. "A Primer on Two-Level Dynamic Structural Equation Models for Intensive Longitudinal Data in Mplus." *Psychological Methods* 25 (5): 610–35. https://doi.org/10.1037/met0000250.

Molenaar, P. C. M., and C. G. Campbell. 2009. "The New Person-Specific Paradigm in Psychology." *Current Directions in Psychological Science* 18 (2): 112–17. https://doi.org/10.1111/j.1467-8721.2009.01619.x.

Plummer, M. 2024. *Rjags: Bayesian Graphical Models Using MCMC*. https://CRAN.R-project.org/package=rjags.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News* 6 (1): 7–11. https://journal.r-project.org/archive/.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schultzberg, M., and B. Muthén. 2018. "Number of Subjects and Time Points Needed for Multilevel Time-Series Analysis: A Simulation Study of Dynamic Structural Equation Modeling." *Structural Equation Modeling: A Multidisciplinary Journal* 25 (4): 495–515. https://doi.org/10.1080/10705511.2017.1392862.

Schuurman, N. K., E. Ferrer, M. de Boer-Sonnenschein, and E. L. Hamaker. 2016. "How to Compare Cross-Lagged Associations in a Multilevel Autoregressive Model." *Psychological Methods* 21 (2): 206–21. https://doi.org/10.1037/met0000062.

Sukpan, C., and R. M. Kuiper. 2024. "How to Evaluate Causal Dominance Hypotheses in Lagged Effects Models." *Structural Equation Modeling: A Multidisciplinary Journal* 31 (3): 404–19. https://doi.org/10.1080/10705511.2023.2265065.

Van de Schoot, R., D. Kaplan, J. Denissen, J. B. Asendorpf, F. J. Neyer, and M. A. G. Van Aken. 2013. "A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research." *Child Development* 85 (3): 842–60. https://doi.org/10.1111/cdev.12169.

Vanbrabant, L., and Y. Rosseel. 2024. *Restriktor: Restricted Statistical Estimation and Inference for Linear Models*. https://github.com/LeonardV/restriktor.