Evaluation of Univariate and Multivariate Dynamic Structural Equation Models with Categorical Outcomes

by

Andrea Savord

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2023 by the
Graduate Supervisory Committee:

Daniel McNeish, Chair
Kevin J. Grimm
Masumi Iida
Roy Levy

ARIZONA STATE UNIVERSITY

August 2023

ABSTRACT

The proliferation of intensive longitudinal datasets has necessitated the development of analytical techniques that are flexible and accessible to researchers collecting dyadic or individual data. Dynamic structural equation models (DSEMs), as implemented in M*plus*, provides the flexibility researchers require by combining components from multilevel modeling, structural equation modeling, and time series analyses. This dissertation project presents a simulation study that evaluates the performance of categorical DSEM using a probit link function across different numbers of clusters (N = 50 or 200), timepoints (T = 14, 28, or 56), categories on the outcome (2, 3, or 5), and distribution of responses on the outcome (symmetric/approximate normal, skewed, or uniform) for both univariate and multivariate models (representing individual data and dyadic longitudinal Actor-Partner Interdependence Model data, respectively). The 3- and 5-category model conditions were also evaluated as continuous DSEMs across the same cluster, timepoint, and distribution conditions to evaluate to what extent ignoring the categorical nature of the outcome impacted model performance. Results indicated that previously-suggested minimums for number of clusters and timepoints from studies evaluating continuous DSEM performance with continuous outcomes are not large enough to produce unbiased and adequately powered models in categorical DSEM. The distribution of responses on the outcome did not have a noticeable impact in model performance for categorical DSEM, but did affect model performance when fitting a continuous DSEM to the same datasets. Ignoring the categorical nature of the outcome lead to underestimated effects across parameters and conditions, and showed large Type-I error rates in the N = 200 cluster conditions.

To Arthur, whose endless support and encouragement got me through the longest nights and hardest days with this project. Thank you for always knowing when I needed coffee and for helping me stay focused. I love you, and look forward to spending the rest of our lives together.

ACKNOWLEDGMENTS

I am thankful for the support and guidance of my advisor, Dr. Daniel McNeish. Thank you for believing in me, for always being willing to meet with me, and for being an incredible teacher, mentor, and collaborator. My growth as a writer, researcher, and quantitative thinker over the past five years is due in no small part to your feedback and encouragement; I could not have asked for a better advisor.

I would also like to thank my other committee members: Drs. Kevin Grimm, Roy Levi, and Masumi Iida, for their feedback and guidance on this project. Masumi, I thank you for teaching an excellent class introducing me to dyadic data analyses and for providing me with so many opportunities to learn and grow since then. Roy, I am grateful that my first semester here was spent in your classroom every Tuesday; Bayes’ theorem still feels a little bit like magic, in the best of ways. Kevin, thank you for your expertise and guidance, and for providing me with foundational path tracing skills several years ago that were critical to a portion of this project.

Thank you to Dr. Thao Ha, and to Selena Quiroz and Olivia Maras, for providing me with the dataset for the motivating data example and for giving me the opportunity to collaborate in your lab. I really appreciate you trusting me with your data and for providing me with invaluable experience with dyadic data.

I am also grateful for Dr. Samantha Anderson and Dr. David MacKinnon, whose courses I took my first semester at ASU; your enthusiasm and mentorship really solidified to me that I had made the correct choice in joining this program and that I could belong here. I thank Dr. Mike Edwards, too, for pushing my thinking on

measurement and for once saying “Only tell a joke if you know it is funny” in a presentation about how to give presentations – very serious advice I hope to one day take.

Thank you to my fellow quant grad students, past and present: Heather Gunn, whose guidance and friendship got me through my first year in Arizona and who I am endlessly grateful to still have as a friend. Heather Smyth, who reminded me to take care of myself (and to eat good sushi whenever possible). Molly Gardner, who has always been there to help me talk through issues (school, life, and otherwise), give me advice, and make me laugh. Thanks also to Russell Houpt, for reminding me that you can still have hobbies (and enjoy them!) even though you are a graduate student.

I am endlessly grateful to Skye Mendes and Renee Benoit for laughing, crying, working, struggling, succeeding, failing, and learning alongside me. Thank you for taking me into your duo and creating the best coven a girl could ask for. I could not have done this without you both, and I cannot wait to celebrate when we are all finished and to see where life takes us next. I will always be down for roasting marshmallows and drinking wine in Skye’s backyard.

Finally, I would like to thank my family: my mother, father, Adrienne, and Jessica, for supporting me on my move across the country and throughout the time I have spent here. Thank you for visiting me and letting me share my favorite parts of the desert with you. Thank you to my mother- and father-in-law for encouraging me and for raising one of my favorite human beings. And finally, thank you again to Arthur. Life is better with you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION TO DYNAMIC STRUCTURAL EQUATION MODELS

Social and behavioral researchers collect longitudinal data as a way to assess change or variability in an outcome, which can be examined at both a within-cluster (individual, dyad) and between-cluster level. Advances in data collection techniques have enabled designs such as experience sampling (Scollon, Kim-Prieto, & Diener, 2003), ambulatory assessment (Fahrenberg et al., 2007), ecological momentary assessment (Shiffman, Stone, & Hufford, 2008), and daily diaries (Bolger, Davis, & Rafaeli, 2003) that allow researchers to gather many timepoints over a relatively short period of time and ask questions about daily life in real time outside of a laboratory setting. Dyadic datasets – where information is collected from two participants who are interdependent (e.g., couples, siblings, friends) – are particularly well suited for ILDs as they allow researchers to examine changes in relationship dynamics over the course of several hours, days, or weeks. These intensive longitudinal datasets (ILDs) and accompanying designs have, in turn, necessitated the development of statistical methods that are appropriate to accommodate the unique elements of ILDs (Jebb et al., 2015).

Dynamic structural equation modeling (DSEM) is a method recently added as a module in M*plus* that is ideally suited for modeling data from ILDs by integrating aspects of multilevel modeling (MLM), structural equation modeling (SEM), and time-series analysis (Asparouhov, Hamaker, & Muthén, 2018; Hamaker et al., 2018; McNeish & Hamaker, 2020). By incorporating these three modeling techniques, DSEM can respectively accommodate repeated measures nested within clusters, multivariate outcomes and latent variables with flexible modeling of variables, and lagged relations for modeling autoregressive effects. This is especially useful in the context of dyadic ILDs, where both the MLM and SEM frameworks often simplify the model so that the collected data conform to the model rather than the model serving the needs of the data and the research questions (Planalp et al., 2017).

Along with providing flexible modeling of ILDs, the DSEM module in M*plus* can accommodate categorical outcomes using a probit link function (Asparouhov & Muthén, 2019). Most research produced using DSEM focuses on continuous outcome measures, though given the popularity of daily diary assessments for intensive longitudinal data collection (Gunthert & Wenze, 2014; Bolger et al., 2003) and the common use of categorical items within them (Nezlek, 2014; Bolger & Laurenceau, 2013, Chapter 6), the ability for a model uniquely suited for modeling ILDs to produce unbiased estimates when modeling binary or categorical outcomes is important.

Given the growing popularity of ILDs, and the use of DSEM to model them, further investigation into how DSEM performs when modeling categorical outcomes from daily-diary style datasets from both individuals and dyads is necessary. For this dissertation project, I propose to examine the behavior and properties of two categorical DSEM models – a univariate lag-1 autoregressive model, and a multivariate (dyadic) lag-1 vector autoregressive model that conforms to the longitudinal Actor-Partner Interdependence Model (L-APIM) framework – across differing sample sizes, number of collected timepoints, number of categories on the outcome variable, and proportion of responses in each category of the outcome. I begin with a formal overview of intensive longitudinal data with an emphasis on DSEM and how it addresses limitations encountered with MLM and SEM, along with the special considerations dyadic data (the

L-APIM in particular) and categorical outcomes present to these analyses. Then, I describe the motivating applied data example that uses DSEM to model intensive longitudinal dyadic data with categorical outcomes as an L-APIM, the results of which will be used to inform the simulation study to follow. Finally, I present the proposed methods for a simulation study that evaluates the behavior and properties of a categorical DSEM model when used with individual and dyadic datasets.

Intensive Longitudinal Data

Researchers with questions that involve the processes of daily life require data collection methods that would adequately capture the dynamics of their variables of interest. This involves collecting intensive longitudinal data that consists of many observations over a short period of time (Laurenceau & Bolger, 2005; Bolger & Laurenceau 2013; Iida et al., 2012). Intensive longitudinal data allows researchers to see outcome dynamics and processes unfold as they happen through frequent (daily, weekly) assessments, and typically do not involve a lab or clinical setting for data collection. By gathering repeated observations in a natural setting, researchers are maximizing ecological validity (i.e., whether results can be generalized to real life), and can examine not only what the typical experience is for their participants, but also how those individuals might systematically change on variables of interest, and how this experience is different within and between individuals (Iida et al., 2012).

Intensive longitudinal datasets (ILDs) provide researchers with the opportunity to ask questions for which panel and cross-sectional designs are not well suited (Bolger & Laurenceau, 2013; Chapter 8). Gathering many timepoints over a relatively short period of time allows researchers to examine the day-to-day or moment-to-moment changes and dynamics of outcomes that single-timepoint snapshots or once-yearly data collection cannot (Bolger, Davis, & Rafaeli, 2003). In this way, researchers with ILDs can ask questions related to within-person variability of stable processes instead of within-person mean change of developmental processes that is common with panel designs (Jongerling et al., 2015). The inherent density of ILDs necessitates the use of a modeling framework that is flexible enough to accommodate many different variable types and accompanying research questions that aim to tease out meaning in their variability; DSEM was developed with this goal in mind (Hamaker, Asparouhov, & Muthén, 2022).

Dynamic Structural Equation Models

The general DSEM model aims to separate observed scores into an individual-specific component and a time-specific component (both between-level variables), as well as an individual- and time-specific deviation (the within-level model). The two-level DSEM model that is the focus of this dissertation omits the time-specific component, retaining the individual-specific between-level and the within-level individual- and time-specific deviations (Asparouhov et al., 2018, p.360). Interdependence from individual-specific effects is modeled using MLM, proximity-based interdependence is modeled using time-series techniques, and between-variable correlations are modeled using SEM (p. 360). A two-level lag-1 DSEM for individual *i* at time *t* can be expressed for an outcome y_{ti} and covariate x_{it} with random intercept and slope terms as shown in Figure 1. This figure corresponds to a system of equations representing the latent decomposition, the within-level, and the between-level:

$$
\text{Latent Decomposition}\begin{cases} y_{ti} = y_{ti}^{w} + \alpha_i \\ x_{ti} = x_{ti}^{w} + x_i^{b} \end{cases}
$$

$$
\text{Within-Level}\begin{cases} y_{ti}^{w} = \varphi_i y_{(t-1)i}^{w} + \beta_i x_{ti}^{w} + e_{ti} \\ x_{ti}^{w} \sim N(0, \omega) \\ e_{ti} \sim N(0, \sigma^2) \end{cases}
$$

$$
\text{Between-Level}\begin{cases} \alpha_i = \gamma_{00} + u_{0i} \\ \varphi_i = \gamma_{10} + u_{1i} \\ \beta_i = \gamma_{20} + u_{2i} \\ x_i^{b} = \gamma_{30} + u_{3i} \\ \\ \begin{bmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \\ u_{3i} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & & & \\ 0 & \tau_{11} & & \\ 0 & 0 & \tau_{22} & \\ 0 & 0 & 0 & \tau_{33} \end{bmatrix} \right) \end{cases} \tag{1}
$$

The latent decomposition is comprised of the equations that latent person-mean center all variables whose effects are allowed to vary at the between-level. α_i is the intercept that is allowed to vary across individuals (and represents the person-specific mean of *y* across timepoints *t*), φ_i is the autoregressive (or carryover) effect for y_{ti} at the previous timepoint (lag-1) that is allowed to vary across individuals, and β_i is the effect of the covariate on the outcome that is allowed to vary across individuals. The within-level residuals e_{ti} are assumed to be normally distributed, with variance σ^2. Each of the terms that include a random effect are denoted with an *i* subscript in the latent decomposition and within-level equations; these effects become outcomes in the between-level equations where they are decomposed into a fixed effect (γ) and random effect (*u*). The random effect *u*-terms are assumed to be normally distributed each with their own variance (τ).

In this example, the random effects are not allowed to covary, and so the off-diagonal terms have been set to 0.

Notably in these equations the predictors $y^{w}_{(t-1)i}$ and x^{w}_{ti} include a *w* superscript – this denotes not only that they represent the disaggregated within-level effects for those variables but also that these variables have been *latent* person-mean centered, which is the default for DSEM in M*plus* for any parameter that is estimated to have a random effect (Asparouhov et al., 2018). Latent person-mean centering (Lüdtke et al., 2008; Asparouhov & Muthén, 2019), assumes that between-level cluster means (which are person means in longitudinal data, centered across all time points for each person) have sampling error. In this way, latent mean centering takes measurement error and unreliability of cluster means into account when disaggregating effects into within-person and between-person components in ways that observed-mean centering cannot (Curran & Bauer, 2011; Wang & Maxwell, 2015; Asparouhov & Muthén, 2019). In equation 1 above, x^{w}_{ti} has been centered around its latent mean x^{b}_{i} (where $x_{ti} = x^{w}_{ti} + x^{b}_{i}$ can be restructured as $x^{w}_{ti} = x_{ti} - x^{b}_{i}$), and the collection of x^{w}_{ti} scores (capturing deviations from the latent person mean across the time series) is assumed to be normally distributed around that latent mean with a variance of ω. Then the between-level in equation 1 includes the trait component x^{b}_{i} (capturing habitual behavior) as it represents the latent mean for each person, and the random effect u_{3i} allows each person to have a unique latent mean. The latent-mean centering process for y_{ti} is captured via the within-person component (y^{w}_{ti}) and the between-person component (α_i), the latter of which represents the latent mean of y_{ti} (as x_i^b is for x_{ti}) and is also the intercept term. Thus, similar to how the expression $x_i^b = \gamma_{30} + u_{3i}$ represents how each person has their own latent mean for x_{it}, the expression $\alpha_i = \gamma_{00} + u_{0i}$ at the between-level allows each person to have their own latent mean for y_{ti} (which in this case also represents a random intercept term). Latent person-mean centering is similar to traditional person-mean centering (also referred to as group-mean centering in a cross-sectional analysis, or more generally as cluster-mean centering) except that the person-mean is treated as an unknown quantity to be estimated (Asparouhov & Muthén, 2019). By treating the person-mean as a latent variable that is estimated in the model researchers can eliminate both Nickell's bias and Lüdtke's bias, giving them more interpretable effects from their model. Nickell's bias refers to a negative bias in the autoregressive effect when the lagged predictor is centered around the observed person-mean and is the result of failing to account for error in the observed mean estimate of the true person-mean (Nickell, 1981; Asparouhov et al., 2018). Relatedly, Lüdtke's bias refers to bias in the between-level regression coefficients due to the difference between the observed person-mean estimate and the true person-mean (Lüdtke et al., 2008).

Centering predictors is an important aspect of multilevel modeling as it disaggregates coefficients into within-person and between-person components for unconfounded interpretation (Hamaker & Grasman, 2015; Raudenbush & Bryk, 2002). DSEM's latent person-mean centering has been shown to eliminate both Nickell's and Lüdtke's bias in simulation studies (Asparouhov & Muthén, 2019) and found to be useful when the number of time-points per person is at least 10 (Gistelinck, Loeys, & Flamant, 2021; Hamaker & Grasman, 2015). In this way, DSEM provides unbiased within-person estimates that are more substantively interpretable relative to uncentered or grand mean centered models along with an intercept that can be interpreted as the true person-mean and therefore more usefully modeled at the between-level (McNeish & Hamaker, 2020). Another important feature of DSEM (and time-series data) is the assumption of *stationarity* which requires that three conditions are met (Chatfield, 2003). First, the expected value of the outcome must be constant for any arbitrary time t (i.e., the series must be mean-reverting and not systematically change or show growth between the first and last measurement occasions; Stroe-Kunold et al., 2012). Second, the variance of the outcome is constant for any arbitrary time t, and third, the autocorrelation must be the same for each same-size interval (e.g., between every measurement in a $t - 1$ or lag-1 model) across the time series. In the following sections, I expand upon the features of DSEM already presented and outline others that provide advantages over traditional modeling frameworks (MLM, SEM, and time-series analyses) with respect to the analysis of ILDs.

DSEM benefits over MLM. The primary advantages DSEM has over multilevel modeling relate to multivariate outcomes, the estimation method's accommodation of missing data, the inclusion of latent variables (both from a modeling standpoint and with respect to centering), and the inclusion of different variable types at different levels of the model. Related to the first point, DSEM is an inherently multivariate framework. This allows researchers who are interested in behavioral and relationship processes to model multiple outcomes simultaneously along with any associated interdependence.

Second, conditional likelihood estimation from MLM is replaced with either joint likelihood or (more often) Bayes methods in DSEM (Asparouhov et al., 2018), so missing values on the predictors do not trigger listwise deletion of the entire individual for the subsequent time point as in MLMs. When applied in the M*plus* software, DSEM also has a built-in Kalman filter in the estimation procedure (e.g., Hamaker & Grasman, 2012; Kim & Nelson, 1999), meaning that repeated measures taken at uneven intervals are naturally accommodated, unlike both the MLM and SEM frameworks where this unequal spacing can be problematic. As a state estimator, the general idea of a Kalman filter is that researchers specify the largest interval in which only one observation can occur, and then any intervening values without observed data are automatically inserted at each multiple of that interval. Then, assuming data are missing at random, these values are predicted from other time-points and parameters in the model (similar to imputation strategies for missing data; Hamaker et al., 2018). This preserves the interpretation of autoregressive parameters by imposing constant spacing between repeated measures, whether or not there are observed data for all times at which repeated measures could potentially be collected.

DSEM also does not require that variables be observed variables; latent variables and measurement models can be incorporated for any variable in the model and invariance of these latent variables over time can be explicitly tested (McNeish et al., 2021). As mentioned earlier, any parameter that is estimated to have a random effect in the within-model is modeled as a latent variable at the between-model. This means that the outcome means (the intercepts for the equations) are treated as *latent means* and allows for latent-mean centering for predictors. Centering is a crucial aspect of MLMs in order to disaggregate effects into between-level and within-level sources (Curran & Bauer, 2011).

Finally, DSEM has fewer constraints on what types of relations can be modeled at the between-level relative to traditional MLMs. As already mentioned, MLMs allow only for observed variables at any level of the model, but particularly require time-invariant variables to be observed variables at the between-level. DSEM, however, allows for path models, and even measurement models at the between-level, and latent variables can serve as predictors or outcomes at any level (McNeish & Hamaker, 2020).

DSEM benefits over SEM. The primary benefits of DSEM over a standard SEM relate to data structure and the accommodation of random effects. First, DSEM operates on data in the long format so extending the model to data with many repeated measures is easily incorporated into the data structure (McNeish & Hamaker, 2020). Another advantage of the long format in DSEM is that time is an explicit variable in the data, which facilitates modeling mean and variance trends over time to satisfy stationarity assumptions. The long format also makes the model explicitly multilevel and facilitates any parameter being modeled as random, which is not permissible in the SEM framework which can only accommodate random intercepts (Gistelinck & Loeys, 2019).

In DSEM, any path can be modeled as random including intercepts, slopes, or even the residual variances to model differences in volatility across individuals with a location-scale model (Hedeker et al., 2008; 2012; Jongerling et al., 2015). The random effect on the residual variance can be particularly interesting to process-oriented researchers as it allows for the investigation of *volatility* (or, on the other hand, *consistency*) in the moment-to-moment dynamics of their variables of interest (Rast, Hofer, & Sparks, 2012).

DSEM benefits over traditional time-series analyses. The main advantages DSEM has over traditional time-series analyses is that traditionally, time-series analyses have been single-sample designs; DSEM, on the other hand, is able to use these methods and accommodate an N > 1 design. DSEM is described as a multilevel extension of time series analyses, which allow the traditionally single-sample methods used to analyze time series data to be applied to samples of individuals (Asparouhov et al., 2018). Time series analyses have been used across multiple disciplines (e.g., econometrics, engineering) to model stable processes from longitudinal datasets with many time points collected closely together, where current outcomes are regressed on past outcomes (McNeish & Hamaker, 2020; Asparouhov & Muthén, 2020). Popular time-series methods include *state-space modeling*, which specifies the relationship between the series of observations and the latent variables (called *states*) that capture how the series develops over time (Durbin & Koopman, 2012; p.1; p.10) and *dynamic factor analysis*, which incorporates factor analysis with time series data (e.g., Zhang, Hamaker, & Nesselroade, 2008). Zhou, Wang, & Zhang (2021) note that ideally, ILD methods would combine both time-series and between-level processes instead of separately examining either time series or growth models (p. 224). DSEM integrates both the state-space modeling and dynamic factor analysis time series techniques, along with multilevel extensions that correspond to the two-level and cross-classified DSEM models discussed above (Hamaker et al., 2018; Asparouhov et al., 2018) which allow for individual-specific and time-specific random effects.

By building off traditional time-series models and longitudinal growth/panel models, DSEM is also able to accommodate both stable and developmental processes.

While the general DSEM model is optimized for examinations of moment-to-moment dynamics thanks to the time series aspects, psychology is often interested in change or growth over time; these trends are problematic for DSEM models as they violate the stationarity assumption discussed above that is necessary for time series analyses (McNeish & Hamaker, 2020; Asparouhov & Muthén, 2020). *Residual* DSEM (RDSEM) is an extension of DSEM where the trend and autoregressive aspects of the data are modeled separately. This involves adding time as an explicit variable in the model (to account for changes across the time series) and the autoregressive path is modeled on the within-level *residuals* rather than on the outcome directly. In this way, the time coefficient models the growth or trend in the series, and the autoregressive moment-to-moment dynamics are then captured in the residuals after that trend is accounted for (called *detrending* the series; McNeish & Hamaker, 2020, p.19). This method has been shown to more accurately recover both the coefficients and variance terms in ILD compared to adding in time as an explicit variable but keeping the rest of the DSEM model the same (i.e., keeping the autoregressive coefficient in the within-level equation as opposed to modeling it in the residual equation; Asparouhov et al., 2018).

Summary. This section covered a basic overview of the DSEM framework, with an emphasis on two-level DSEM for ILDs. It was compared to the three modeling frameworks that contribute to the features of DSEM: MLM, SEM, and time series analyses. In the following chapters, additional issues that must be considered when conducting an analysis using DSEM with dyadic data and categorical outcomes are presented.

CHAPTER 2

SPECIAL CONSIDERATIONS FOR DSEM WITH DYADIC DATA

Dyadic datasets are particularly well suited for ILDs as they allow researchers to examine changes in relationship dynamics over the course of several hours, days, or weeks; a popular method for intensive longitudinal dyadic designs is the daily diary method, which involves collecting once-daily assessments from both partners in a dyad (Laurenceau & Bolger, 2012; Iida et al., 2012; Bolger et al., 2003). Dyadic data does provide challenges for modeling ILDs, however, and considerations are needed to select the appropriate method to address to the nonindependence of people within dyads in addition to dependence attributable to repeated measures (Gistelinck & Loeys, 2019).

Despite being aware of the importance of handling the multiple sources of dependence present in dyadic data, researchers face challenges with respect to implementation. Empirical analyses of longitudinal dyadic data often make simplifications to reduce sources of dependence such as averaging over dyad members or modeling each dyad member separately (which is reported in over 80% of empirical longitudinal dyadic studies based on the review in Planalp et al., 2017). Both practices limit the ability to answer research questions and can misrepresent the dyadic relationship. For instance, a dyad where each partner scores at opposite extremes and a dyad where each partner scores in the middle of a scale look the same when averaged, despite representing different situations (Cook & Kenny, 2005). This failure to capitalize on the information available in dyadic data underscores the need for methods that avoid these simplifications such that the model serves the needs of the data structure and research questions rather than the restrictions of modeling frameworks driving the analysis.

Duncan et al. (1984) describe the concept of *pseudounilaterality* as situations where interdependent tasks are measured using only individual data (and omitting the partner). Ignoring interdependence, however, can introduce considerable bias in between- and within-level variances and affect significance tests (Kenny & Judd, 1986; Kenny, Kashy, & Cook, 2006; Laurenceau & Bolger, 2012). Gonzalez and Griffin (2012) write "Interdependence is not a problem with the data but is a limitation of the standard statistical techniques psychologists tend to use" (pg. 450), pointing out that the issue lies not with the researcher interested in dyadic processes, but with the *methods* that preclude them from analyzing their data in a way that optimally addresses their research questions. Because of the statistical and theoretical considerations outlined above, statistical techniques have been developed to acknowledge interdependence instead of treating it as a limitation or nuisance.

Intensive Longitudinal Dyadic Data: 2- or 3-Level Models?

ILD from dyads conceptually involve time points nested within people nested within dyads. While this conceptual framework appears to lend itself nicely to a 3-level model specification (e.g., Figure 2a), there are methodological issues to consider that preclude a 3-level model from being the best option to fit to longitudinal dyadic data (Laurenceau & Bolger, 2012; Kenny & Kashy, 2011; Gistelinck & Loeys, 2019). First, multilevel models assume random variability at each level which, when considering models with *distinguishable* dyads, would not be the case at level 2 – once the role of the first dyad member is known, the role of the second is known as well (the model is saturated; Diggle, et al., 2002, p.65). Second, individual time points within each person are assumed to be independent from observations at the same time point for other people; the interdependent nature of dyads not only violates this assumption in most cases, but this correlation is in fact a feature of the data that researchers are interested in capturing (Kenny & Kashy, 2011). Finally, in the three-level case the correlation between each dyad member is equal to the variance of the random effect on the intercept (an effect commonly estimated in dyadic multilevel models; Kenny, Kashy, & Cook, 2006; Chapter 4), constraining the correlation to be positive between dyad members. Though it is not frequently the case, nonindependence can be negative, such as when assessing zero-sum variables (e.g., percentage of a conversation each partner spends talking), which is not permissible as a random effect variance value and leads to estimation issues (see Kenny, Kashy, & Cook, 2006, p. 26 for more examples of negative nonindependence).

Because of the issues outlined above, methodologists suggest that longitudinal dyadic models should be specified as two-level models as shown in Figure 2b (Laurenceau & Bolger, 2012 p. 412; Bolger & Laurenceau, 2013, pp.147-149). This is possible when the observations for each dyad member occur at the same time (Day 1 for person 1 is the same as Day 1 for person 2), which means that within-person observations can be thought of as being *crossed* rather than nested, which reduces the hierarchy to two levels. Level-1 then includes a multivariate system of timepoints for both dyad members (allowing them to covary to acknowledge interdependence) and Level-2 includes *both* between-dyad effects and person-specific random effects on intercepts or slope terms (allowing for variability at Level 2). In this way, stable sources of non-independence are now captured by the random intercept terms and within-dyad non-independence can be captured by residual correlations at each timepoint (Bolger & Shrout, 2007; Bolger & Laurenceau, 2013; Gistelinck & Loeys, 2019).

In DSEM, any path can be modeled as random including intercepts, slopes, or even the residual variances to model differences in volatility across individuals with a location-scale model (Hedeker et al., 2008; 2012; Jongerling et al., 2015). This longitudinal form maps onto a lag-1 vector autoregressive model (VAR-1), which is typically used to model moment-to-moment dynamics of two variables simultaneously (e.g., Bringmann et al., 2018; Chen et al., 2020; Savord et al., 2022). The random effect on the residual variance can be particularly interesting to process-oriented researchers as it allows for the investigation of *volatility* (or, on the other hand, *consistency*) in the moment-to-moment dynamics of their variables of interest (Rast, Hofer, & Sparks, 2012).

The Longitudinal Actor-Partner Interdependence Model

One of the most popular models for dyadic data is the Actor-Partner Interdependence Model (APIM; Kenny, 2018). The APIM's popularity is evident given the flexibility of its design: when collecting the same set of variables for each dyad member, researchers can specify a *multivariate* model and avoid having to simplify their dyadic data for the purposes of accommodating a single outcome (e.g., summing together the scores from each dyad member to have a single outcome; Cook & Kenny, 2005). The APIM uses mixed variable outcomes (meaning they can vary both within and between dyads) denoted Y1 and Y2 in Figure 3, predicted by *actor* effects (a person's predictor on their own outcome) and *partner* effects (a person's predictor on their partner's outcome) for predictors denoted X1 and X2. Correlations between the two predictors (one from each dyad member) and the residuals of each outcome are also estimated to acknowledge the interdependence of observations (Kashy & Kenny, 2000; Kashy & Snyder, 1995; Kenny, 1996; Kenny & Cook, 1999). First described by Kenny (1996), the APIM was developed as a cross-sectional or two time point model. A brief introduction of the APIM using a two time point example is presented below; longitudinal extensions will be covered in more detail in subsequent sections.

The multivariate specification for the APIM allows researchers to estimate individual effects (i.e., for both partners) while preserving the *dyad* as the level of analysis for the model through the addition of correlated predictors and correlated residuals (the double-headed arrows shown in Figure 3) to specify the interdependent nature of dyad members (Kashy & Snyder, 1995; Kenny, 1996).

Similar to other types of nested data, the APIM has traditionally been fit as either an MLM or SEM. Despite different origins and minor differences of these frameworks (e.g., Curran, 2003), each provides the same results for traditional cross-sectional APIMs (Wendorf, 2002). Nonetheless, there are some distinctions that differentiate MLM and SEM in cross-sectional designs and especially in longitudinal designs (Ledermann & Kenny, 2017). The following sections provide a brief overview of those distinctions and discusses the and limitations of the L-APIM in the MLM and SEM frameworks, followed by an overview of how DSEM addresses those limitations.

L-APIM as a multilevel model. MLMs relax independence assumptions of the general(ized) linear model and are designed to explicitly model dependence due to context, with classic examples being children clustered within classrooms or repeated measures clustered within an individual (e.g., Hox et al., 2017; Raudenbush & Bryk, 2002). As previous discussed, dyadic data involves people clustered within dyads, with timepoints conceptually clustered within people for longitudinal designs. Most standard software implementations of MLMs (e.g., SPSS MIXED, lme4 in R, SAS Proc Mixed) are *univariate* extensions of the general linear model such that only one outcome can be modeled. This means that data are structured in the *long* format such that there is a single column corresponding to the observed outcome variable (latent variables are not permitted as outcomes or predictors in standard MLM software) and observations are identified by index variables signifying cluster membership and within-cluster identifiers (e.g., dyad ID number for cross-sectional data and/or day of data collection for daily diary data). Hypothetical dyadic data in the long format would be structured as

Dyad	Partner	Y	X
1	0	5	2
1	1	7	4
2	0	6	3
2	1	8	1

Where the Dyad column signifies cluster membership and the Partner column identifies the dyad member (assuming dyad members are *distinguishable*[1], meaning they can be differentiated based on an identifying variable such as sex or role; Kenny, Kashy, & Cook, 2006). This long-format structure is referred to as the *dyad data structure* (Ledermann & Kenny, 2015; 2017; Kenny, Kashy, & Cook, 2006).

One hurdle with the APIM in the MLM framework is that the model is inherently multivariate (i.e., there are two outcomes, one for each dyad member), but most implementations of MLMs are restricted to be univariate (i.e., there can only be a single

[1] Indistinguishable dyads are also appropriate for an APIM, see Ledermann & Kenny (2017) and Kenny et al. (2006; Chapter 7) for a discussion on the pros and cons of using MLM or SEM in this case. The DSEM framework can also accommodate indistinguishable dyads, though is not the focus of this project.

outcome variable in the model). In the hypothetical data above, it is easy to see how to incorporate actor effects because all data from the same person exist within the same row (Laurenceau & Bolger, 2005). However, with the dyad data structure the variables involved in partner effects are located in different rows making it unclear how to estimate these effects. Therefore, the data are typically restructured to a *pairwise data structure* before being suitable for the MLM framework (Ledermann & Kenny, 2015).

The pairwise structure reformats the data such that extra columns are added to include data corresponding to the other partner from the same dyad, so each person's observations show up as the "actor" variables for their own row of data and the "partner" variables for their partner's row of data. The same hypothetical data above in a pairwise structure is shown below. The ActorX "4" value in the second row becomes the value in the PartnerX column of the first row. Similarly, the ActorX "2" value in the first row becomes the value for the PartnerX column in the second row. This is similar to *stacked* data structures used in multilevel mediation models (e.g., Bauer, Preacher, & Gil, 2006), which is another model that is inherently multivariate but coerced into a univariate MLM framework. Additionally, there are multiple columns that act as indexes for either the distinguishing variable ("Dist. Var") or which partner is coded as the "actor" for the row ("Partner1" and "Partner2"). The utility of these variables and when to use which columns when fitting the APIM as an MLM is discussed below.

Dyad	Dist. Var	Partner1	Partner2	Y	ActorX	PartnerX
1	-1	0	1	5	2	4
1	1	1	0	7	4	2
2	-1	0	1	6	3	1
2	1	1	0	8	1	3

A benefit of structuring data in the long format is that additional timepoints are straightforward to include by simply adding more rows per dyad member followed by the addition of variable that codes the time at which data were collected. Given that the pairwise structure is already complicated compared to a univariate wide-format data structure, data with many repeated measures such as those coming from ILDs may become more burdensome to manage from a logistical standpoint. Ledermann and Kenny (2015), however, created a toolbox that allows researchers to input their data and have it restructured to pairwise or dyad structure by the program, so this does not necessarily affect the ability to fit the model if certain conditions are met. The potential pitfalls of extending the MLM to longitudinal designs are whether these conditions, discussed below, are satisfactorily met.

To reframe Figure 3 as a repeated measures design, the actor effects from X to Y can be thought of as autoregressions, where the previous repeated measure (at time t-1) directly predicts the next repeated measure (at time t). Autoregressive models are popular to use with ILDs to model stability and carryover over time, provided that the stationarity assumption is met so there is no unmodeled systematic trend in the mean or variance over time (e.g., any cyclical, predictable changes in mean and variance over time are accounted for in the model). In this way, autoregressive models are a natural way to extend the APIM to many repeated measures in the MLM framework, though have three noted complications.

The first limitation of autoregressive models fit in the MLM framework is that the autoregressive parameters such as the X to Y actor effects are typically assumed to be constant across time (Gollob & Reichardt, 1987). This can cause issues when the spacing of the repeated measures changes across the window of observations. Repeated measures that are closer together in time tend to be more related to one another, but this distinction is lost in MLMs when there is uneven spacing (Ryan et al., 2018; Voelkle et al., 2012). Though some daily diary designs may not encounter this issue because data are consistently collected one day apart, other ILDs like EMA or experience sampling purposefully collect data at unequal intervals to prevent participants from anticipating the next measurement occasion to maximize external validity (McNeish & Hamaker, 2020).

A second limitation relates to the covariance structure. That is, when coercing multivariate data into a univariate framework when the data contain many repeated measures, there may be autocorrelation in the residuals for the same variable across time but there are *also* residual covariances between members of the same data. In a univariate model, both sources of dependence must be modeled within a single matrix. Properly structuring the covariance in such a case typically requires a Kronecker product of an unstructured matrix capturing dependence between dyad members and an autoregressive matrix capturing dependence over time (Bolger & Shrout, 2007; Galecki, 1994). Fortunately, Kronecker product covariance structures are available in two of the most mainstream MLM programs, SAS Proc Mixed and the SPSS MIXED command (Gistelinck & Loeys, 2019). Recall, however, that the first limitation related to equally spaced intervals means that autoregressive covariance structures are also only applicable when data are equally spaced; alternative structures such as a Markov structure (a.k.a. spatial power structure; Wolfinger, 1993) would be needed to preserve information about unequal spacing. Markov structures are available in some software (in SAS Proc Mixed, this possible with TYPE= SP(POW)). However, SAS does not permit Kronecker products of Markov structures in its mixed model procedures (SAS Institute, 2018, p. 6613).

The third noted limitation for the MLM framework for longitudinal APIMs is that most MLM software use conditional likelihood estimation (Bauer, 2003). This means that the predictors are (a) conditioned out of the likelihood function, (b) treated exogenously, and (c) not formally part of the model, so missing data on the predictors is not permitted (Allison, 2012). This is not necessarily a problem for longitudinal models in general but poses problems for models with missing time-varying covariates such as predicting the current outcome from the immediately preceding value in an autoregressive model (e.g., McNeish & Matta, 2020). Missing data are common in ILDs given the frequency of measurement, so the ramification in a pairwise data structure is that missing data for one dyad member's outcome (resulting in a 'single'; Ledermann & Kenny, 2017) will cause the entire dyad to be listwise deleted at the next time-point, even if the other member of the dyad has complete data. It is recommended that researchers use multiple imputation techniques in the presence of missing data in order to avoid biased parameter estimates that can result from listwise deletion (Enders, 2010; Schoemann & Sakaluk, 2016; Ledermann & Kenny, 2017).

Overall, modeling an L-APIM in the MLM framework for dyadic ILDs has the benefit of allowing for a long-format data structure, but exacerbates the limitations of MLM's univariate structure while introducing additional complications around data collection intervals, missing data, and specifying the correct covariance structure.

L-APIM as a structural equation model. SEM is a multivariate framework that structures data in the *wide* format such that each dyad only occupies a single row and each variable is a unique column. The same data shown in the previous section on MLMs can be structured in the wide format as follows,

Dyad	Y1	X1	Y2	X2
1	5	2	7	4
2	6	3	8	1

There is no longer an index variable for which dyad member the data correspond; all data for a dyad appear on the same row. Each outcome and each predictor for each dyad member are instead represented by separate columns in the data. SEM then allows the fitted model to have multiple equations, meaning that the path diagram for the distinguishable APIM in the SEM framework is identical to the model presented in Figure 3 and there is no need to restructure the data into a pairwise format. SEM has advantages over MLM including accommodation of latent variables as outcomes or predictors and joint likelihood estimation, which allows for missing data on any variable in the model without listwise deletion (Bauer, 2003).

When extending to longitudinal data, additional repeated measures can be included by adding additional columns to the data. The multiple sources of dependence can be accounted for by covarying the residuals of the same outcome across time (which could be constrained to create an autoregressive structure if one wished to match the MLM implementation) and covarying the residuals of different outcomes at the same time if necessary. Figure 4 shows the SEM path diagram of an APIM extended to 3 repeated measures. This path diagram assumes a single variable is of interest, but the model can be expanded to include repeatedly measured *X* and *Y* variables (Gistelinck & Loeys, 2019). The SEM framework does not possess the missing data weakness present with MLMs because the joint likelihood is used for estimation (by default in lavaan and by request in M*plus*), but other considerations exist for fitting L-APIMs from ILDs in the SEM framework.

First, because data are in the wide format, measures are assumed to be time-structured such that all people within dyads are measured at the same time points, and that all observations are collected at equal intervals (e.g., once per day; Hamaker et al., 2018). This will not likely be problematic for daily diary studies where all people typically respond each day but will be an issue for designs that incorporate some degree of randomness into the timing of measurements to maximize ecological validity such as EMA or experience sampling (e.g., Kuiper & Ryan, 2018). Use of the wide format also assumes that everyone is measured at equal time points, which is also not necessarily true of ILDs other than daily diary. When dyads are measured at irregular intervals, or certain time points are missing for certain dyads, however, 'phantom' latent variables can be included in the model as placeholders for those missing timepoints to impose an equal-timepoint structure (Rindskopf, 1984; e.g., Voelkle & Oud, 2015).

Second, the wide format tends to encounter estimation issues once the number of timepoints exceeds about 10 because the dimensions of the sample and model-implied mean and covariance structures become quite large (Asparouhov & Muthén, 2015, p. 182) or hundreds of parameter constraints may be necessary (Ledermann & Kenny, 2017 p. 448). Many intensive longitudinal designs contain many more measures (e.g., Collins, 2006 defines "intensive longitudinal" as roughly 20 or more repeated measures) and may have estimation issues if fit in the SEM framework. Gistelick and Loeys (2019) note similar issues (p. 335) but propose a method that was able to converge with data containing 21 repeated measures. Their method involves avoiding the excessive number of constraints necessary to fit the appropriate Kronecker product covariance structure discussed above by decomposing the effects into a stable trait, an autoregressive trait, and a state. They note, however, that their method cannot yet handle latent outcomes, latent decomposition, or indistinguishability tests (p. 342).

Third, SEM treats the model multivariately and accounts for dependence with covariances rather than fitting a multilevel model where people are nested within dyads (treating data multivariately reduces the data hierarchy by one level relative to a MLM, so the SEM framework treats dyadic data as a single level model; Muthén, 1997). Adding features like random intercepts can be accommodated with a model similar to the random intercepts cross-lagged panel model (Hamaker, Kuiper, & Grasman, 2015) but can introduce complications with many repeated measures. As a further complication, there is currently no way to incorporate random slopes in the wide SEM framework (Gistelinck & Loeys, 2019, p. 342). This is because with a wide-format SEM, the slopes for actor and partner effects are not variables themselves (contrasted with MLM, where slopes can be treated as outcomes at the between-level); as a result, a random effect of slopes across dyads cannot be modeled. Relatedly, as with other multivariate models for longitudinal data, time is represented in the data by the number and ordering of columns and not by an explicit variable. Modeling trends over time is therefore less straightforward in the SEM framework and involves embedding the APIM into a latent growth model (Ledermann & Macho, 2014) or possibly using a structured residual model (Curran et al., 2014).

L-APIM as a dynamic structural equation model. Specific to fitting the L-APIM to ILDs, the MLM aspects of DSEM allow data to be clustered within dyads (i.e., data are in the long format), the SEM aspects allow the model to be multivariate (i.e., different variables are different columns in the data and a pairwise data structure is unnecessary), and the time-series aspects allow for lagged relations to model moment-to-moment dynamics over time.

Figure 5 shows the conceptual one-variable longitudinal APIM with random intercepts and slopes in the DSEM framework. The within-person model for Figure 5 looks similar to Figure 3 but generalizes the idea to an arbitrary number of repeated measures; also because of random effects at the within-level, Figure 5 includes a latent decomposition and between-level model as well. The L-APIM can be viewed as a variation of the VAR-1 model such that the two time-series correspond to two members of a dyad rather than two variables. The outcome for Partner 1 at time t -1 predicts the outcome for Partner 1 (actor effect; φ_{1d}) and Partner 2 (partner effect; φ_{4d}) at subsequent time point t. Similarly, the outcome for Partner 2 at time t -1 predicts the outcome for Partner 2 (actor effect; φ_{2d}) and Partner 1 (partner effect; φ_{3d}) at subsequent time point t. These four effects are allowed to vary across dyads, so they have a d subscript and appear in the between-level model as well. Similarly, the intercept terms for Partner 1 (α_{1d}) and Partner 2 (α_{2d}) are allowed to vary across dyads (i.e., each dyad member gets a unique estimate for their person-mean, which is made more accurate thanks to the latent-mean centering of the predictors) and as such also have a d subscript and appear in the between-level model.

The latent decomposition along with within- and between-level model can be expressed as a system of equations as:

$$
\text{Latent Decomposition}\begin{cases} y1_{td} = y1_{td}^{w} + \alpha_{1d} \\ y2_{td} = y2_{td}^{w} + \alpha_{2d} \end{cases}
$$

$$
\text{Within-Level}\begin{cases} y1_{td}^{w} = \varphi_{1d}\, y1_{(t-1)d}^{w} + \varphi_{4d}\, y2_{(t-1)d}^{w} + e_{1td} \\ y2_{td}^{w} = \varphi_{2d}\, y2_{(t-1)d}^{w} + \varphi_{3d}\, y1_{(t-1)d}^{w} + e_{2td} \\ \begin{bmatrix} e_{1td} \\ e_{2td} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right) \end{cases}
$$

$$
\text{Between-Level}\begin{cases} \alpha_{1d} = \gamma_{10} + u_{1d} \\ \alpha_{2d} = \gamma_{20} + u_{2d} \\ \varphi_{1d} = \gamma_{30} + u_{3d} \\ \varphi_{2d} = \gamma_{40} + u_{4d} \\ \varphi_{3d} = \gamma_{50} + u_{5d} \\ \varphi_{4d} = \gamma_{60} + u_{6d} \\ \mathbf{u}_d \sim MVN\left(\mathbf{0}_6, diag\left[\tau_{11} \ldots \tau_{66}\right]\right) \end{cases} \tag{2}
$$

With actor, partner, and intercept effects as specified above. Notice, too, that each predictor has a *w* superscript denoting that it has been latent-mean centered; because the inclusion of a random effect for that variable's coefficient places that coefficient in the between-level model, it gets modeled as a latent variable to be estimated (Asparouhov & Muthén, 2019). The error terms are assumed to follow a normal distribution, with unique variances and a shared covariance.

Where each random effect specified at the within-level model appears as an outcome at the between-level, predicted by a fixed effect (e.g., γ_{10} for the intercept for Partner 1) and a random effect (e.g., u_{1d} for the intercept for Partner 1). The random effect terms follow a multivariate normal distribution each with a unique variance (e.g., τ_{11} for the intercept for Partner 1). In this model, no covariances between random effects are specified, though there may be theoretical reasons to include a covariance between the intercept random effect terms, for example, to capture any relationship between the random effect for a Partner 1's latent person-mean and the random effect for Partner 2's latent person-mean.

Taken together, consolidating aspects of MLM, SEM, and time-series analysis into a single framework facilitates longitudinal APIM because noted weaknesses of MLM and SEM are avoided. Because DSEM can accommodate multivariate outcomes, the model in Figure 5 can be fit without any data structure manipulation, and weaknesses introduced by a pairwise data structure in MLMs are removed. Additionally, having time explicitly in the dataset in the long format facilitates satisfying the stationarity assumption without having to embed the L-APIM into a latent growth model as one would have to do if working in the SEM framework. The ability to model random effects on the residual variances opens up new potential for dyadic researchers to examine the interdependence across their ILDs – does residual variance (volatility) for each dyad member vary across dyads? Does this volatility covary with their partner's volatility, or their intercept (person-mean) parameter? Furthermore, DSEM subsumes MLM and SEM as special cases. So even in situations where MLM or SEM may be suitable for fitting an L-APIM from an ILD (e.g., relatively few equally spaced repeated measures with no missing data and few random effects), MLM and SEM provide relatively little tangible benefit over DSEM aside from being able to be fit in more general software programs.

Summary

DSEM can be used when researchers with dyadic ILDs have research questions traditionally answered using an APIM. By including aspects of MLM, SEM, and time-series models, along with Bayesian estimation methods, DSEM can act as a more flexible and general framework which consequently allows researchers to answer more nuanced questions with their datasets. By allowing for multivariate outcomes and a long-format data structure, adding additional variables (predictors or outcomes) or timepoints to a model involves less data manipulation. The Bayesian estimation methods and Kalman filter to handle missingness allows for more complex models to be fit to smaller datasets with fewer convergence issues. There may be situations where methods besides DSEM are necessary to analyze particular data structures, such as those where the stationarity assumption may not be appropriate (see Bringmann et al., 2018 for an example on emotion dynamics and time-varying vector-autoregressive models). Overall, DSEM is a flexible method well-equipped to meet the need for methodological options when researchers have dyadic intensive longitudinal data.

CHAPTER 3

SPECIAL CONSIDERATIONS FOR DSEM WITH CATEGORICAL OUTCOMES

While some forms of intensive longitudinal data collection can be passive, such as physiological data gathered by smartwatches that might be used for EMAs (e.g., Kop et al., 2001), most involve participants' direct self-reports to items such as the ones above. Gunthert & Wenze (2014) argue that daily diary methods strike a balance between less frequent measurement occasions (which can introduce higher recall bias) and more frequent measurement occasions (which can be costlier and result in higher attrition rates), resulting in a dataset that has more ecological validity than less frequent assessments but higher compliance rates for self-report than more frequent assessments. Furthermore, when considering the measures for daily diary-style ILD collection, it is often the case that items with categorical responses are used, with items such as "Did a symptom occur today?" prompting a binary response, or "How stressed were you today?" prompting a response from a list of several ordered categories (see Bolger & Laurenceau, 2013, Chapter 6). The focus of this dissertation project is on daily diary-style ILDs that contain categorical item responses.

Modeling ILDs with categorical outcomes in the DSEM framework in M*plus* involves a generalized linear model using a probit link function. This link function serves to maintain a linear relationship between predictors and the outcome through a standard normal cumulative distribution function (typically denoted as $\Phi(\cdot)$). The probit model assumes, then, that there is an underlying normal distribution for the variable of interest that is observed as a collection of categorical responses (Agresti, 2012). For example, to revisit the question above, "How stressed were you today?", one can imagine a situation where participants are asked to respond to that question on a 5-category scale ranging from 0 (not stressed at all) to 4 (extremely stressed). While not everyone who assigns themselves a value of “3” have the *exact* same level of stress, one may assume they fall within the same *range* of values on the standard normal distribution (bounded above by those who respond at “4” and below by those who respond at “2”).

As with many of our beloved methodologies, the probit model was developed to improve agricultural crop yield. Toxicologist Bliss (1935), who surely never felt his name ironic, proposed the probit model as a way to assess the status of pests poisoned by increasing doses of pesticide using a binary variable (alive or dead); the effects of the pesticide follow a continuous process, but these effects were imperceptible until a threshold was passed (in this case, manifesting as the death of the insect). And thus, the futures of grapevines and categorical outcome models were permanently improved just as the life of a certain grape leaf-loving pest was being meticulously reduced from a continuum of experiences to a cold binary truth: alive or dead.

While the idea of using a probit link function may seem daunting, its use of the cumulative normal distribution puts the outcome on the Z-score metric which can be easily converted into a probability by taking the area under a Z-distribution to the left of the Z-score. When using DSEM with a probit link, one can re-imagine the model specified in equation 1 above as though y_{ti} is a binary outcome (with covariate x_{ti} as continuous). The relationship between the binary variable y_{ti} and its underlying continuous process y^*_{ti} can then be specified as

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

And the subsequent system of equations would then be:

$$
\begin{array}{ll}
\text{Latent Decomposition} & \begin{cases} \Pr(y_{ti}=1)=\Phi(y_{ti}^{*}) \\ y_{ti}^{*}=y_{ti}^{*w}+\alpha_i \\ x_{ti}=x_{ti}^{w}+x_i^{b} \end{cases} \\
\text{Within-Person} & \begin{cases} y_{ti}^{*w}=\phi_i y_{(t-1)i}^{*w}+\beta_i x_{ti}^{w}+e_{ti} \\ x_{ti}^{w}\sim N(0,\omega) \\ e_{ti}\sim N(0,1) \end{cases} \\
\text{Between-Person} & \begin{cases} \alpha_i=-\tau_1+u_{0i} \\ \phi_i=\gamma_{10}+u_{1i} \\ \beta_i=\gamma_{20}+u_{2i} \\ x_i^{b}=\gamma_{30}+u_{3i} \\ \mathbf{u}_i\sim MVN(\mathbf{0}_4, diag[\sigma_{00},\sigma_{11},\sigma_{22},\sigma_{33}]) \end{cases}
\end{array}
\tag{4}
$$

Notice how, while the majority of the model looks similar to equation 1, there are three differences. First, the additional expression at the top: $\Pr(y_{ti}=1)=\Phi(y_{ti}^{*})$. This expression can be read as "the probability of y_{ti} taking on a value of 1 is equal to the probability to the left of the Z-score specified by y_{ti}^{*}," where y_{ti}^{*} represents the underlying continuous process and is defined in the subsequent equations similar to before. Second, the autoregressive path that was denoted as φ_i in equation1 is now ϕ_i, which represents a polychoric (more specifically tetrachoric for binary variables) autocorrelation and is more appropriate to use when modeling the underlying continuous processes of two categorical variables (Asparouhov & Muthén, 2019). Third, the residual error term e_{ti} now has a fixed variance of 1 instead of an estimated variance of σ^2. This is standard in two-level probit regression for identifiability (Asparouhov & Muthén, 2019). By parameterizing the model in this way – that is, with y_{ti}^{*} directly modeling the underlying continuous process, instead of embedding the second expression above entirely within the parenthetical as $\Pr(y_{ti} = 1) = \Phi(\alpha_i + \phi_i y_{(t-1)i} + \beta_i x_{ti})$ – the model can be estimated more quickly in M*plus* and requires fewer timepoints to yield reliable estimates (can have as few as 20 observations per person; Asparouhov & Muthén, 2019, pg. 136).

The within-level equations are only half the story, though. The between-level equations for this model have two major differences from what was presented above in equation 1. First, because the autoregressive effect at the within-level is denoted as ϕ_i, it changes to ϕ_i here as well. Second (and more majorly), the fixed effect for the intercept equation that was previously γ_{00} is now $-\tau_0$. While still representing a fixed effect, this term is the result of how M*plus* parameterizes probit models. Specifically, M*plus* uses a latent intercept (α_i) but defines the fixed effect in terms of a threshold (τ_1), which represents the probability that the outcome is 0 (instead of 1 as is specified in the latent decomposition at the top of equation 4). This means that the fixed effect for the intercept is equal to the opposite of the threshold (i.e., $-\tau_1$) because the outcome being modeled is the probability that y_{ti} = 1 instead of y_{ti} = 0 (see Muthén & Muthén, 1998-2017, pg. 552).

A path diagram for the model outlined in equation 4 above is shown in Figure 6. In this figure, observed variables are represented with rectangles, latent variables are represented with circles, and constants are represented with triangles. The left-hand panel shows both the latent decomposition of y_{ti}^* and x_{ti} into within- and between-level components, along with the relationship between the binary observed y_{ti} and the assumed underlying continuous process y^*_{ti} (represented by the wavy line; de Boeck & Wilson, 2004). The right-hand panel depicts the within and between level models for y^*_{ti} with effects corresponding to the equations presented above. Apart from the changes described above, the terms in the within- and between-level model can largely be interpreted the same way as described for equation 1, with the caveat that they represent relationships with the *underlying continuous process* that manifests as the categorical outcome. Variables with random effects still get latent person-mean centered to provide more interpretable effects that are free from Nickell's and Lüdtke's bias, and effects can be partitioned into within- and between-level components. The other desirable DSEM features, such as the Kalman filter to handle missing observations, is also present in the categorical outcome model.

The model outlined here is for a univariate DSEM with a binary outcome, but can be generalized to polytomous items with *C* categories and *C*-1 thresholds. For example, revisiting the scenario above of asking people to rate their stress on a 5-category scale, those five groups of people might fall on an underlying normal distribution of stress levels as displayed in Figure 7. These groups are bounded by Z-score threshold values that characterize which portion of the distribution of stress scores is predicted to manifest as a particular category – for 5 categories on the outcome, there would be 4 threshold values (in Figure 7 shown at Z scores of -1.5, -0.5, 0.5, and 1.5) to partition the underlying distribution, just as how in the case of binary outcomes there was only 1 threshold value. These threshold values provide an idea of what proportion of the sample selected each potential categorical response based on the density of the standard normal distribution. For example, the probability of someone selecting 0 as their stress score

("Not stressed at all") would be approximately 0.067 as this is the area to the left of -1.5 on a Z-distribution. The probability of someone selecting 1 as their stress score would be approximately 0.242, as this is the area to the left of -0.5 *minus* the area to the left of -1.5, leaving just the slice of the distribution corresponding to the categorical response of 1.

With this in mind, equations 3 and 4 above can expand to model a 5-category outcome y_{ti} and continuous covariate x_{ti} as the following system of equations:

$$y_{ti} = \begin{cases} 4 \text{ if } \tau_4 \le y_{ti}^* < \infty \\ 3 \text{ if } \tau_3 \le y_{ti}^* < \tau_4 \\ 2 \text{ if } \tau_2 \le y_{ti}^* < \tau_3 \\ 1 \text{ if } \tau_1 \le y_{ti}^* < \tau_2 \\ 0 \text{ if } -\infty \le y_{ti}^* < \tau_1 \end{cases}$$

$$\text{Latent Decomposition} \begin{cases} \Pr(y_{ti} = 0) = \Phi(\tau_1 - y_{ti}^*) \\ \Pr(y_{ti} = 1) = \Phi(\tau_2 - y_{ti}^*) - \Phi(\tau_1 - y_{ti}^*) \\ \Pr(y_{ti} = 2) = \Phi(\tau_3 - y_{ti}^*) - \Phi(\tau_2 - y_{ti}^*) \\ \Pr(y_{ti} = 3) = \Phi(\tau_4 - y_{ti}^*) - \Phi(\tau_3 - y_{ti}^*) \\ \Pr(y_{ti} = 4) = \Phi(-\tau_4 + y_{ti}^*) \\ y_{ti}^* = y_{ti}^{*w} + \alpha_i \\ x_{ti} = x_{ti}^w + x_i^b \end{cases}$$

$$\text{Within-Person} \begin{cases} y_{ti}^{*w} = \phi_i y_{(t-1)i}^w + \beta_i x_{ti}^w + e_{ti} \\ x_{ti}^w \sim N(0, \omega) \\ e_{ti} \sim N(0,1) \end{cases}$$

$$\text{Between-Person} \begin{cases} \alpha_i = u_{0i} \\ \phi_i = \gamma_{10} + u_{1i} \\ \beta_i = \gamma_{20} + u_{2i} \\ x_i^b = \gamma_{30} + u_{3i} \\ \mathbf{u}_i \sim MVN(\mathbf{0}_4, diag[\sigma_{00}, \sigma_{11}, \sigma_{22}, \sigma_{33}]) \end{cases} \quad (5)$$

Here, the major differences between equation 5 and equations 3 and 4 above are how the 5-category outcome y_{ti} relates to the continuous underlying process y_{ti}^{*} (equation 4 compared to the top of equation 5), the latent decomposition of y_{ti}^{*}, and the between-level equation for the intercept α_i. Modeling the relationship between a categorical observed variable and its underlying continuous process can be more generally expressed for each category C as $y_{ti} = \{C \text{ if } \tau_{C-1} \le y_{ti}^{*} < \tau_C$, where $\tau_0 = -\infty$ and $\tau_C = \infty$ (Long, 1997, pg. 121; Asparouhov et al., 2018, pg. 374); note that in this example, the first category gets an observed value of 0, so here C = 1 when y_{ti} = 0.

In the latent decomposition of equation 6 the probability of *each* outcome gets expressed in terms of the underlying distribution y_{ti}^{*}, whereas the latent decomposition shown in equation 6 only has to model one outcome; knowing the probability of one response automatically gives the probability of the other response in the binary setting. Additionally, in equation 5 subtraction is used to isolate just the portion of the underlying normal distribution that corresponds to a particular category. For example, looking just at $\Pr(y_{ti} = 0) = \Phi(\tau_1 - y_{ti}^{*})$, notice that it is expressing the probability that the categorical outcome is 0. Because this is the lowest category option, the probability is equal to all of the underlying continuous distribution y_{ti}^{*} that falls below τ_1 (Threshold 1 in Figure 7). Similarly, the probability that the categorical outcome is 1 is equal to taking the portion of y_{ti}^{*} that falls below τ_2 and subtracting from that the portion of y_{ti}^{*} that falls below τ_1.

Another aspect of polytomous probit models that was not discussed when presenting a binary outcome model has to do with the relationship between the intercept and threshold parameters themselves. In the binary case, the intercept parameter α_i was defined as having a fixed effect of $-\tau_1$ and random effect u_{0i}. While this specification was a helpful way to think about the relationship between the intercept and threshold values, it is no longer useful when generalizing to polytomous items. Instead, for polytomous categorical outcome models like the one specified in equation 5 above, the intercept has a fixed effect of 0 and a random effect of u_{0i} allowing for identification of the model (Long, 1997, pp. 122-123). As a result, the threshold parameters remain fixed in the same place on the underlying continuous normal distribution around the fixed effect for the intercept (i.e., around a mean of 0 as in a standard normal distribution) and are therefore not part of the latent decomposition, within-person, or between-person model. However, each *person-specific* underlying continuous distribution y_{ti}^* can shift its mean based on the intercept random effect (though its variance is still fixed at 1; Asparouhov et al., 2018, pg. 363; Long, 1997). In this way, even though the threshold values themselves do not vary from person to person, allowing the intercept random effect u_{0i} to act as a shifting parameter provides them with unique probabilities for responding in the categories provided on the observed response variable.

Along with the nuances of model specification and interpretation when faced with a polytomous probit model, researchers have differing opinions on when (if ever) categorical models should be used in lieu of their continuous counterparts. While some argue that a categorical model is not necessary at all relative to treating the outcome as continuous (see Robitzsch, 2020), others believe that there are benchmarks – say, five categories – that demarcate an upper threshold of when categorical methods should be used, or a lower threshold for when continuous methods should work well enough (see Rhemtulla et al., 2012). While the potential ease of using a more familiar continuous modeling framework is tempting, ignoring the categorical nature of the outcome could lead to biased parameter estimates and incorrect standard errors; the extent to which this is true for various-sized categorical outcomes in the DSEM framework has yet to be fully explored.

Summary

As discussed above, the extension to polytomous outcomes is especially relevant when considering the Likert-scale types of questions typically found in daily diary studies. Cook et al. (2010) report that five to seven response categories is the optimal range to obtain adequate reliability and validity of inferences from scores, which is a separate consideration from whether the variable can then be modeled as continuous rather than categorical. Furthermore, dyadic models are inherently *multivariate* and so would need to account for multiple categorical outcomes within the same model. Luckily, Asparouhov & Muthén (2019) report that the probit DSEM for categorical outcomes "easily extends to ordered categorical variables" and "can easily accommodate multivariate time-series modeling with categorical variables" as well (p. 136). The issue, however, is that the DSEM simulations assessing model performance with categorical outcomes (see Asparouhov & Muthén, 2019, pp. 135-136, and Asparouhov & Muthén, 2020, pp. 285-286) have focused only on the univariate binary outcome case.

Overall Summary

The DSEM framework as implemented in M*plus* is well-suited to accommodate intensive longitudinal datasets. Its combination of MLM, SEM, and time-series analyses make it a more general version of all three, and is as a result flexible enough to fit models with many different variable types and accompanying research questions to examine moment-to-moment changes in stable processes. This framework works particularly well for multivariate models such as those frequently used by dyadic researchers, though most simulations evaluating categorical outcomes thus far have focused on univariate binary outcomes. More work is needed to investigate how categorical outcomes might impact estimates from both multivariate and univariate DSEM models with polytomous outcomes.

CHAPTER 4

MOTIVATING DATA EXAMPLE

Description of Motivating Data

The dataset was collected from young adult heterosexual couples in the US to study the use of technology in relationships during the COVID-19 pandemic. Couples completed a baseline survey and were randomized to one of two conditions. In Condition 1, couples only completed a follow-up survey 6 weeks after the baseline survey. In Condition 2, couples completed a 21-day daily diary immediately following the baseline survey prior to completing the 6 post-baseline follow-up survey. In Condition 2, each day couples responded to questions about their relationship, emotions, and COVID-19 related feelings and behaviors, among other topics. In total, 36 couples were randomly assigned to Condition 1 and 55 were randomly assigned to Condition 2.

The data included in the example analyses focuses on the 55 couples in Condition 2 that mirror the type of ILDs that are increasingly popular with dyadic data. In this 110-person sample, the average age is 20.8 years, average relationship duration is 21.44 months, 24.5% of participants live with their romantic partner, and 54.5% are white. Couples were modeled as distinguishable dyads based on sex.

Methods

Measures. *Emotions Towards Partner.* Participants were asked 12 questions related to feelings and emotions within their relationship with their partner. Each item provided five response options for the extent to which they had experienced each feeling or emotion that day; response options were “Not at all”, “A little”, “Moderately”, “Quite a bit", and "Extremely". The item of interest from this section asks the extent to which participants felt supported by their partner that day.

Relationship Quality. Participants were asked 8 questions related to characterizing their romantic relationship over the past 24 hours. Each item provided a scale slider that ranged from 0 to 100 for participants to use. Two items were of interest from this section related to closeness: a physical intimacy item (scale ranged from 0 = not physically intimate, to 100 = physically intimate) and an emotional closeness item (scale ranged from 0 = emotionally distant, to 100 = emotionally close). These two responses were summed together and divided by 10 to create an item representing overall relationship connectedness that ranged from 0 to 20.

Analysis. The research questions for this example are based on an L-APIM of previous- and present-day assessments of the 5-category emotions item ("support"), as well as the disaggregation of within- and between-level effects of the continuous relationship connectedness covariate ("connect"): Does a person's perception of whether they felt supported by their partner from the previous day affect their own or their partner's present-day perception of feeling supported (modeled as actor and partner effects in an L-APIM)? Does a person's day-to day perception of their relationship connectedness affect their same-day perception of support (within-level connect predicting support)? How does a person's habitual level of connectedness affect their habitual level of perception of support (between-level connect predicting between-level care)? The overall goals of this analysis are to describe the effects produced by an L-APIM DSEM analysis with a polytomous outcome, and to determine plausible effects values that can be used for the simulation study described in subsequent sections of this document.

For this model, the intercepts, actor, partner, and covariate effects will be allowed to vary across dyads (i.e., random intercepts and random slopes). Statistically, the model can be written as,

$$SupportM_{td} = \begin{cases} 4 \text{ if } \tau M_4 \leq SupportM^*_{td} < \infty \\ 3 \text{ if } \tau M_3 \leq SupportM^*_{td} < \tau M_4 \\ 2 \text{ if } \tau M_2 \leq SupportM^*_{td} < \tau M_3 \\ 1 \text{ if } \tau M_1 \leq SupportM^*_{td} < \tau M_2 \\ 0 \text{ if } -\infty \leq SupportM^*_{td} < \tau M_1 \end{cases}$$

$$SupportF_{td} = \begin{cases} 4 \text{ if } \tau F_4 \leq SupportF^*_{td} < \infty \\ 3 \text{ if } \tau F_3 \leq SupportF^*_{td} < \tau F_4 \\ 2 \text{ if } \tau F_2 \leq SupportF^*_{td} < \tau F_3 \\ 1 \text{ if } \tau F_1 \leq SupportF^*_{td} < \tau F_2 \\ 0 \text{ if } -\infty \leq SupportF^*_{td} < \tau F_1 \end{cases}$$

$$\text{Latent Decomposition} \begin{cases} \Pr(SupportM_{td} = 0) = \Phi(\tau M_1 - SupportM^*_{td}) \\ \Pr(SupportM_{td} = 1) = \Phi(\tau M_2 - SupportM^*_{td}) - \Phi(\tau M_1 - SupportM^*_{td}) \\ \Pr(SupportM_{td} = 2) = \Phi(\tau M_3 - SupportM^*_{td}) - \Phi(\tau M_2 - SupportM^*_{td}) \\ \Pr(SupportM_{td} = 3) = \Phi(\tau M_4 - SupportM^*_{td}) - \Phi(\tau M_3 - SupportM^*_{td}) \\ \Pr(SupportM_{td} = 4) = \Phi(-\tau M_4 + SupportM^*_{td}) \\ \\ \Pr(SupportF_{td} = 0) = \Phi(\tau F_1 - SupportF^*_{td}) \\ \Pr(SupportF_{td} = 1) = \Phi(\tau F_2 - SupportF^*_{td}) - \Phi(\tau F_1 - SupportF^*_{td}) \\ \Pr(SupportF_{td} = 2) = \Phi(\tau F_3 - SupportF^*_{td}) - \Phi(\tau F_2 - SupportF^*_{td}) \\ \Pr(SupportF_{td} = 3) = \Phi(\tau F_4 - SupportF^*_{td}) - \Phi(\tau F_3 - SupportF^*_{td}) \\ \Pr(SupportF_{td} = 4) = \Phi(-\tau F_4 + SupportF^*_{td}) \\ \\ SupportM^*_{td} = SupportM^{*w}_{td} + \alpha_{1d} \\ ConnectM_{td} = ConnectM^w_{td} + ConnectM^b_d \\ SupportF^*_{td} = SupportF^{*w}_{td} + \alpha_{2d} \\ ConnectF_{td} = ConnectF^w_{td} + ConnectF^b_d \end{cases} \quad (6)$$

$$
\text{Within-Person}\quad
\begin{cases}
SupportM_{td}^{*w} = \phi_{1d}SupportM_{(t-1)d}^{*w} + \phi_{3d}SupportF_{(t-1)d}^{*w} + \beta_{1d}ConnectM_{td}^{w} + eM_{td} \\
SupportF_{td}^{*w} = \phi_{2d}SupportF_{(t-1)d}^{*w} + \phi_{4d}SupportM_{(t-1)d}^{*w} + \beta_{2d}ConnectF_{td}^{w} + eF_{td} \\
\begin{bmatrix} ConnectM_{td}^{w} \\ ConnectF_{td}^{w} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \omega_1 & \\ 0 & \omega_2 \end{bmatrix}\right) \\
\begin{bmatrix} eM_{td} \\ eF_{td} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \\ \rho_{21} & 1 \end{bmatrix}\right)
\end{cases}
$$

$$
\text{Between-Person}\quad
\begin{cases}
\alpha_{1d} = \gamma_{11}ConnectM_{d}^{b} + u_{1d} \\
\alpha_{2d} = \gamma_{21}ConnectF_{d}^{b} + u_{2d} \\
\phi_{1d} = \gamma_{30} + u_{3d} \\
\phi_{2d} = \gamma_{40} + u_{4d} \\
\phi_{3d} = \gamma_{50} + u_{5d} \\
\phi_{4d} = \gamma_{60} + u_{6d} \\
\beta_{1d} = \gamma_{70} + u_{7d} \\
\beta_{2d} = \gamma_{80} + u_{8d} \\
ConnectM_{d}^{b} = \gamma_{90} + u_{9d} \\
ConnectF_{d}^{b} = \gamma_{10,0} + u_{10,d} \\
\\
\mathbf{u}_d \sim MVN\left(\mathbf{0}_{10}, \begin{bmatrix} \tau_{11} & & & & \\ \tau_{21} & \tau_{22} & & & \\ & & \tau_{33} & & \\ & & & \ddots & \\ & & & & \tau_{10,10} \end{bmatrix}\right)
\end{cases}
$$

This model is also represented in Figure 8. At the top of equation 6, observed outcomes $SupportM_{td}$ and $SupportF_{td}$ are defined in terms of their respective underlying continuous distributions $SupportM_{td}^{*}$ and $SupportF_{td}^{*}$. The latent decomposition section further specifies these underlying continuous outcomes into person-specific latent means (α_{1d} and α_{2d}) reflecting a person's baseline proportion of feeling supported, and within-person effects ($SupportM_{td}^{*w}$ and $SupportF_{td}^{*w}$) reflecting person-specific momentary or day-to-day deviations from that baseline. This same composition happens for the continuous covariate ($ConnectM_{td}$ and $ConnectF_{td}$, both of which have been person-mean centered) where person-specific within-subject effects are denoted with a *w* superscript and the person specific latent means are denoted with a *b* superscript.

In the within-person equations, actor effects for male and female dyad members are denoted ϕ_{1d} and ϕ_{2d} respectively, and partner effects for male and female dyad members are denoted ϕ_{3d} and ϕ_{4d} respectively. The effect of each person's perception of their overall relationship connectedness on their same-day feelings of being supported (within-level connect predicting support) is denoted as β_{1d} for males and β_{2d} for females. The within-subject effects for connectedness for male and female partners follow a multivariate normal distribution with mean vector 0 and variances ω_1 and ω_2, respectively. Similarly, the within-level residuals for $SupportM_{td}^{*w}$ and $SupportF_{td}^{*w}$ (eM_{td} and eF_{td}) follow a multivariate normal distribution with a mean vector of zero and variances fixed to 1 for identification. While the variances for these residuals are fixed, the covariance between them is not; it is denoted ρ_{21} and represents the correlation between male and female errors (or volatility) around their latent person-means for the underlying continuous process of support. Both the residual variances and the residual correlation values are assumed to be constant across the time series.

Each effect that has a *d* subscript in the latent decomposition or within-level models is allowed to vary across dyads, and appear in the between-level equations as outcomes. This includes the latent means for the underlying continuous outcomes and actor and partner effects, along with the within-level covariate effects (β_{1d} and β_{2d}) and the covariate latent means ($ConnectM_d^b$ and $ConnectF_d^b$). Each of these outcomes are specified by a fixed effect (γ terms for actor, partner, and within-level covariate effects) and random effects (u terms for all equations). The latent means for the underlying continuous outcomes are also predicted by the latent means for the continuous covariate ($\gamma_{11}ConnectM_d^b$ for male partners and $\gamma_{21}ConnectF_d^b$ for female partners). Finally, the random effects are specified as following a multivariate normal distribution with mean vector of 0 and variances represented by a covariance matrix with variances on the diagonal, a covariance between male and female intercept random effects on the off-diagonal (τ_{21}) and all other off-diagonal covariances not estimated (set to 0).

This model was specified in M*plus* 8.8 (Muthén & Muthén, 1998-2017) as a two-level random effects model with Bayesian estimation with 2 chains and a specified range of iterations from 25,000 – 1,000,000. The first half of the total iterations used to estimate the model were used as burn-in, with the second half of iterations used to construct posterior distributions for each estimated parameter. Prior distributions were set to the M*plus* defaults, which are $\mathcal{N}\left(0,\infty\right)$ for the actor, partner, and within-level covariate fixed effects (where "infinity" is defined as 10^{10}; Muthén & Muthén, 1998-2017, pg. 775), $\mathcal{N}\left(0,5\right)$ for the threshold fixed effects, $\mathcal{W}^{-1}\left(\mathbf{I},3\right)$ for the between- and within-level residual variances and covariances of the outcomes, and $\Gamma^{-1}\left(-1,0\right)$ for the within- and between-level variances for connectedness for male and female partners as well as all 6 of the between-level random effect variances.

Results

The median of the posterior distribution for each parameter along with its 95% credible interval (constructed using the 2.5 and 97.5 percentiles) are reported in Table 1.

Fixed effect and variability parameters whose credible interval did not contain 0 are bolded. Not shown in Table 1 is the observed percentage of male and female participants in each of the 5 categories on the outcome across the 21-day time series:

	Males	**Females**
Not at all	5.0%	5.2%
A little	7.6%	6.3%
Moderately	19.1%	14.6%
Quite a bit	25.0%	21.5%
Extremely	43.3%	52.3%

Note that both males and females reported feeling "Extremely" supported by their partner around half the time. The fixed effect threshold parameters for this response for male and female partners are $\tau M_4 = 0.243$ and $\tau F_4 = 0.052$ respectively, and both of their 95% credible intervals contain 0; this makes sense, as a value of 0 in this case corresponds to a 50% probability of that outcome.

Recall that in probit models in M*plus*, the parameterization is such that the average intercept value is fixed to 0 (and thus does not appear in the results). However, the intercept is still allowed to capture variability in latent person-means (i.e., people having different baseline levels of feeling supported), and as such can have both a random effect variance and (in the multivariate case) a random effect covariance. These random effects, can be interpreted as shifting parameters for the threshold values for individuals (rather than shifting the intercept value itself) to retain the underlying normal distribution mean of 0 for each person. In practical terms, the intercept random effect terms can be interpreted as giving people different probabilities of responding to the categories on the outcome regardless of whether it is viewed as shifting the thresholds and keeping the mean at 0 or shifting the mean and keeping the threshold values the same.

The variance parameter for the thresholds/intercept for both males and females indicate that there is variability across dyads for these latent person-mean values. The posterior median for the male partner's variability is $\sigma_{11} = 1.912$ with a 95% credible interval of [0.914, 3.544]. Assuming normality of the random effects, most person-specific shifts across male partners are estimated to fall between $0 \pm 1.96\sqrt{1.912} = \left[-2.710, 2.710\right]$, which represents a uniform shifting of the thresholds that discretize their underlying continuous distribution of support ($SupportM^*_{td}$). As such, for a male participant whose latent-person mean for feeling supported falls one standard deviation above the mean, the threshold values are shifted left (become more negative), making them more likely to respond in categories on the higher end of the response scale (i.e., they have higher baseline values of support and are therefore more likely to feel "Extremely" supported by their partner). To illustrate this, Figure 9 shows the average threshold locations across participants (center panel with "0" header) as well as the thresholds for a person who falls one standard deviation $\left(\sqrt{1.912} = 1.383\right)$ below (top panel with "-1" header) and one standard deviation above (bottom panel with "1" header) the average thresholds. Looking specifically at the fourth threshold (represented by the blue line furthest to the right), notice how just under half of men at the average latent person-mean for feeling supported fall above that threshold – this matches the 43.3% overall endorsement rate for feeling "Extremely" supported. Men who are 1 standard deviation above the latent person-mean for feeling supported, however, are *more* likely to report feeling “Extremely” supported by their partner, as evidenced by the larger portion of the distribution that falls above the fourth threshold in the bottom panel of Figure 9. These shifts above and below the average are conditional on having a latent person-mean rating for connectedness that is the average across all male latent person-means in the sample ($ConnectM_d^b = 0$ where $ConnectM_d^b$ is latent person-mean centered).

Looking at the within-level residual correlation parameter (ρ_{21}) which captures the relationship between the day-to-day volatility for male and female partners that is not explained by other effects in the model, the value is 0.169 with a 95% credible interval of [0.034, 0.300]. This indicates that the unexplained volatility around person-specific latent means are positively related for male and female partners: a female partner with small day-to-day unexplained variations from their latent person-mean for support is more likely to have a male partner who also has small day-to-day unexplained variations from their latent person-mean for support.

When considering actor effects in this model, the posterior median for the male actor effect (representing the polychoric autocorrelation fixed effect for males) is $\gamma_{30} = 0.245$ and the 95% credible interval for this effect does not contain 0 [0.107, 0.403], indicating that there is a positive relationship between feeling supported at time *t* – 1 and feeling supported at time *t* for male partners. This actor effect variability is $\sigma_{33} = 0.087$, indicating that this relationship varies across male partners ($0.245 \pm 1.96\sqrt{0.087} = \left[-0.333, 0.823\right]$) where it is a moderate negative relationship for some people, and a strong positive relationship for others. The actor effect for female partners is $\gamma_{40} = 0.107$ and is plausibly null (0 falls within the credible interval); the variability around that parameter ($\sigma_{44} = 0.103$), however, indicates that while the overall effect might be 0, it varies across female partners: $0.107 \pm 1.96\sqrt{0.103} = [-0.522, 0.736]$ where some have a strong negative association between previous-day feelings of support and present-day feelings of support, while others have a strong positive association.

When considering the posterior medians and 95% credible intervals for the partner effects, the male partner effect ($\gamma_{50} = 0.018$, [-0.134, 0.151]) is plausibly null, but the female partner effect ($\gamma_{60} = 0.141$, [0.026, 0.290]) is not. This indicates that overall, a male reporting feeling supported at time *t – 1* corresponds with a female partner being more likely to report feeling supported at time *t*, but not vice versa. Looking at the variability of these estimates, there is some variability in the male partner effect ($\sigma_{55} = 0.097$, [0.009, 0.304]) and the female partner effect ($\sigma_{66} = 0.043$, [0.002, 0.199]). This indicates that estimates of the male partner effect varies across dyads $0.018 \pm 1.96\sqrt{0.097} = [-0.592, 0.628]$ even though the overall fixed effect is 0, and estimates of the female partner effect also varies across dyads $0.141 \pm 1.96\sqrt{0.043} = [-0.265, 0.547]$ indicating that while the average relationship is positive, some women are negatively influenced by their partner's previous-day feelings of support and others are more positively influenced relative to the fixed effect. Note that, unlike with the actor effects, these partner effect values cannot be interpreted directly as correlations because the variability for the female outcome is not equal to the variability for the male outcome, and are instead indicative of changes on the underlying normal distribution for support.

Finally, when looking at the effect of relationship quality as assessed by relationship connectedness on whether a person felt cared for by their partner, there are three separate effects for each partner to consider. First, the person-specific latent means ($ConnectM_d^b$ and $ConnectF_d^b$) represent each person's baseline levels of connectedness and appear as effects γ_{90} and $\gamma_{10,10}$ for males and females, respectively. Next, the effect is disaggregated into within- and between-person effects. The within-person effect of relationship connectedness ($ConnectM_{td}^w$ and $ConnectF_{td}^w$) represents the effect of day-to-day deviations from their person-specific latent mean for connectedness on the probability that they felt supported by their partner that day (represented by fixed effects γ_{70} and γ_{80} for males and females, respectively); this effect is allowed to vary across dyads. The between-person effect of relationship connectedness represents the effect of a person's baseline level of connectedness on their baseline level of feeling supported by their partner (represented by effects γ_{11} and γ_{21} for males and females, respectively).

Notice that the effect for the latent means for both male and female partners (γ_{90} and $\gamma_{10,10}$) have 95% credible intervals that contain 0. This is unsurprising as the connectedness variable, originally on a 0 to 20 scale, was centered for both groups prior to analysis (raw score male group mean = 13.645, female group mean = 12.216); as a result, both the male and female average latent person-means are close to their group mean value, though not perfectly at 0. The variability around these latent means is large for both male and female participants – for males the posterior median value of variability is $\sigma_{99} = 13.526$ meaning person-specific latent means fall mostly within the range $-0.212 \pm 1.96\sqrt{13.526} = [-7.420, 6.996]$ (or between approximately 6 and 20 on the 0 to 20 scale), and for females the posterior median value of variability is $\sigma_{10,10} = 13.532$ meaning person-specific latent means fall mostly within the range $0.246 \pm 1.96\sqrt{13.532} = [-6.964, 7.456]$ (or between approximately 5 and 20 on the 0 to 20 scale). Additionally, the within-dyad variability around the person-specific latent means for connectedness is large ($\omega_1 = 12.786$ for males and $\omega_2 = 17.929$ for females) indicating that both male and female participants show a large amount of variability around their person-specific latent mean of connectedness across the timeseries.

The within-person effect of momentary deviations from their latent-person mean of connectedness on the probability that they felt cared for by their partner that day was non-null for female participants ($\gamma_{80} = 0.150$) and for male participants ($\gamma_{70} = 0.138$). This means that reporting higher connectedness resulted in a higher probability on the underlying continuous distribution of feeling cared for by their partner that day. The variability around both of these effects suggest that most female participants vary around the fixed effect $0.150 \pm 1.96\sqrt{0.020} = [-0.127, 0.427]$ and most male participants vary around their fixed effect $0.138 \pm 1.96\sqrt{0.014} = [-0.094, 0.370]$.

The disaggregated between-person effects for male and female partners reflect how different latent means of connectedness affect the probability of feeling cared for by their partner. The fixed effects of both males and females have credible intervals that do not contain zero ($\gamma_{11} = 0.304$ for males and $\gamma_{21} = 0.354$ for females) indicating that people who have a higher latent person-mean of connectedness have a higher probability of reporting feeling cared for by their partner. To illustrate this relationship, Table 2 shows the predicted probability that a person would respond to a certain category on the outcome given their latent person-mean level of the covariate. The heat map coloration shows low-probability categories in darker shading and high-probability categories in lighter shading. Notice how as the latent person-mean of connectedness increases, so does the probability of responding in higher/more positive categories of support: people who feel more connected to their partners are more likely to also report that they feel supported by their partner.

Discussion

Overall, this model suggests that higher latent person-means for connectedness are related to higher probabilities of feeling supported for both men and women, and increases in momentary day-to-day changes in connectedness are related to higher probabilities of feeling supported. Additionally, the male actor effect and female partner effect were non-null in the core APIM within-person model, indicating that men's reports of feeling supported by their partner at time $t - 1$ affected both their own and their partner's reports of support at time t. Both actor and partner effects appeared to vary across dyads regardless of whether their fixed effect was different from 0 or not. When considering the latent decomposition of the underlying continuous process for the outcome, both men and women showed a negative skew in their response distributions (represented by the first three threshold values being negative and non-zero with the fourth threshold being positive but not different from 0). This indicates that people in general were more likely to endorse feeling "Extremely" supported relative to feeling

“Not at all” supported by their partner. In the following section, these values are used as starting points for population values in the simulation.

CHAPTER 5

METHODS

Proposed Study and Hypotheses

I conducted a Monte Carlo simulation to evaluate how DSEMs perform when the outcome is a 2-, 3-, or 5-category ordinal variable. To do so, I generated a univariate lag-1 autoregressive model representing individual data and a lag-1 vector autoregressive model as an L-APIM representing dyadic data with respect to four manipulated conditions: number of categories on the outcome, total number of clusters (people in the univariate case, dyads in the multivariate case), number of timepoints, and proportion of responses in each of the categories on the outcome. Additionally, for the 3- and 5-category outcomes I fit the model treating the outcome as both categorical and continuous. The univariate and multivariate models are formally compared in the results section, though trends are considered informally. I have four research questions and accompanying hypotheses related to these conditions.

Research Question 1: How do number of clusters and number of timepoints affect estimates? When considering sample size and number of timepoints, there are a few issues that can have an impact on how these two factors affect the performance of a categorical DSEM. While Schultzberg & Muthén (2018) examined only univariate continuous DSEM models, they found that a larger number of people is better able to compensate for potential issues with a small number of timepoints than a large number of timepoints is able to compensate for issues due to having a dataset with relatively fewer people. They add a caveat, however, that the random autocorrelation in particular requires a larger number of timepoints to accurately estimate (pg. 511). Additionally, Gistelinck et al. (2021) found that continuous DSEM does not perform well (relative to an uncentered approach) when the number of timepoints are as few as 4, but works well as long as the number of timepoints is 10 or more – this result was the same when sample size was 50 or 200. Finally, Asparouhov et al. (2018) found that a small number of timepoints (less than 20) could lead to biased estimates in a continuous DSEM as a result of the default uninformative Bayesian priors becoming unintentionally informative in the analysis, though this bias was not present when the number of timepoints was 30 or greater (pg. 362).

All of the studies discussed above focused on continuous outcome DSEM models. To my knowledge, no work has been done to a similar degree on sample size requirements when there is instead a categorical outcome, perhaps due to this being a relatively new module in M*plus.* I hypothesized that, similar to the continuous outcome findings, the number of clusters (individuals or dyads) will have a large impact on the performance of the model, and number of timepoints will have an impact on the recovery of the autoregressive parameter(s) in particular. The sample size and timepoint requirements for categorical outcomes may even be larger than those for continuous outcomes because the observed data may show less variability within and between clusters.

Research Question 2: How does the proportion of responses on the outcome affect estimates? Categorical items that have a uniform or approximate normal distribution of responses across all categories will show more variability within a time series and between individuals' time series than items that have a skewed distribution of responses. For example, if a skewed distribution has 70% of scores in the highest category, the dataset will contain many more occurrences of that value than any other value relative to a uniform or even a normal distribution. Low variability within a time series is particularly problematic when using DSEM, as within-person variability is a main focus of the analysis. Additionally, a categorical outcome in general can further limit variability in responses relative to a continuous outcome covering the same range of values, which can impact the "optimal" number of time points for different cluster sizes between categorical and continuous outcomes. Because of this, I hypothesized that larger numbers of clusters and more timepoints will be necessary to recover effects from models with skewed and normal distributions relative to uniform distributions.

Research Question 3: How does the number of categories on the outcome affect estimates? Asparouhov et al. (2018, pg. 363) discuss how models with ordered polytomous outcomes generally provide more information in an analysis, providing more precise estimation relative to binary outcomes. While it can be possible to simplify the categorical DSEM when the outcome is binary (by replacing the threshold parameterization with a mean parameter for the outcome y_{it}^{*}), making estimation more efficient, it does not appear that M*plus* employs this reparameterization technique in its estimation. Whether the outcome is binary or polytomous, the parameters in the output are always thresholds as opposed to intercept means. For these reasons, I hypothesized that there will be a difference in how the number of categories on the outcome affects estimates, where the 5-category outcome will provide more precise estimates relative to the binary and 3-category outcomes, as it will provide more information and be better approximated by an underlying continuous normal distribution.

Research Question 4: How does treating the 3- and 5-category outcomes as continuous compare to using a categorical model? As mentioned earlier, the extent to which researchers actually use categorical models over continuous ones varies from "sometimes" to "never" (e.g., Robitzsch, 2020), though outcomes with 5 or more categories tend to be a popular benchmark for using continuous methods. Rhemtulla et al. (2012) recommended using categorical methods when there were fewer than 5 categories, but noted that continuous methods worked well when there were 5 or more categories, sample sizes were small, or category thresholds were symmetric. While their results suggest that a 3-category outcome would be best suited to a categorical model, the 5-category condition seemed to be right on the threshold for whether categorical or continuous methods performed better. Given that Rhemtulla et al. (2012) focuses on CFA rather than intensive longitudinal data, it is worth investigating the extent to which a focus on variances and autocorrelations rather than factor loadings warrants different recommendations on how a 5-category outcome can be treated with DSEM. I hypothesized that treating the outcome as categorical will be more precise and show less bias relative to treating the outcome as continuous for both the 3- and 5-category conditions, especially for the skewed distributions with large sample sizes.

Manipulated Simulation Factors

A Monte Carlo simulation with five factors was implemented for both a multivariate (dyadic) and univariate model. The four manipulated factors imposed on all conditions are total sample size (two levels), number of timepoints (three levels), number of categories on the outcome (three levels), and proportion of responses in each category of the outcome (two levels for binary outcome, three levels for other outcomes).

Additionally, the 3- and 5-category outcome conditions were fit with both categorical and continuous models. In total for both univariate and multivariate models, there were 12 design cells for the 2-category and 36 design cells for the 3- and 5-category outcomes. For each cell, 300 replications were simulated with a goal of having at least 250 converged replications for analysis.

The total sample sizes investigated were 50 and 200. For the univariate model, the sample size represents individuals, while in the multivariate model it represents number of dyads. Asparouhov & Muthén (2019) evaluated a large sample size of 5000 (with timepoints of 20, 50, and 200) to show that the autoregressive parameter was least biased when latent centering was used relative to other centering options, and in a single-sentence follow-up found that the autoregressive parameter was unbiased when the sample size was 100 (across the same timepoint conditions of 20, 50, and 200). Social science research often involves sample sizes smaller than 100 (see McNeish, 2019), and so the simulation conditions chosen for this project aimed to fill in how categorical DSEM performs with more realistic sample sizes – 50 chosen to approximate the motivating data example, and 200 chosen to represent a "larger" sample size in intensive longitudinal contexts.

The number of timepoints investigated were 14, 28, and 56. In Asparouhov & Muthén's (2019) categorical DSEM simulation study, they evaluated 20, 50, and 200 timepoints and found that each size produced unbiased effects (albeit with sample sizes of 100 and 5,000). Asparouhov et al. (2018) specify that fewer than 20 timepoints might result in effects that have bias due to the default Bayesian prior distributions, but this bias goes away with 30 or more timepoints (pg. 362). Gistelinck et al. (2021) found that

DSEM with latent-mean centering performs well as long as timepoints were at least 10 (and this was the same whether the examined sample size was 50 or 200). Furthermore, Gunthert & Wenze (2014) specify that most daily diary studies range from 7 to 30 days, with a modal number of timepoints around 14 days (pg. 150). With this past research in mind, these categories of timepoints were chosen to represent the modal number of timepoints (14 days; possibly small enough to see bias due to priors), the upper end of most studies (28 days, or four weeks) which may be large enough to provide unbiased estimates (though falls at the upper end of a nebulous zone based on Asparouhov et al., 2018 recommendations above), and a longer daily diary study (56 days, or about 8 weeks; should not be biased based on past research).

The three different numbers of categories on the outcome investigated were 2, 3, and 5. Binary outcomes are what have been modeled in past DSEM simulation studies (Asparouhov & Muthén, 2019, for DSEM; Asparouhov & Muthén, 2020, for RDSEM) with follow-up text stating that these models were generalizable to polytomous outcome models. Asparouhov & Muthén (2018) did examine a polytomous outcome DSEM in a simulation study, but it was using a less optimal specification for the model that is no longer recommended (Asparouhov & Muthén, 2019, pg. 136) as it required a large number of timepoints per person to evaluate a subject-specific autocorrelation (Asparouhov & Muthén, 2018, pg. 373). Though the categorical DSEM framework specified in equation 5 follows the preferred method for categorical DSEM outlined by Asparouhov & Muthén (2019), no formal simulation study has evaluated how this model performs with a polytomous outcome. Because binary and 5-category items are commonly used in daily diary designs (e.g., motivating example above, Cook et al., 2010), this simulation expands on existing simulations for categorical DSEM by examining binary and 5-category outcome models, along with a 3-category outcome to look for patterns in the range between 2- and 5-categories.

The final manipulated simulation factor is the proportion of responses in each category of the outcome. For this project, three different situations were examined: a uniform distribution on the outcome, an approximate normal/symmetric distribution (hereafter referred to as the "normal" distribution condition), and a skewed distribution of responses, though the 2-category outcome conditions were limited to just a uniform or skewed distribution. The uniform distribution put an equal proportion of responses in each category, the normal distribution split the Z-score range from -3 to +3 into equal intervals based on number of categories while also ensuring no category has fewer than 10% of all observations, and the skewed distribution split the percentage of responses on the outcome such that the lowest category has a majority of observations and each subsequent category has fewer, though again ensuring that no category has fewer than 10% of all observations. The skewed response distribution mirrors both the motivating data example outcome for feeling supported (in which most people reported feeling "Extremely" supported, fewer reported feeling "Quite a bit", fewer still "Moderately" supported, etc.) and studies that might focus on relatively infrequent behaviors where you might expect, for example, a majority of participants to respond in the lowest category for level of depression or substance use experienced that day. The approximate percentage of scores in each category across different conditions and total number of categories is shown in Table 3. These values were selected as the uniform distribution is used in the binary DSEM simulation outlined in Asparouhov & Muthén (2019), the normal distribution reflects the assumption of an underlying normal distribution, and the skewed categories were similar to those observed in the 2-, 3-, and 5-category items used as outcomes in the motivating examples. In the multivariate conditions, both outcomes followed the distributions specified above.

Data Generation

There were two broad groups of generating models: those with univariate and those with multivariate outcomes, both of which also had two continuous covariates. Data generation and model fitting all happened in M*plus* Version 8.8 (Muthén & Muthén, 1998-2017).

Fixed and random effects for intercept and slope parameters were based on the motivating example as discussed above (see Table 1) for the multivariate conditions, though were modified such that one actor, partner, within- and between-level covariate effect was set to 0 and one was non-0 (retaining the larger value from the motivating data example for each effect), and the random effect variances were rounded to either 0.10 or 0.05; the univariate conditions retained as many population values from the multivariate conditions as possible, splitting the two covariates into one that had all non-null population values and one that had all zeroes as population values (including the random effect variance). These population values within the multivariate and univariate conditions were kept consistent across other manipulated factors to allow for comparisons across conditions (i.e., even though the motivating example presented a 5-category outcome that had a skewed distribution, the actor, partner, and intercept effect values were kept the same for the normal and uniform conditions as well, as well as for the corresponding 3- and 2-category outcome effects).

Univariate DSEM. The generating model for all (2-, 3-, and 5-cateogory) univariate DSEM models is presented in Figure 10 and the non-threshold population values are shown in Table 4. The corresponding series of equations that were used to generate data can be expressed generally for categorical variable y_{ti} and continuous covariates $x1_{ti}$ and $x2_{ti}$ as follows:

$$
\begin{array}{ll}
\text{Latent Decomposition} & \begin{cases} y_{ti}^{*} = y_{ti}^{*w} + \alpha_i \\ x1_{ti} = x1_{ti}^{w} + x1_{i}^{b} \\ x2_{ti} = x2_{ti}^{w} + x2_{i}^{b} \end{cases} \\
\text{Within-Person} & \begin{cases} y_{ti}^{*w} = \phi_i y_{(t-1)i}^{w} + \beta_{1i} x1_{ti}^{w} + \beta_{2i} x2_{ti}^{w} + e_{ti} \\ \begin{bmatrix} x1_{ti}^{w} \\ x2_{ti}^{w} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \omega_1 & \\ 0 & \omega_2 \end{bmatrix} \right) \\ e_{ti} \sim N(0,1) \end{cases} \\
\text{Between-Person} & \begin{cases} \alpha_i = \gamma_{01} x1_{i}^{b} + \gamma_{02} x2_{i}^{b} + u_{0i} \\ \phi_i = \gamma_{10} + u_{1i} \\ \beta_{1i} = \gamma_{20} + u_{2i} \\ \beta_{2i} = \gamma_{30} + u_{3i} \\ x1_{i}^{b} = \gamma_{40} + u_{4i} \\ x2_{i}^{b} = \gamma_{50} + u_{5i} \\ \mathbf{u}_i \sim MVN(\mathbf{0}_6, diag[\sigma_{00}, \sigma_{11}, \ldots, \sigma_{55}]) \end{cases}
\end{array} \tag{7}
$$

As mentioned previously, these values and equations are the same for the 2-, 3-, and 5-category univariate models, with the exception of the threshold parameters (discussed below). Note that the within-level residual variance for the outcome does not appear in Table 4 as it is fixed to 1 by default and is therefore not a specified population value.

Multivariate DSEM. The generating model for all (2-, 3-, and 5-category) multivariate DSEM models is presented in Figure 11 and the non-threshold population values are shown in Table 5. The corresponding series of equations that were used to generate the data can be expressed generally for categorical variables $y1_{td}$ and $y2_{td}$ and continuous covariates $x1_{td}$ and $x2_{td}$ as follows:

$$\text{Latent Decomposition}\begin{cases} y1^{*}_{td} = y1^{*w}_{td} + \alpha_{1d} \\ x1_{td} = x1^{w}_{td} + x1^{b}_{d} \\ y2^{*}_{td} = y2^{*w}_{td} + \alpha_{2d} \\ x2_{td} = x2^{w}_{td} + x2^{b}_{d} \end{cases}$$

$$\text{Within-Person}\quad\begin{cases} y1^{*w}_{td} = \phi_{1d} y1^{w}_{(t-1)d} + \phi_{3d} y2^{w}_{(t-1)d} + \beta_{1d} x1^{w}_{td} + e_{1td} \\ y2^{*w}_{td} = \phi_{2d} y2^{w}_{(t-1)d} + \phi_{4d} y1^{w}_{(t-1)d} + \beta_{2d} x2^{w}_{td} + e_{2td} \\ \begin{bmatrix} x1^{w}_{td} \\ x2^{w}_{td} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \omega_1 & \\ 0 & \omega_2 \end{bmatrix}\right) \\ \begin{bmatrix} e1_{td} \\ e2_{td} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \\ \rho_{21} & 1 \end{bmatrix}\right) \end{cases}$$

$$\text{Between-Person}\quad\begin{cases} \alpha_{1d} = \gamma_{11} x1^{b}_{d} + u_{1d} \\ \alpha_{2d} = \gamma_{21} x2^{b}_{d} + u_{2d} \\ \phi_{1d} = \gamma_{30} + u_{3d} \\ \phi_{2d} = \gamma_{40} + u_{4d} \\ \phi_{3d} = \gamma_{50} + u_{5d} \\ \phi_{4d} = \gamma_{60} + u_{6d} \\ \beta_{1d} = \gamma_{70} + u_{7d} \\ \beta_{2d} = \gamma_{80} + u_{8d} \\ x1^{b}_{d} = \gamma_{90} + u_{9d} \\ x2^{b}_{d} = \gamma_{10,0} + u_{10,d} \\ \mathbf{u}_d \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & & & & \\ \sigma_{21} & \sigma_{22} & & & \\ & & \sigma_{33} & & \\ & & & \ddots & \\ & & & & \sigma_{10,10} \end{bmatrix}\right) \end{cases} \qquad (8)$$

As with the univariate models, the equations and population values are the same across multivariate conditions with the exception of the threshold values, and the within-level residual variances have a fixed value of 1 and thus do not appear in Table 5.

Threshold values. The threshold population values for all generating models are shown in Table 6. In order to obtain the percentage distributions specified in Table 3, these threshold values were adjusted from their marginal Z-score values (as would be seen in a traditional theta parameterization of the continuous underlying distribution of the outcome) to accommodate both the autocorrelation effect of their corresponding outcome and the total variance of the outcome. For example, note that the value for the first threshold of the 5-category multivariate Y1 outcome is -2.417 (the top-leftmost value of Table 6). This is different from the *marginal* Z-score needed to specify 10% of the underlying normal distribution manifesting as the lowest observed response category (-1.282). The discrepancy between the *marginal* threshold and the *actual* threshold needed to obtain 10% of all Y1 responses in category 1 is due to having a non-zero autocorrelation value (0.25, as specified in Table 5) along with the total variability of Y1 being around 1.886 instead of being a standard normal distribution with a variance of exactly 1 as would be expected in a theta parameterization. Because Y1 has more variability than the standard normal, the threshold value gets pulled further into the tail of the distribution by a factor of 1.886, thus changing the value from -1.282 to -2.417. This same process was followed for all threshold values across the different conditions to adjust for both the autocorrelation and the total variance of each outcome within each condition.

While not an explicit part of the generating models, Table 6 also displays the average score for each outcome across conditions. These values served as the intercept population values (taking the place of the threshold values) when the generated data were fit as a continuous DSEM model. All other population values were specified as shown in Tables 4 and 5 regardless of whether the outcome was treated as categorical or continuous.

Evaluation Criteria

For each condition, a model equivalent to the generating model was fit to each of the 300 replications generated (treating the outcome as categorical). The replications for the 3- and 5-category conditions were also fit to a DSEM model that treated the outcome as continuous (but had no other meaningful changes). The results for each converged replication were saved, comprised of the posterior median and standard deviation for each estimate. The within- and between-level variances for each outcome in each condition were also calculated by generating a large (N = 10,000 people or dyads, T = 56 timepoints) so that the within- and between-level covariate parameters, partner effect fixed effects, level-1 residual covariance, and level-2 outcome residual variances and covariance could be standardized to more easily compare between the categorical and continuous conditions in the univariate and multivariate models. To evaluate the performance of each model, I used those estimates and the population values (on the raw or standardized scale) to calculate relative bias (for non-null population parameters) or absolute bias (for null population parameters). Using information provided by the M*plus* output files for each condition I compiled the 95% coverage measure and either power or Type-I error for each parameter for each condition. Because coverage can be affected if the parameters are biased, I also calculated an efficacy measure. These measures were calculated based on the following equations:

$$\text{Relative bias} = \frac{\text{Median of posterior distribution - Population value}}{\text{Population value}}$$

$$\text{95\% Coverage} = \frac{\text{Number of credible intervals including population value}}{\text{Number of replications}}$$

$$\text{Power or Type-I Error} = \frac{\text{Number of credible intervals not covering zero}}{\text{Number of replications}}$$

$$\text{Efficacy} = \frac{\text{Average SD of each replication's posterior distribution for a parameter}}{\text{SD of posterior median values across replications}}$$

Relative bias indicates how much the sample medians from the posterior distributions differ from the population values specified in the generating model. A relative bias of 0 indicates that the values are the same, a negative bias indicates that the samples tend to produce larger estimates than the generating model, and a positive bias indicates that the samples produced smaller estimates than the generating model. Absolute bias involves just the numerator of the relative bias equation, and is used for parameters whose population values are 0.

The 95% coverage measure looks at the 95% credible intervals (based on percentiles from the Bayesian posterior distribution for each estimate) and assesses how many of those intervals contain the population value for each design cell. Importantly, this measure is affected by both the bias of the estimates as well as the number of iterations in the Bayesian Gibbs sampler used to estimate the posterior distributions for each parameter. To resolve the latter issue, Schultzberg & Muthén (2018) recommend that 95% coverage values between 0.92 and 0.98 can be considered as adequate for those measures when using less than 10,000 iterations (pg. 498). Coverage values below 0.90 were considered inadequate (Collins, Schafer, & Kam, 2001; Enders & Peugh, 2004). To address the former issue, if a parameter was shown to be biased then efficacy was examined instead. Efficacy compares the standard deviation of the posterior distribution made up of parameter estimates from all replications (e.g., the standard deviation of all 300 posterior distribution median values for the Y1 actor effect; the denominator of the equation), and the average standard deviation value for a parameter across replications (e.g., the average standard deviation of the posterior distribution for all 300 replications of the Y1 actor effect; the numerator of the equation). In this way, the numerator can be considered an *estimate* of the variability in parameter values in the posterior distribution akin to the standard error in a frequentist model, and the denominator can be considered the *observed* variability in parameter values across replications. Values that are close to 1.0 indicate that estimated variability in a parameter matches the observed variability in a parameter. Values above 1.1 indicate that the estimated variability is larger than the observed variability in a parameter, and values below 0.9 indicate that the estimated variability is smaller than the observed variability in a parameter.

The final measure that will be used to examine the performance of these models is power/Type-I error, which is calculated by evaluating how many 95% credible intervals contain 0. When an effect is specified to be non-zero in the population, this measure is power; when an effect is specified to be 0 in the population, this is instead a measure of Type-I error. The population values were designed such that as many different parameter types as possible (e.g., actor/carryover effect, partner effect, within- and between-level covariate effects) had one 0 and one non-0 population value in the generating models so that both Type-I error and power could be assessed. The exception to this is the single autocorrelation effect in the univariate model, which was specified to be non-0 and thus provided power but not Type-I error information. For this parameter, an *estimated* Type-I error rate was examined by evaluating 1 – coverage, which would indicate how often replications did *not* contain the population value of interest in their 95% credible intervals, similar to how the Type-I error rate for parameters with a population value of 0 assesses how often replications did not contain 0 in their 95% credible interval.

Data Analyses

Bias, coverage/efficacy, power, and Type-I error values for each condition are graphed in panel plots as appropriate to evaluate trends and relationships between design factors (e.g., continuous and categorical results for different conditions of the 5-category models, or all categorical results for the multivariate models). These results from the multivariate and univariate models will not be directly compared in panel plots, but general trends will be discussed as appropriate.

CHAPTER 6

RESULTS

A minimum of 297 of 300 replications converged across all conditions. To keep the number of replications consistent across conditions, the results presented below are based on the first 297 replications to converge within a condition. Results are presented first within outcome categories (2, 3, and 5) to discuss within-category trends, and then collated between outcome categories for the categorical and continuous conditions for both the multivariate and univariate models. Only results for the fixed and random effect variances (when appropriate) for the autoregressive/actor, partner, and level-1 and level-2 covariate effects are discussed below.

2-Category Models

Recall that for the 2-category outcome models, only skewed and uniform distributions of responses were specified and replications were fit only as a categorical DSEM. Because there was no continuous model fit, all results are presented in their original units (nothing has been standardized for comparison across continuous and categorical models). Additionally, because bias was present for some conditions for some parameters across the autoregressive/actor, partner, and covariate effects; efficacy results are presented in lieu of coverage.

Bias for Fixed and Random Effect Variances. The average relative or absolute bias values for the multivariate actor and partner fixed effects for both outcomes across conditions are presented in Figure 12. Bars around the average bias value represent the standard error across all 297 replications. The population value appears in the top right-hand corner of each cell. When the population value is zero (Y1 partner effect and Y2 actor effect, the middle two rows of the panel), the bias presented is absolute bias; for these effects, there is very little bias across sample size (50 and 200), distribution on the outcome (where “SK” refers to the skewed distribution condition and “UN” refers to the uniform distribution condition), and number of timepoints (denoted on x-axis). The two non-zero effects (Y1 actor and Y2 partner) are represented by relative bias, and show an average of between 30 and 40% bias at the smallest timepoint condition (T = 14) across both sample size and response outcome distribution conditions. This bias is greatly diminished in the T=28 timepoint conditions, and is essentially eliminated in the T=56 timepoint conditions.

Looking at the univariate model’s relative bias for the autoregressive effect (Figure 13), this trend is largely replicated: there is an average of between 25 and 45% bias at the smallest timepoint condition (T = 14) regardless of sample size (N = 50 individuals on top row, N = 200 individuals on bottom row) and distribution on the outcome (skewed distribution on left-hand column, uniform distribution on right-hand column). Overall for these parameters (autoregressive, actor, partner), the number of timepoints appears to have the greatest (and perhaps the only meaningful) effect on bias.

The average bias values for the multivariate actor and partner random effect variances for both outcomes across conditions are presented in Figure 14. Notice how for Y1, both the actor and partner random effect variance had a population value of 0.10, and for Y2 they both had a population value of 0.05. In general, the line within each cell of the panel plot shows a downward trajectory, indicating that bias is highest when T = 14. This effect is especially prominent for partner effects, and particularly for partner effects when N = 50. Even when the sample size is larger (N = 200) for the partner effects, however, the smaller random effect variance (Y2 partner effect, bottom row of Figure 14) still shows bias when T = 28. Actor effect random effect variances overall show less bias compared to the partner effects, though still have considerable bias when T = 14 regardless of the population value, and when T = 28 for the smaller random effect value (Y2 actor effect, third row of figure) when sample size is smaller (N = 50).

When considering relative bias values for these conditions, which are interpreted as proportions or percentages above or below the population value, it is important to note that when values are small (as is the case with most random effect variances in autoregressive or even cross-lagged relations), a large proportion on the y-axis might correspond to a small increase in actual value. For example, when the Y2 partner random effect variance bias is around 0.7 or 70% when T = 28 timepoints (see first cell of fourth row in Figure 14), this means that the average random effect variance *median of the posterior distribution value* is around 0.085 for that condition instead of the population value of 0.05. The practical implications of overestimating the random effect variance as 0.085 instead of 0.05 for these conditions can vary depending on what purpose the model is serving.

Looking at the univariate model's average relative bias for the autoregressive random effect variance (Figure 15), again a similar trend emerges relative to its multivariate actor effect counterpart. Namely, the main factor affecting bias appears to be number of timepoints, with the largest bias observed with the T = 14 timepoint conditions. Overall, these parameters show the most bias when the random effect variance is small, the parameter is a cross-lagged/partner effect, and both timepoints and sample size is small. Because the random effect variances cannot be negative, this larger bias as the population value decreases (from 0.10 to 0.05) is not particularly surprising: because M*plus* will not allow the value to be negative, it introduces positive bias in the parameter as the population value gets closer to 0.

The average bias for the multivariate models' level-1 and level-2 covariate effects for both outcomes across conditions are presented in Figure 16. The covariate effects that were specified as 0 (level-1 covariate for Y2, denoted as "L1 Y2" in the second row of the figure, and level-2 covariate for Y2, denoted as "L2 Y2" in the fourth row of the figure) show little to no bias across conditions, much like the actor and partner fixed effects. The level-1 covariate for Y1 (denoted as "L1 Y1", top row of Figure 16) shows bias results similar to the non-zero actor and partner effects displayed in Figure 12: largest bias for T = 14 timepoints condition, essentially no bias for the T = 28 and T = 56 timepoint conditions. A difference here, however, is that there also appears to be an effect of sample size: bias for the N = 200 condition (last two columns of the top row) is less than half that of its corresponding N = 50 condition when T = 14 (around 4% versus 12% for the skewed distribution conditions and around 5% versus 11% for the uniform distribution conditions, respectively). Additionally, the level-2 covariate for Y1 (denoted as "L2 Y1", third row of Figure 16) shows bias for T = 14 when N = 50 and the distribution is skewed (first column, third row), but essentially no bias for any other condition across that row.

The average absolute and relative bias for the univariate models' level-1 and level-2 covariate effects across conditions are presented in Figure 17. Once again, the null covariate effects show little to no bias, while the non-zero covariate effects show bias that varies across both timepoint and sample size conditions, with smaller sample size and smaller timepoint conditions showing the highest bias values. The level-1 covariate ("L1 Cov1", top row of Figure 17) also appears to show higher bias in the skewed distribution conditions when both sample size and timepoints are small (N = 50 and T = 14) relative to the uniform distribution condition. Unlike the multivariate level-1 covariate ("L1 Y1", top row of Figure 16), the univariate level-1 covariate still shows bias when T = 28 and N = 50. Looking to the non-zero level-2 covariate ("L2 Cov1", third row of Figure 17), it appears to not be as affected by number of timepoints relative to its multivariate counterpart ("L2 Y1", third row of Figure 16), and instead shows differences only between sample size conditions. Overall, number of timepoints and sample size affect the non-zero level-1 covariate effects for both univariate and multivariate models. For level-2 covariate effects, however, timepoints appear to have a noticeable effect at small sample sizes in the multivariate model, but this is not reflected in the univariate model which seems to only be affected by sample size. Across conditions and parameters, when differences between distributions were detected, it was always that the skewed distribution showed larger bias.

When looking at the average bias for random effect variances of the multivariate level-1 covariate effects in Figure 18, the pattern of results mirror those of the partner random effect variances in Figure 14. Smaller population effect values, lower timepoints, and lower sample size all have higher bias values than other conditions. The level-1 covariate random effect variances do tend to be less biased than their partner effect counterparts though: for example, look at the top-left cell of Figure 18 ("L1 Y1" for N = 50 and a skewed distribution, population value of 0.05) and the bottom-left cell of Figure 14 ("Y2 Part" for N = 50 and a skewed distribution, population value of 0.05). When T =

14, the partner random effect variance relative bias has an average around 185%, while the level-1 covariate random effect variance relative bias average is around 150%. Moving across the timepoint conditions in these cells, the relative average relative bias has values of around 70% and 50% respectively for T = 28, and around 20% and 25% respectively for T = 56, so at larger timepoints these conditions' relative bias values become more similar.

For the univariate models, the average bias for random effect variances of the level-1 covariate effects are shown in Figure 19. Importantly, the two covariates specified in the univariate model were given random effect variance population values of 0.05 ("Cov 1", top row of Figure 19) and 0 ("Cov 2", bottom row of Figure 19), representing covariates with little between-person variability or no between-person variability (see Table 4). As with the recovery of null effects discussed previously, there again seems to be little to no bias for the second covariate across conditions. For the first covariate, however, bias is affected by both sample size and number of timepoints, and to a lesser extent the distribution on the outcome; this mirrors the trends in the multivariate models' covariate random effect variance whose population value was 0.05 (see top row of Figure 18 compared with top row of Figure 19). Generally, there is more bias when N = 50 than when N = 200 for the same timepoint conditions, and there is also more bias when T = 14 timepoints. The N = 50 sample size condition also shows bias even at T = 56 timepoints for the non-zero covariate random effect variance. The skewed distribution is slightly more biased than the uniform distribution for this parameter when N = 50 and T = 28 (middle points on lines in first two columns of top row of Figure 19), though no differences at T = 14 and T = 56. Overall, the random effect variances for the covariate effects across models are more biased when sample size is small (N = 50 dyads or individuals) and when number of timepoints is small (T = 14 timepoints) relative to other conditions, and the skewed distribution shows slightly more bias than the uniform distribution.

Efficacy for Fixed and Random Effect Variances. Overall, the efficacy values for the multivariate actor and partner fixed effects (Figure 20) and multivariate covariate fixed effects (Figure 21) suggest that, in general, the estimate of parameter value variability matches the observed variability in parameter values. The parameter with the most out of bounds values is the Y1 partner effect (second row of Figure 20) when sample size N = 50 and timepoints are T = 14 and/or T = 28. In these cases (particularly when the distribution is uniform), the variability in the partner effect is overestimated, meaning that an individual analysis fitting those conditions (N = 50 dyads, 14 or 28 timepoints, binary outcome with uniform distribution) will tend to report a standard deviation of a null partner effect that is larger than it would be in the population. While the level-1 and level-2 covariate effects with a population value of 0 (in the second and fourth rows of Figure 21) also show efficacy values slightly above the acceptable range, these deviations are much smaller than those of the Y1 partner effect.

The univariate models' autoregressive fixed effect showed acceptable efficacy values across both sample sizes, all three timepoints, and both distribution conditions (Figure 22). The covariate effects (Figure 23) also showed relatively acceptable efficacy values across conditions for all parameters except for the null level-1 covariate effect (second row); this is similar to the efficacy results in the multivariate covariate fixed effects in Figure 21. The null level-1 covariate effects in Figure 23 show efficacy values beyond the acceptable range when the outcome has a uniform distribution; this is seen even when both sample size and number of timepoints is large (N = 200 and T = 56, fourth column of second row), along with for all timepoint conditions when sample size is small (second column of second row). Overall, the fixed effects efficacy results from the multivariate and univariate models suggest that the variability is largely being accurately assessed across conditions, though a uniform distribution, small sample size, and null effect increases the likelihood that the posterior standard deviation will be overestimated for a partner or level-1 covariate effect.

The efficacy results for the multivariate actor and partner random effect variances (Figure 24) and level-1 covariate random effect variances (Figure 25) show more conditions with out of range values relative to their fixed effects counterparts. These happen almost exclusively at the smaller sample size of N = 50 dyads. Of note in Figure 24, however, is the Y1 partner effect at N = 50 and a skewed distribution (far left column in second row) which has both a black and a gray line. The black line represents the results from all 297 replications that converged for that condition, and has an efficacy value that falls way below the acceptable range of values (around 0.40) when number of timepoints is T = 14. Further investigation revealed that one of the 297 replications had a reported posterior median for the Y1 partner random effect variance of 11.88; for context, the next highest reported posterior median for that effect is 1.56. This likely contributed to a larger posterior standard deviation for that parameter for that replication (8.63, with next highest being 1.87 – 8.63 was more than 15 standard deviations above average for those 297 replications), leading to a larger estimated variability and a much larger observed variability for that parameter in that condition. This single replication, in turn, led to a much smaller efficacy ratio for that particular parameter for that particular condition. The gray line represents the efficacy value for that condition if this replication is removed. When considering the gray line and other partner effect efficacy values (second and fourth row of Figure 24), overall the values are within bounds or only slightly out of bounds.

Actor effects in Figure 24 (first and third row) show large efficacy values when N = 50 and T = 14 or 28. This effect is present for both skewed and uniform distribution conditions, and is slightly more pronounced when the population random effect variance value is smaller (Y2 actor effect, third row of Figure 24). For the covariate random effect variances in Figure 25, again the smaller random effect variance population value (0.05, represented by “L1 Y1” in the top row of Figure 25) has more efficacy values that are out of bounds at smaller sample sizes. For these covariate effects it also appears that a uniform distribution leads to righter rates of overestimating the variability of the random effect variance parameter relative to a skewed distribution, which is true for both the Y1 and the Y2 covariate as shown in the second column of Figure 25.

Looking to the random effect variance efficacy results for the univariate models, unlike with the autoregressive fixed effect efficacy shown in Figure 22, the autoregressive effect’s random effect variance has efficacy values that are almost exclusively too high (Figure 26). While this is true even at the larger sample size conditions (N = 200, bottom row of Figure 26), efficacy is generally highest when timepoints are small (T = 14) and sample size is small (N = 50, top row of Figure 26). These efficacy values are also generally higher than the multivariate actor effect random effect variance of the same magnitude (population value of 0.10, see top row of Figure 24 for a comparison) when sample size is N = 50 and about the same when sample size is N = 200. The level-1 covariate random effect variance efficacy results (Figure 27) also shows similar trends to its multivariate counterpart of the same population value (top row of Figure 27 and top row of Figure 25). One exception to the similarities, however, is that efficacy for the univariate effects do not show the same differences between a skewed and uniform distribution when both timepoints and sample size are small (see first two columns of top rows for both Figures 25 and 27). The bottom row of Figure 27 shows the efficacy results for the level-1 covariate random effect variance that was specified to be 0. All sample size, timepoint, and outcome distribution show efficacy values that are way above the range of acceptable values. As mentioned previously, however, this is not altogether unsurprising given that M*plus* will not allow for that parameter to be exactly zero; by producing estimates for this parameter that are necessarily larger than the population value, it makes sense that the estimated variability will exceed that of the observed variability.

Overall, the multivariate and univariate random effect variance efficacy results again suggest that a lower population value for random effect variances leads to higher rates of overestimation of that parameter's variability, particularly for actor effects and continuous covariate effects. A small sample size, smaller number of timepoints, or having a continuous covariate with a uniformly distributed outcome also contribute to this overestimation. The default prior distributions for the random effect variance parameters in M*plus* may be contributing to the overestimation of the variability in these parameters. The default for these parameters is $\Gamma^{-1}\left(-1,0\right)$ which is meant to be diffuse or non-informative and acts as a uniform distribution across the range of "negative infinity to positive infinity" (Muthén & Muthén, 1998-2017, pg. 775). Because the population values for these parameters are so small, this prior that is meant to be non-informative or diffuse may instead be influencing the variability such that the random effect variances have slightly higher efficacy values relative to their fixed effects counterparts, which in turn leads to having more out-of-range efficacy values.

Power/Type-I Error Rate for Fixed Effects. Because random effect variances are always positive, non-zero values, power is necessarily always 100% and the Type-I error rate cannot be calculated. Therefore, only power and Type-I error rates for the fixed effects will be discussed in this section.

Power to detect significant actor and partner fixed effects from the multivariate models is shown in Figure 28. While the actor effect has adequate power across all conditions (top row of Figure 28), the partner effect does not reach the 80% threshold for adequate power when the sample size is N = 50 dyads (bottom row of Figure 28, first two columns) but has a slight advantage when the outcome has a uniform distribution. The Type-I error rate for actor and partner fixed effects is shown in Figure 29. Values fall mostly within the range of acceptable values around 0.05 (0.025 to 0.075), though are slightly too high for the Y1 partner effect when the outcome has a skewed distribution and T = 14 timepoints; this is true whether the sample size is N = 50 or N = 200 dyads, though is more pronounced (over 8% versus right around 8%) at the larger sample size.

For univariate models, the power to detect significant autoregressive effects is adequate across almost all conditions (Figure 30). The one exception is when T = 14 timepoints and N = 50 sample size for a skewed outcome distribution (top-left cell in Figure 30), where power is slightly below the 80% threshold. Because there was only one autoregressive effect, which was specified to be non-null, I cannot directly measure the Type-I error rate for this parameter. However, as mentioned in the methods section above, I can *approximate* the Type-I error rate by calculating 1 – coverage for that parameter; the results for this estimated Type-I error rate are shown in Figure 31. These values are estimated to be well above the acceptable range when sample size is large and number of timepoints is T = 14 (see bottom row of Figure 31). It is important to keep in mind, however, that because the autoregressive effect was biased (see Figure 13, note large bias for T = 14 timepoints and N = 200) the coverage rates are likely biased as well, which in turn could be inflating the estimated Type-I error rates shown in Figure 31.

Power to detect significant multivariate level-1 or level-2 covariate fixed effects is shown in Figure 32. Again, the larger sample size of N = 200 dyads shows adequate power for all conditions, while the smaller sample size of N = 50 dyads falls below the 80% threshold for T = 14 and T = 28 timepoints for both level-1 and level-2 covariates. Looking at the largest timepoint condition (T = 56) for N = 50 dyads, the level-1 covariate exceeds the 80% power threshold when the outcome has a uniform distribution, and comes very close for the skewed distribution. The level-2 covariate, however, does not reach 80% power for any of the N = 50 dyad conditions. This is not surprising, however, considering level-2 effects have a much smaller total number of observations (just N = 50) to detect an effect relative to the total number of observations to detect a level-1 effect (N multiplied by T), keeping power low even in conditions with more timepoints. Conversely, the Type-I error rates for level-1 and level-2 null covariate effects shown in Figure 33 fall within the acceptable range of values across all conditions.

The power to detect significant univariate level-1 or level-2 covariate fixed effects is shown in Figure 34, and largely follow the trends shown in Figure 32. The Type-I error rates for level-1 and level-2 covariates shown in Figure 35 also closely resemble the results displayed in Figure 33, with the exception of the level-1 covariate having a Type-I error rate that is slightly too low when sample size is N = 50 and the outcome has a uniform distribution (see top row, second cell of Figure 35). Overall, these results show that the autoregressive or actor effects have the most power across conditions, with partner effects and level-1 covariate effects typically achieving adequate power when sample size is small but number of timepoints is large, and level-2 covariates failing to meet the 80% power threshold across any conditions when sample size is N = 50.

2-Category Results Summary. Overall, the 2-category results provided support for the hypotheses for both research questions 1 (how do the number of clusters and number of timepoints affect estimates?) and 2 (how does the proportion of responses on the outcome affect estimates?). The key takeaways from the results for the conditions that had a binary outcome can be summarized as follows:

- Number of timepoints had a greater impact on the autoregressive effects than the sample size; T = 14 timepoints and some T = 28 timepoint conditions showed bias in both the fixed effect and random effect variances
- Both number of timepoints and sample size had an impact on the partner and both level-1 and level-2 covariate effects; more bias, lower power, and higher efficacy values tended to be in the T = 14 and 28 timepoints and N = 50 sample size categories

- When there were differences between skewed and uniform distributions, typically skewed distribution showed more bias or less power. An exception is efficacy, where uniform distribution tended to have higher values

3-Category Models

For the 3-category models, approximate normal/symmetric distributions of responses were specified along with skewed and uniform, and replications were fit as both a categorical and continuous DSEM. As a result, the continuous covariate effects were standardized across models and conditions to facilitate comparison between continuous and categorical methods. Additionally, as with the 2-category models, efficacy results are presented in lieu of coverage due to bias in parameter estimates.

Bias for Fixed and Random Effect Variances. The average relative or absolute bias for the multivariate actor and partner fixed effects are shown in Figure 36. In these figures (and subsequent figures in the 3-category section), results from the models fit as categorical DSEM are shown with a solid line, those fit as a continuous DSEM are shown with a dashed line. Each row represents a different effect (in this case, Y1 actor effect, Y1 partner effect, Y2 actor effect, Y2 partner effect, respectively); the first three columns are for the normal, skewed, and uniform distribution conditions for the N = 50 sample size conditions, and the last three columns are for the normal, skewed, and uniform distribution conditions for the N = 200 sample size conditions, respectively. The population value is shown in the upper-right corner of each cell.

As with the 2-category model results (Figure 14), the null actor and partner effects have little to no bias across conditions (see the middle two rows of Figure 36). Additionally, for the non-zero actor effect ("Y1 Act", top row of Figure 36), once again for the categorical model, number of timepoints appears to have a greater effect on bias than sample size or distribution of responses on the outcome. When the number of timepoints is small (T = 14) there is the most bias in the autoregressive actor effect. This trend is generally true for the non-zero partner effect's categorical results ("Y2 Part", solid lines across bottom row of Figure 36) as well. When considering the bias in the continuous model results, however, a different pattern emerges for both the non-zero actor and partner effects. First, all of the average bias values are *negative*, contrasted with the positive bias values presented up to this point. This is because the autoregressive and cross-lagged effects are attenuated in continuous models when the outcomes are categorical; for the autoregressive effects specifically, this is because the categorical model treats these effects as polychoric correlations while the continuous model will assess them as Pearson correlations. While the cross-lagged / partner effects are not truly correlations (because the variances are not equal), the categorical nature of both variables still leads to an attenuation of the effect when fit as a continuous model. This attenuation also leads to an interesting pattern for the non-zero actor effect when the number of timepoints is T = 14: it would appear that for most conditions across the top row of Figure 36, the continuous model actually shows *less bias* on average than the categorical model (see for example, the second and fifth columns of the top row at T = 14 timepoints). While it may be tempting to believe that the continuous model is better able to recover the parameters of the generating model in those cases, I believe it instead to be an artifact of being an attenuated version of a positively biased effect. This belief is reinforced by the general downward trajectory of the continuous bias values within each cell for the actor effect, mirroring the categorical results, suggesting that if the categorical model had instead showed no bias in those conditions, the continuous model would then reflect a larger negative bias.

Another noticeable result when examining the continuous model results is the differences between the distribution conditions: for both the non-zero actor and partner effects (first and fourth row of Figure 36), the normal distribution conditions ("NO", first and fourth columns) show the most bias, followed by the skewed distribution conditions ("SK", second and fifth columns), with the uniform distribution conditions showing the least amount of bias generally ("UN", third and sixth columns). This is especially prominent in the partner effect in the bottom row of Figure 36, while the actor effect has more similar bias values for skewed and uniform. In fact, the average absolute bias for the partner effect in the continuous framework appears to only be influenced by distribution on the outcome, and not by timepoints or sample size – notice how the lines are almost flat, and follow the same respective patterns when N = 50 and when N = 200. The non-zero actor effect, on the other hand, appears to be affected by both distribution and number of timepoints, though not sample size – the dashed lines in the top row of Figure 36 have a downward trajectory that is about the same when N = 50 as when N = 200 across conditions.

The autoregressive or carryover fixed effect from the univariate model has average relative bias values displayed in Figure 37. These results closely mirror those of the non-zero actor effect from Figure 36, almost down to the average bias values themselves for both the categorical and continuous model results.

Turning to the level-1 and level-2 covariate fixed effect bias for the multivariate models (results shown in Figure 38), recall that the posterior median values used to calculate bias, along with the population values themselves, have all been standardized to allow for a more direct comparison between the categorical and continuous model results. The null covariate effects (level-1 and level-2 covariate effects for Y2, second and fourth rows of Figure 38) show essentially no bias for both the continuous and categorical models, across all conditions. The non-zero level-2 covariate effect ("L2 Y1", third row of figure) also shows very little bias for the categorical models across conditions. This parameter does have slightly more negative bias across conditions for the continuous models, with the most prominent differences being when the outcome has a skewed distribution (see the second and fifth columns for "L2 Y1", fourth row). There do not appear to be trends across number of timepoints or sample sizes for this effect. The non-zero level-1 covariate effect, however ("L1 Y1", top row of Figure 38), shows effects of timepoints and sample size in the categorical results and of timepoints, sample size, and distribution on the outcome in the continuous results. Specifically, the categorical results show more bias when number of timepoints is $T = 14$ and when sample size is $N = 50$ (first three columns of top row), and shows almost no bias for $T = 28$ and 56 timepoints for $N = 50$, and for all timepoints when $N = 200$. The continuous model results, however, show prominent negative bias in each cell, and appear to show the same attenuation effects from Figure 36: the trajectory of the bias for the dashed-line continuous conditions follows the same general pattern of the categorical, solid-line conditions across the conditions in the top row. However, the attenuation is smaller here (i.e., the distance between the dashed line and solid line is smallest) when the outcome has a uniform distribution (see the third and sixth columns, first row). Despite the categorical model results when $N = 200$ showing almost no bias across conditions for the level-1 covariate fixed effect, the continuous model results show the *least* amount of bias when the outcome's distribution is uniform.

As with the actor and partner effects, the univariate level-1 and level-2 covariate fixed effect biases (shown in Figure 39) follow similar patterns to the multivariate results discussed above. There is arguably slightly more bias in the non-zero level-1 covariate effect ("L1 Cov1", top row of Figure 39) for the categorical model when the outcome has a normal distribution relative to the other distributions at T = 14 timepoints and N = 50, though this difference is small. The continuous model results still show the least amount of bias in the level-1 covariate when the distribution on the outcome is uniform, and shows the greatest difference in bias values from categorical results for the level-2 non-zero covariate effect ("L2 Cov1", third row of Figure 39) when the distribution on the outcome is skewed, regardless of other simulation factors.

Overall, the bias results for the fixed effects suggest that, when 3-category data are fit using a categorical DSEM, small timepoints leads to more bias for the non-zero autoregressive/actor and partner effects and level-1 covariate effects, and small sample size also leads to increased bias for level-1 covariate effects. When 3-category data are fit using a continuous DSEM, all non-zero fixed effects examined (autoregressive/actor, partner, level-1 covariate, and level-2 covariate) are affected by the distribution of responses on the outcome; namely, bias tends to be larger when the distribution is skewed or normal, and smaller when the distribution is uniform. The autoregressive/actor effects also show more bias when the number of timepoints is small for the continuous model.

When considering the random effect variance bias for the 3-category models, the covariate effects presented an issue. Because the level-1 fixed effects were presented on a standardized metric, the random effect variances would also potentially need to be standardized in some way to make them more comparable. Instead, I opted to display bias results on the raw scale for the random effect variances. This has the benefit of not only contextualizing the level-1 covariate random effect variance values across the categorical and continuous models, but also contextualizing those values for the autoregressive/actor and partner effects as well. As mentioned in the 2-category model section, the bias metric for small values can be misleading (e.g., a 40% bias might not be a large increase on the raw metric for a random effect variance).

The average values of the random effect variance for the multivariate actor and partner effects across conditions is shown in Figure 40. For these figures, instead of the dashed line indicating the "no/zero bias" mark, it instead marks the population value (which is the same value relative to the average random effect variance values). For these parameters, there is one actor effect and one partner effect with a population value of 0.1 ("Y1 Act" and "Y1 Part", first two rows of Figure 40) and the other actor and partner effects have population values of 0.05 ("Y2 Act" and "Y2 Part", bottom two rows of Figure 40). In general, the actor random effect variances for categorical models (solid black lines) show almost no bias at T = 28 and 56 timepoints, and slight bias when T = 14 timepoints – this bias at T = 14 timepoints is approximately the same in terms of relative bias values (between 10 and 30% for at T = 14 timepoints across conditions), but looks smaller for the Y2 actor effect (third row of Figure 40) because the population value is smaller. For both actor effects, sample size and outcome distribution does not appear to have an effect on the random effect variance parameters in the categorical models.

When looking at the continuous models' values for these same two parameters, there is an effect of both timepoints and response distribution: there is an attenuation effect similar to the actor fixed effect (top row, Figure 36, compared with top row of Figure 40), but this attenuation is larger when the distribution of responses is normal (the first and fourth columns, top row of Figure 40). This effect is similar for the Y2 actor effect (third row of Figure 40), but because the variance values cannot be zero or negative there is less difference between the continuous and categorical models' results across conditions (i.e., there is still attenuation, but there is a floor effect to that attenuation). The gap between random effect values from the categorical and continuous models is still slightly larger when the distribution is normal (first and fourth columns, third row of Figure 40) relative to other distribution conditions.

The two partner effects (second and fourth rows of Figure 40) show an effect of timepoints, sample size, and response distribution for the categorical model results, but primarily response distribution only for the continuous model results. For both partner effects, the random effect variance values in the categorical models are too high when the number of timepoints is small, and this is more pronounced when the sample size is small (N = 50, first three columns) or when the outcomes have normal or skewed distributions ("NO" or "SK", first and second columns). The T = 14 timepoint conditions when N = 200 for the skewed distribution (column five) also shows a difference, which again has a higher bias than other distributions at that same timepoint and sample size (last three columns). The continuous results primarily show that the random effect variance is most attenuated when the distribution is normal; this is most prominent when the population value is 0.10 (Y1 partner effect, second row) rather than 0.05 (Y2 partner effect, fourth row), which makes sense given how much closer to 0 the smaller population value is.

The random effect variance bias / values for the univariate autoregressive effect across conditions are shown in Figure 41. As with the multivariate actor effect with the same population value (0.10, see top row of Figure 40), these conditions show differences across timepoints only for the categorical models: T = 14 timepoints shows the largest difference from the population value. They also show the same differences across timepoints and distribution conditions for the continuous models: the trajectories follow the same patterns within-cells, across timepoints, and have a larger gap between the categorical results when the outcome distribution is normal (see first column of Figure 41).

Finally, the results for the level-1 covariate random effect variances for the multivariate models are shown in Figure 42, and for the univariate models are shown in Figure 43. Notably in these two figures, the multivariate level-1 covariate random effect variance population values are 0.05 (top row, Figure 42) and 0.10 (bottom row, Figure 42), while the univariate level-1 covariate random effect variance population values are 0.05 (top row, Figure 43) and 0 (bottom row, Figure 43). Interestingly, all four level-1 covariate effects show the same patterns across conditions for both categorical and continuous models. The categorical models tend to overestimate the random effect variance when sample size is small (N = 50, first three columns of both figures), and when number of timepoints is small (T = 14, and T = 28). Even the T = 56 timepoint condition when the sample size is small tends to have either the same or slightly higher estimated values compared to the T = 14 timepoint conditions when N = 200 (compare last point of lines in first three columns to the first point of lines in last three columns for each figure). For the categorical model results, there also seems to be less bias at T = 14 timepoints and N = 50 sample size when the outcome also has a uniform distribution, relative to the normal and skewed distributions (third column versus first and second columns of both figures).

The continuous model results for these all show an effect of outcome distribution, but not of sample size or timepoints. Looking across the N = 50 and N = 200 sets of conditions (first three columns, last three columns), there is a step-like pattern to the random effect variance values such that the normal distribution has the lowest values, then the skewed distribution, and the uniform distribution has values closest to the population value (though still quite a bit below). An exception here is the “L1 Cov2” in the univariate models (bottom row, Figure 43), which shows the continuous results at about zero across all conditions. Again, this is because the value for that parameter cannot be below zero (or at exactly zero), and so the continuous model cannot underestimate that parameter to the same extent as the other conditions; this is also why the random effect variance with the largest population value (0.10, bottom row of Figure 42) shows a larger discrepancy between the population value and the results from the continuous models – that parameter has more room in the lower range of potential values to be underestimated relative to the 0.05 population value, and certainly relative to the 0 population value conditions.

Overall, the bias for the random effect variances showed effects of timepoints for all parameters in the categorical models, where smaller timepoints led to a larger overestimation of the effect. The partner and level-1 covariate random effect variances also showed differences with respect to sample size and response distribution in the categorical models, where smaller sample sizes showed greater overestimations and a uniform distribution showed in general smaller overestimations. Within the continuous models, all parameters showed an effect of response distribution wherein normally distributed outcomes saw the greatest degree of underestimation of the effect, and uniformly distributed outcomes showed the lowest degree of underestimation. The autocorrelation/actor effects also showed an effect of timepoints, where the continuous models' average median values were highest when number of timepoints were small (T = 14) and lowest when number of timepoints were large (T = 56).

Efficacy for Fixed and Random Effect Variances. Overall, the efficacy results for the multivariate actor and partner effects (Figure 44) and the univariate autoregressive effect (Figure 45) show largely "well-behaved" values that do not go above 1.1 or below 0.9 for both the categorical and continuous models. There are a couple conditions where the efficacy value goes slightly above 1.1, but never by much. Importantly, as seen in the 2-category results above, efficacy can be affected by the number of replications are run for a simulation condition (more replications lead to more posterior median values making up the distribution of values for the denominator of the equation, and more posterior standard deviation values per parameter to smooth out the contribution of any outliers in the numerator), so these slight deviations above 1.1 are not concerning.

When considering the efficacy results for the level-1 and level-2 covariate fixed effects for the multivariate (Figure 46) and univariate (Figure 47) models, the level-2 covariate fixed effects (bottom two rows for both figures) are also largely well-behaved, with any values outside of the 0.9 to 1.1 range being slight. When looking at the level-1 covariate fixed effect efficacy results, the categorical model values are all mostly within range, while the continuous model values tend to be larger and across more conditions. For example, the second level-1 covariate in the univariate model (“L1 Cov2”, second row of Figure 47) shows the dashed categorical model line falling beyond the 0.9 – 1.1 range for ten conditions of the 18 total shown in that row (three conditions per cell), and another three conditions falling slightly above the acceptable range. This is again does not seem to be particularly problematic, as having only 297 replications per cell is most likely contributing to this pattern, along with the default prior distribution assigned to that parameter. The M*plus* default priors for these level-1 covariate fixed effects are a normal distribution prior with a mean of 0 and variance of “infinity” (clarified as being a value of 10^{10} - see Muthén & Muthén, 1998-2017, page 775). Having a prior distribution with a large variance likely contributes to the efficacy results being larger and more out-of-bounds for the continuous conditions, though overall these values did not reach problematic levels.

Moving to the efficacy results for the random effect variances of the multivariate actor and partner effects (Figure 48) and univariate autoregressive effect (Figure 49), in general a different pattern emerges compared to the fixed effects: here, the categorical models tend to have higher efficacy values than their continuous model counterparts. Number of replications again probably contributes somewhat to these large efficacy values, along with the default prior for the random effect variances – they all get the $\Gamma^{-1}\left(-1,0\right)$ prior distribution, regardless of whether the model is categorical or continuous. This prior distribution, as discussed in the efficacy results for the 2-category models, can be acting as an unintentionally informative prior for these conditions and contributing to an overestimation of the posterior standard deviation for the random effect variance parameters. However, that is true for both categorical and continuous and therefore would not lead to the categorical efficacy results being *larger* than the continuous ones within the same conditions. A possible explanation is that, because the random effect variances themselves tend to be *underestimated* when the outcome is treated as continuous (see Figures 40 and 41), the random effect variances around these parameters have to be smaller to keep the range of values from going below 0. This could, in turn, be counter-acting the inflation of efficacy values due to the smaller number of replications and the default prior distribution, leading the continuous models' efficacy values for the random effect variances to look "okay" (or at least closer to the range of acceptable values) while the categorical models' values remain high.

Finally, the efficacy results for the random effect variances for the level-1 covariates for the multivariate model are shown in Figure 50 and for the univariate model are shown in Figure 51. As with previous efficacy results, these values go outside of the 0.9 to 1.1 range, but these deviations can again be explained by other factors in the simulation rather than a problematic degree of overestimation or underestimation of the variability of these parameters. This assertion might feel out of place for the "L1 Cov2" results for the univariate model (bottom row, Figure 51). While these values at first seem out of control, the population value for this random effect variance parameter is 0. As a result, these efficacy values are trying to assess over/underestimation of incredibly small values: for example, the spike in the N = 200, normal distribution, T = 28 timepoints condition (fourth column) is because the average standard deviation across the 297 posterior distributions for that parameter was 0.0004 and standard deviation of all the medians of the 297 posterior distributions for that parameter was 0.0001. Arguably, this condition's efficacy results should perhaps be ignored.

Overall, the efficacy results for the fixed effects were largely within the acceptable range of values, while the efficacy results for the random effect variances were more diverse. The number of replications, the use of the M*plus* default prior distributions, and the differences in prior distributions when the model is treated as categorical rather than continuous likely all contributed to the diversity of random effect variance results, along with the population values themselves being very small and therefore difficult to estimate variability around.

Power/Type-I Error Rate for Fixed Effects. The power to detect a significant actor or partner effect in a multivariate model is shown in Figure 52, and to detect a significant autoregressive effect in a univariate model is shown in Figure 53. The actor effect from the multivariate model (top row of Figure 52) and the autoregressive effect from the univariate model show similar patterns: all conditions have exceeded the 80% threshold for power, with the largest difference between the categorical and continuous models when sample size is N = 50, number of timepoints are T = 14, and the outcome has a skewed distribution (see second column, top row for both Figures 52 and 53). The partner effect for the multivariate model (bottom row, Figure 52) shows adequate power when sample size is N = 200, but at the smaller sample size only crosses the threshold at T = 56 timepoints for the normal and skewed distributions, or at T = 28 timepoints for the uniform distribution conditions. This is in contrast to the power for the same partner effect in the 2-category model (see bottom row, Figure 28) which never achieved adequate power at the smaller N = 50 sample size, even when T = 56 timepoints. Partner effects tend to be smaller than actor effects so they generally take more power to detect, and having the additional information provided by 3 categories instead of 2 likely contribute to the higher power values for the partner effect in these results. The larger variability in responses in the uniform distribution (over the skewed and normal distribution conditions) also likely contributes to that condition achieving over 80% power when T = 28 timepoints instead of T = 56 timepoints.

Power to detect significant level-1 and level-2 covariate effects for a multivariate model are shown in Figure 54 and for a univariate model are shown in Figure 55. Here again, both level-1 and level-2 covariate effects have adequate power across all conditions when sample size is large (N = 200; see last three columns of Figures 54 and 55). The level-1 covariate when sample size is small (N = 50) looks very similar in both the univariate and multivariate conditions (top row, first three columns of Figures 54 and 55) and to the power for a partner effect when N = 50 (first three columns, bottom row of Figure 52). Adequate power is achieved at this smaller sample size when timepoints are large (T = 56) for the normal, skewed, and uniform distributions in the multivariate model for both categorical and continuous conditions (see first three columns, top row of Figure 54). For the univariate model, adequate power is achieved at the smaller sample size when T = 28 timepoints for the normal distribution, continuous condition, as well as for both the categorical and continuous uniform distribution conditions, and at T = 56 timepoints for all other conditions (skewed categorical and continuous, normal categorical; see first three columns, top row of Figure 55). The level-2 covariate effects never achieve adequate power when sample size is small, though as mentioned in the 2-

category results section, this is likely due to the smaller sample size available to detect level-2 effects relative to level-1 effects.

Turning to Type-I error rates for actor and partner effects in the multivariate model (Figure 56) and the estimated Type-I error rate for the autoregressive effect in the univariate model (Figure 57), the differences between categorical and continuous models is apparent. Across all multivariate conditions, the Type-I error rate remains within a relatively reasonable range for all categorical conditions (see solid lines in Figure 56, both rows). When sample size is large (N = 200) the Type-I error rates for the actor and partner effects in the continuous model conditions becomes much larger than the upper threshold of acceptable values, reaching almost 40% when N = 200, T = 56, and the outcome has a skewed distribution for the Y2 actor effect (fifth column, second row of Figure 56). This is problematic if model results will be evaluated by significance of effects alone, and suggests that treating a 3-category outcome as continuous for these conditions can lead to researchers claiming effects that do not actually exist in the population.

The estimated Type-I error rate for the univariate autoregressive effect (Figure 57) needs to be interpreted in the context of the bias in the autoregressive effect (see Figure 37). Because the estimated Type-I error rate is calculated using coverage, it is important to remember that coverage can be misleading if the estimate is biased; in this case, because the categorical autoregressive effect has a positive bias in all conditions when T = 14 timepoints (see first point of all lines in Figure 37), then coverage (which assesses how many posterior distribution credible intervals contain the population value) will be *lower* than expected if the parameter was unbiased. Then, because estimated

Type-I error is calculated as (1 – coverage), having a lower coverage value in turn leads to a *higher* estimated Type-I error rate. This is on display in Figure 57, where the Type-I error rate that goes above the upper boundary of 0.075 for the categorical models when N = 200 and T = 14 (bottom row), but is back within the 0.025 to 0.075 boundaries when T = 28 and 56, at the same time that the autoregressive effect no longer showed bias in Figure 37. The same logic can be applied to the continuous results: here, because the autoregressive effect was consistently underestimated, coverage values were again lower than would be expected if the effect was unbiased. This was more extreme in the continuous conditions (see dashed lines in Figure 37) and bias became larger as timepoints increased, leading to the steep increase in estimated Type-I error rates within most cells of Figure 57 for the continuous conditions. The closest to "unbiased" the continuous conditions come in Figure 37 are when distribution = "SK", and T = 14 (see both rows, second column of Figure 37) – the corresponding estimated Type-I error rates for those conditions fall within the range of acceptable values (see second column, both rows of Figure 57).

Finally, the Type-I error rates for the level-1 and level-2 covariate effects are shown in Figure 58 for the multivariate model and Figure 59 for the univariate model. Unlike the autoregressive and cross-lagged effects, these parameters show acceptable Type-I error rates across all sample size, timepoint, and distribution conditions. Unlike the discrepancy between level-1 and level-2 effects in the power results (bottom row of Figures 54 and 55), here the level-1 and level-2 effects show similar rates across conditions.

Overall, power results suggest that autoregressive effects are able to be detected when using both a continuous and categorical DSEM, even when sample size and number of timepoints are low. There are issues with an inflated Type-I error rate for these effects, however, when sample size is large and the model is fit as a continuous DSEM. The cross-lagged/partner effects showed low power unless sample size was large or number of timepoints was 56 (or 28 for a uniformly distributed outcome), and also had inflated Type-I error rates for large sample size conditions with a continuous DSEM. The level-1 and level-2 covariate effects avoid the inflated Type-I error rate issue, but also saw inadequate power levels when the sample size was small; while the level-1 covariate conditions reached 80% when number of timepoints was 28 or 56 across conditions, the level-2 covariate effect was never adequately powered, reaching only around 60% in the highest-power condition.

3-Category Results Summary. Overall, the 3-category results provided support for the hypotheses for research questions 1 (how do the number of clusters and number of timepoints affect out estimates?), 2 (how does the proportion of responses on the outcome affect out estimates?) and 4 (how does treating the 3-category outcomes as continuous compare to using a categorical model?). The key takeaways from the results for the conditions that had a 3-category outcome can be summarized as follows:

- Categorical model conditions were less biased and had lower Type-I error rates than their continuous model counterparts. Power was typically similar between the two model conditions, and did not conclusively favor the categorical or continuous models

- Number of timepoints had a greater impact on the autoregressive effects than other simulation conditions; T = 14 timepoints showed bias for both the fixed effects and random effect variances for the autoregressive effects.
- An exception to the autoregressive effects trend is sample size and outcome distribution for Type-I error rates of continuous models, for which N = 200 sample size and the skewed distribution conditions also affected how inflated the rate was along with number of timepoints.
- Partner/cross-lagged effects' and level-1 and level-2 covariate effects' performance on the evaluation criteria varied more across sample size and response distribution categories than the autoregressive effects, with lower sample sizes and skewed distribution conditions performing worse (more bias, less power, higher Type-I error rates, etc.)
- Parameters in the continuous model conditions were primarily affected by response distributions, with the uniform distribution conditions performing better (less bias, more power, etc.) than both normal and skewed distribution conditions. Sample size also affected power results, with larger sample sizes having higher power values in all continuous model conditions

5-Category Models

The 5-category model conditions matched those described above in the 3-category model section: approximate normal/symmetric, skewed, and uniform distributions of responses; replications were fit as both a categorical and continuous DSEM. The continuous covariate effects were again standardized across models and conditions, and efficacy results are presented in lieu of coverage due to bias in parameter estimates.

Bias for Fixed and Random Effect Variances. The average relative or absolute bias for the multivariate actor and partner fixed effects are shown in Figure 60, and the average relative bias for the univariate autoregressive fixed effect is shown in Figure 61. Once again, the null actor and partner effects have little to no bias across conditions (see the middle two rows of Figure 60). For both the non-zero actor effect ("Y1 Act", top row of Figure 60) and autoregressive effect (Figure 61), number of timepoints appears to have an effect on bias for the categorical conditions, though in this case the distribution of responses on the outcome also seem to have an effect. When the number of timepoints is small (T = 14) not only is bias highest in the autoregressive/actor effects, the skewed distribution also shows higher average bias values compared to both the normal and uniform distributions. This difference between distributions is not apparent at T = 28 or 56 timepoints, and sample size does not appear to influence bias for these effects. For the non-zero partner effect's categorical results ("Y2 Part", solid lines across bottom row of Figure 60), smaller timepoints again generally mean more bias across conditions, though sample size and response distribution do not appear to be influential.

When considering the bias in the continuous model results, however, a different pattern emerges for both the non-zero actor/autoregressive and partner effects. While the attenuation that was present in the 3-category results is here as well, so is the autoregressive effect's tendency to have a larger discrepancy between categorical and continuous conditions when the outcome is normal than when it is skewed or uniform (top row, first and fourth column of Figure 60, first column of Figure 61). Interestingly, both the multivariate and univariate autoregressive effect show the relative bias values being closest when the outcome has a skewed distribution. For the partner effect, however, the largest discrepancy between categorical and continuous conditions is when the outcome is skewed, and the number of timepoints appears to also have an effect on the continuous results (note the negative slope in most of the dashed lines, bottom row of Figure 60). This is different from the 3-category outcome results that showed the largest negative bias in continuous conditions when the distribution was *normal*, and showed no effect of timepoints (bottom row, Figure 36). Here in the 5-category results, the non-zero partner effect shows very similar bias results in the continuous model for the normal and uniform distribution conditions.

The average absolute or relative bias results for the standardized level-1 and level-2 covariate fixed effects are shown in Figure 62 for multivariate model conditions and Figure 63 for univariate model conditions. The null effects (middle two rows of Figure 62, second and fourth rows of Figure 63) continue the trend of being unbiased for both categorical and continuous models. The non-zero level-2 effects (bottom row of Figure 62, third row of Figure 63) follow similar trends both to each other and to their analogous 3-category model results (see Figure 38, third row for multivariate and Figure 39, third row for univariate effect). Namely, both the categorical and continuous conditions show very little bias, but the largest difference between average bias values for these conditions is when the outcome has a skewed distribution. There also again appears to be little to no effect of timepoints or sample size.

The non-zero level-1 covariate effects (top rows of Figures 62 and 63) show an effect of timepoints in the categorical conditions when N = 50 dyads for the multivariate model (Figure 62) where T = 14 timepoints shows more bias than others; this does not appear in the univariate model, which shows almost no bias for the categorical conditions across sample size, timepoints, and outcome distribution conditions. When considering the performance of the continuous model conditions, both univariate and multivariate models show the most negative bias when the outcome is skewed (second and fourth columns of top row for both Figures 62 and 63), but no discernable difference in bias results between the normal and uniform distributions. There also appears to be a small effect due to number of timepoints when the sample size is N = 50 dyads in the multivariate model results, where the bias for the continuous model results follow a similar downward trajectory as the categorical results across the three outcome distribution conditions (first three columns, top row Figure 62). This effect is not apparent in the univariate results, which show no difference in bias across sample size and number of timepoints for the continuous conditions (top row, Figure 63).

Overall, the bias results for the fixed effects suggest that, when 5-category data are fit using a categorical DSEM, small timepoints leads to more bias for the non-zero autoregressive/actor and partner effects, as well as for multivariate level-1 covariate effects when sample size is small. When 5-category data are fit using a continuous DSEM, all non-zero fixed effects examined (autoregressive/actor, partner, level-1 covariate, and level-2 covariate) are affected by the distribution of responses on the outcome; namely, bias tends to be larger when the distribution is skewed (for partner, level-1, and level-2 covariates) or normal (for autoregressive/actor). While the actor/autoregressive and level-2 covariate effects' bias trends are similar to the 3-category model, the partner and level-1 covariate effects' bias trends show that the normal distribution conditions have less bias here than in the 3-category model: while they more closely resemble the skewed distribution results in the 3-category models, here they more closely resemble the uniform distribution results. The autoregressive/actor effects also show more bias (around their presumed attenuated value) when the number of timepoints is small for the continuous model.

The random effect variance results for the 5-category models are shown on the raw scale rather than as average relative bias values. The average values for the multivariate actor and partner random effect variances are shown in Figure 64, and values for the univariate autoregressive random effect variance are shown in Figure 65. The average values for the multivariate level-1 covariate random effect variances are shown in Figure 66, and values for the univariate level-1 covariate random effect variances are shown in Figure 67. I present these altogether here because the results are extremely similar to the trends covered in the 3-category model results for the same parameters (Figures 40 to 43 corresponding to Figures 64 to 67, respectively). There is a similar (if slightly less extreme) effect of both timepoints and outcome distribution for all parameters in the categorical models, where smaller timepoints led to a larger overestimation of the effect and this is particularly true when the outcome is skewed. The partner and level-1 covariate random effect variances again show smaller sample sizes lead to larger overestimations.

Within the continuous models, parameters tended to show an effect of response distribution wherein normally distributed outcomes saw the greatest degree of underestimation of the effect (this is particularly visible when the random effect variance is larger – see top two rows of Figure 64 where the population value is 0.10, versus the bottom two rows where the population value is 0.05). The autoregressive/actor random effect variances also show an effect of timepoints (T = 14 timepoints has highest values relative to T = 28 and T = 56), particularly when the outcome is skewed or uniform. The largest discrepancies between population values and average values are for continuous models when the partner random effect variance is larger (0.10, dashed lines in second row of Figure 64), regardless of other simulation factors, and for the non-zero level-1 covariate random effect variances in the normal distribution conditions (Figure 66, dashed lines in both rows, and in top row of Figure 67). In all of these cases, the continuous models' average values for the random effect variance is about half that of the population value.

Efficacy for Fixed and Random Effect Variances. The efficacy results for the multivariate actor and partner fixed effects are shown in Figure 68, and for the univariate autoregressive fixed effects in Figure 69. Efficacy results for the multivariate and univariate covariate fixed effects are shown in Figures 70 and 71, respectively. As with the 3-category results, the efficacy for all of these conditions fall within an acceptable range of values, with deviations outside the 0.9 – 1.1 range likely being a result of sampling variability, the M*plus* default prior distributions assigned to the parameters, or the overall size of the effect being small.

A similar summary can be given for the efficacy results for the random effect variances. The multivariate actor and partner random effect variance efficacy values are shown in Figure 72, univariate autoregressive random effect variance efficacy values are shown in Figure 73, and the level-1 covariate random effect variance efficacy values for the multivariate and univariate models are shown in Figures 74 and 75, respectively. As with the 3-category results, efficacy values for the level-1 covariate random effect variance that had a population value of 0 are the largest across all parameters, because even small differences between the observed variability and estimated variability appear as large ratios when values are close to zero.

Power/Type-I Error Rate for Fixed Effects. The power to detect a significant actor or partner effect in a multivariate model is shown in Figure 76, and to detect a significant autoregressive effect in a univariate model is shown in Figure 77. The autoregressive effects have more than 80% power across all model, sample size, timepoint, and outcome distribution conditions (top row of Figure 76, all of Figure 77). The power to detect a significant partner effect is below 80% when sample size is N = 50 dyads and number of timepoints is T = 14 for all outcome distribution and model conditions, but within that category is lowest when the distribution is skewed (around 60% power, compared to around 70% for normal and uniform distributions; bottom row, first three columns of Figure 76). For both categorical and continuous model conditions, when N = 200 dyads at all timepoints or when N = 50 dyads and at least T = 28 timepoints, power is above 80% for the partner effect.

Power to detect significant level-1 and level-2 covariate effects for a multivariate model are shown in Figure 78 and for a univariate model are shown in Figure 79. Here for the first time, the level-2 covariate effect achieves at least 80% power at the N = 50 sample size: in the multivariate model (Figure 78, bottom row, first three columns), for all categorical model N = 50 conditions *except* when there is a skewed distribution and T = 14 timepoints. The continuous model results in those same three cells shows over 80% power by T = 56 timepoints for all outcome distribution conditions, as well as for T = 28 timepoints when the outcome is normal (first column, bottom row Figure 78). The univariate level-2 covariate effect (Figure 79, bottom row), despite having the same population value, never reaches 80% power in the smaller N = 50 sample size conditions. While these level-2 effects generally have less power (as discussed in previous sections), I think this discrepancy between multivariate and univariate model conditions here is likely the result of the non-zero level-2 covariate effect being the same variable that had the null level-1 covariate effect. This is only the case in the 5-category, multivariate models (the 5-category univariate level-2 effect is the same variable that has the level-1 effect, as is the case with the 2- and 3-category univariate and multivariate models). As a result, all of the variability in the covariate is attributed to the level-2 effect giving it higher power. It is likely that if this effect was specified for the same variable that had the non-null level-1 effect, power would be much lower.

When looking at the Type-I error rates for actor and partner effects in the multivariate model (Figure 80), the continuous model conditions again show inflated rates when sample size is N = 200 for both effects (last three columns). This is particularly true when the outcome distribution is skewed (fifth column), where Type-I error rates are above 20% for all timepoints for the null actor effect (bottom row) and above 10% for the partner effect (top row). Meanwhile, the categorical model conditions are within the 0.025 and 0.075 range of acceptable values for all sample size, timepoint, and outcome distribution conditions (solid lines in Figure 80).

The estimated Type-I error rates for the autoregressive effect (Figure 81) have to be viewed with respect to the bias for that parameter for the same model, timepoint, sample size, and outcome distribution conditions (see Figure 61). When the sample size is N = 200, a bias value of at least 10% (positive or negative) produces estimated Type-I error rates that are out of bounds for both the categorical and continuous model conditions. As mentioned in previous sections, however, this is because a biased parameter in either direction leads to lower coverage rates, which in turn leads to an increase in estimated Type-I error rate.

The Type-I error rates for level-1 and level-2 covariate effects are more straightforward than the estimated Type-I rates in Figure 81. All of these values are within bounds across all conditions. The multivariate models' level-1 and level-2 covariate effect Type-I error rates are shown in Figure 82; the univariate covariate effect rates are shown in Figure 83.

Overall, power results suggest that autoregressive effects are able to be detected when using both a continuous and categorical DSEM, even when sample size and number of timepoints are low. There are issues with an inflated Type-I error rate for these effects, however, particularly when sample size is large, the outcome has a skewed distribution, and the model is fit as a continuous DSEM. The cross-lagged/partner effects showed low power for both categorical and continuous conditions unless sample size was large or number of timepoints was at least 28, and also had inflated Type-I error rates for large sample size conditions and a skewed distribution with a continuous DSEM. The level-1 and level-2 covariate effects avoided the inflated Type-I error rate issues. The level-1 covariate effects however did have inadequate power when sample size was small (N = 50) and number of timepoints was also small (T = 14); the level-2 covariate effects showed inadequate power in the univariate model at all N = 50 sample size conditions, but only some N = 50 sample size conditions in the multivariate models with continuous outcomes (where the skewed distribution and T = 14 timepoint conditions had the lowest power).

5-Category Results Summary. Overall, the 5-category results provided support for my hypotheses given for research question 1 (how do the number of clusters and number of timepoints affect out estimates?), research question 2 (how does the proportion of responses on the outcome affect out estimates?), and research question 4 (how does treating the 5-category outcomes as continuous compare to using a categorical model?). The key takeaways from the results for the conditions that had a 5-category outcome can be summarized as follows:

- Power for within-level non-autoregressive effects required at least T = 28 timepoints when sample size was N = 50 to reach 80% for both the categorical and continuous conditions
- Type-I error rates for the autoregressive and cross-lagged/partner effects in continuous model conditions were inflated when N = 200, particularly for the skewed outcome distributions
- Number of timepoints had a larger and more consistent impact on categorical autoregressive fixed effects than on other parameters, with T = 14 timepoints having larger bias. The skewed outcome also had higher bias values at T = 14 timepoints than other outcome conditions
- Partner and level-1 and level-2 covariate fixed effects showed greatest differences between categorical and continuous model conditions when the outcome distribution was skewed; the autoregressive/actor fixed effects showed greatest difference when outcome distribution was normal/symmetric
- Random effect variances for the continuous partner effects across all simulation conditions, as well as for the continuous level-1 covariate effects when the

distribution is normal/symmetric, were consistently half of the population value (50% negatively biased) while the same categorical conditions were either unbiased or much closer to the population value

- Parameters in both categorical and continuous model overall showed less bias, higher power, lower Type-I error rates compared to 3-category parameters in the same conditions

All Categorical Models

While the previous sections touched on comparisons between categories while presenting results, their primary function was to cover results within each category condition. Here, I have collated the univariate and multivariate results across category conditions for the categorical DSEM models. Because these results have been presented separately above, I will focus here on a more concise overview of the differences between 2-, 3-, and 5-category models across parameters and evaluation criteria. This involves presenting only collated figures for bias, power, and Type-I error rate, as the efficacy values were all admissible. Additionally, because the estimated Type-I error rates were often affected by bias, only Type-I error rates calculated from null parameters are presented (i.e., the univariate autoregressive Type-I error rates will not be shown here). Finally, because the null effects showed no bias, only non-zero parameter effects will be compared in the bias figures.

Bias for Fixed and Random Effect Variances. The average relative bias results for all non-zero autoregressive/actor and partner fixed effects for univariate and multivariate models across all category conditions are shown in Figure 84. For this figure, the labels for each row signify the parameter (e.g., “Act sig” refers to the non-zero actor effect for multivariate models), then the sample size, then the outcome distribution condition. The light gray dashed-line shows results for the 2-category models (note that this line does not appear in the symmetric/approximate normal "NO" condition cells, first and fourth columns); the solid dark gray line shows results for the 3-category models, and the dashed black line shows results for the 5-category models. Overall, these trends indicate that in general, T = 28 and T = 56 timepoint conditions show little bias across parameters and other conditions. Within the T = 14 timepoint conditions, 2-category models have the largest relative bias, and while 3- and 5-category model conditions tend to be closer together than the 2-category model conditions, the 5-category models generally have the smallest relative bias.

The average relative bias for the standardized level-1 and level-2 non-zero covariate fixed effects for multivariate and univariate models across all category conditions are shown in Figure 85. In this figure, the labels for each row first signify whether the effect came from the multivariate model ("DYAD") or univariate model ("IND" for individuals), followed by the parameter (e.g., "L1 sig" refers to the non-zero level-1 covariate fixed effect), then sample size and outcome distribution conditions. The light gray dashed-line shows results for the 2-category models (again not in the "NO" condition cells, first and fourth columns); the solid dark gray line shows results for the 3-category models, and the dashed black line shows results for the 5-category models. Trends in this figure indicate that overall, when N = 200 sample size, there is little bias in any of the covariate parameters (last three columns). When sample size is smaller (N = 50, first three columns), 2-category models again tend to have higher bias values, along with there being generally higher bias when T = 14 timepoints and the outcome has a skewed distribution. The 3- and 5-category results overlap here much more than they did in the autoregressive and partner effect bias results.

The average values of the random effect variances for the autoregressive/actor and partner effects with the larger population value (0.10) are shown in Figure 86, and with the smaller population value (0.05) are shown in Figure 87. The average values of the random effect variances for the level-1 covariate effects (regardless of population value) are shown in Figure 88. In Figure 86, the overall trend suggests that the random effect variance for partner effects tends to be overestimated to a greater degree than actor effects when sample size is small (see third versus first row, first three columns), but they are more similar when N = 200 (last three columns). The carryover/autoregressive effect in the univariate model ("Carry", second row) also generally has a larger degree of overestimation of the random effect variance when N = 50 relative to the multivariate autoregressive effect (top row), though only when number of timepoints is small (T = 14). When the outcome has a skewed or uniform distribution (second, third, fifth, and sixth columns), the 2-category models tend to show more bias than 3- and 5-category models when T = 14 timepoints, particularly for the partner effect. These trends largely hold up when the random effect variance is smaller as well (i.e., 0.05; Figure 87).

In Figure 88, there is once again a label denoting whether the parameter was part of the dyadic or individual data simulation conditions, followed by the parameter (e.g., "L1 0.05" meaning the level-1 covariate random effect variance parameter with a population value of 0.05), sample size, and outcome distribution conditions. Once again, there is a larger overestimation of the random effect variance when T = 14 and N = 50 (first three columns), and 2-category models tend to overestimate to a larger degree while the 3- and 5-category models tend to be both closer together and closer to the population value.

Power/Type-I Error Rate for Fixed Effects. The power to detect a significant autoregressive/actor or partner effect across all categorical model conditions is shown in Figure 89, and to detect a significant level-1 or level-2 covariate effect across all categorical model conditions is shown in Figure 90. In Figure 89, note again that 2-category model results are not available for the approximate normal ("NO") outcome distribution conditions. Overall, Figure 89 shows that autoregressive/actor effects (which are generally larger in magnitude than partner effects) have higher power when N = 50. All three parameters had adequate power when N = 200 across all other conditions. 2-category models had the lowest power, followed by 3-, and then 5-category models have the highest power across parameters. Depending on the distribution on the outcome, the number of timepoints needed when N = 50 to have at least 80% power varied between T = 28 and T = 56 timepoints, importantly both higher than the modal length of daily diary studies reported by Gunthert & Wenze (2014), which was 14 days.

In Figure 90, the power to detect level-1 or level-2 covariates was above 80% again when N = 200. The level-2 covariate effects when N = 50 had the lowest power overall and never reached the 80% threshold, with the exception of the 5-category conditions in the multivariate model (second row, first three columns of Figure 80; black dashed line). As previously mentioned, however, this is likely due to the setup of the simulation in those conditions being different from the 2- and 3-category models for both the univariate and multivariate conditions, as well as from the 5-category univariate model conditions. For the level-1 effects when N = 50, 2-category models had the lowest power, followed by 3-category models, with 5-category models having the highest power (first and third rows, first three columns). While most conditions for the level-1 covariate effects achieved 80% power by T = 56 timepoints, very few had crossed that threshold by T = 28 timepoints, and none when T = 14 timepoints. Again, given the average length of daily diary studies being 14 days, these results indicate that those are insufficient to find significant level-1 or level-2 continuous covariate effects in a categorical DSEM unless the number of individuals or dyads is large.

The Type-I error rates for actor and partner effects are shown in Figure 91 and for level-1 and level-2 covariate effects are shown in Figure 92 for all categorical model conditions. As discussed in the results sections above for 2-, 3-, and 5-category models separately, all Type-I error rates were acceptable across sample size, timepoint, and outcome distribution conditions for each parameter.

All Categorical Results Summary. Overall, these collated categorical model results were summarized here to provide support for my hypothesis for research question 3 (how does the number of categories on the outcome affect estimates?). The key takeaways from the results for all the categorical model conditions can be summarized as follows:

- 5-category outcomes showed the least bias and had the most power on average over the 2- and 3-category outcome conditions
- 3- and 5-category outcomes tended to perform more similarly to each other than they did to the 2-category outcome conditions, but still showed differences especially when N = 50 and/or T = 14
- All outcome category conditions had acceptable Type-I error rates for null effects

All Continuous Models

This final results section contains the collated results for all continuous models. Only the 3- and 5-category conditions were run as continuous DSEM, so only those results will be presented here. As with the previous section, there will be no efficacy results or estimated Type-I error rates, and bias will be presented only for the non-zero effects.

Bias for Fixed and Random Effect Variances. The average relative bias results for all non-zero autoregressive/actor and partner fixed effects for all continuous model conditions are shown in Figure 93. Similar to the previous section, the labels for each row of this figure signify first the parameter (e.g., "Act sig" refers to the non-zero actor effect for multivariate models), then the sample size, then the outcome distribution condition. The solid dark gray line shows results for the 3-category models, and the dashed black line shows results for the 5-category models. When considering the pattern of results in Figure 93, it is important to remember that these are the average relative bias results for *attenuated* effects. So, as discussed in the 3- and 5-category model individual results sections, just because the continuous model results appear to be unbiased at some timepoints does not mean that the effect is truly unbiased – it is instead biased around an attenuated population value. However, this figure does help to delineate what sample size, timepoint, and distribution conditions make the most difference when treating a categorical outcome as continuous in a DSEM model.

For example, the 5-category results are more consistent than the 3-category results across sample size and distributions for the partner effect (bottom row) – they show slightly lower values when the distribution is skewed (and are likely more biased in those conditions), but overall stay more consistent than the solid gray lines denoting the 3-category results do across the same six conditions in the bottom row. Additionally, the 5-category results for the partner effect seem to suggest there would be more bias when the outcome is skewed, while the 3-category results suggest there would be more bias when the outcome is normal/symmetric. For the carryover/actor effects, it seems that both the 3- and 5- category conditions for the continuous models also might have a harder time accurately estimating those parameters when the distribution of responses on the outcome is normal/symmetric, particularly when the outcome has only three categories (first and fourth columns of top two rows). Generally, the sample size does not matter as much as the number of timepoints (for actor/autoregressive effects) or the distribution of the outcome (both actor/autoregressive and partner effects).

Figure 94 shows the average relative bias for the standardized non-zero level-1 and level-2 covariate effects (again denoted as “DYAD” for the multivariate models and “IND” for the univariate models). Unlike with the actor/autoregressive and partner effects, because these have been standardized they can be more easily compared to the dashed line at 0.0 on the y-axis that denotes no bias. Generally, it seems that the level-2 covariate effects have less bias than the level-1 covariate effects when the categorical outcomes are treated as continuous. The largest bias for the level-2 effects (second and fourth rows) are when the outcome distribution is skewed (second and fifth columns). Within the level-1 covariate results (first and third rows), the 3-category models show the largest (and roughly equal amounts of) bias when the outcome is either normal/symmetric or skewed (first, second, fourth, and fifth columns) while the 5-category models show the largest bias just when the outcome is skewed. For the 5-category level-1 effects, the normal/symmetric conditions show bias roughly equivalent to the bias levels of the uniform distribution conditions. This might indicate that items with a normal/symmetric distribution of responses are more sensitive to number of categories on the outcome (because it showed the most change between the category conditions), or that the skewed distribution takes more information than 5 categories to reduce bias to the levels of the normal or uniform conditions for that effect (as it consistently had larger bias).

Looking at the random effect variance values for the autoregressive/actor and partner effects (Figures 95 and 96 for the 0.10 and 0.05 population values, respectively, the most noticeable trend is that the continuous DSEM models tend to underestimate these effects (unlike the categorical model results in Figures 86 and 87 above, that largely showed an overestimation of effects). Then the population value is larger (0.10, Figure 95), this is particularly true: the partner effect especially (bottom row) hovers around an average value of 0.06 for the 5-category model conditions, and between 0.03 and 0.05 for the 3-category model conditions. Across the 5 parameters displayed in these two figures, the skewed distribution conditions tend to show the largest change across timepoints (have the steepest slopes, see second and fifth column of Figures 95 and 96), and the symmetric/normal distributions tend to have the lowest overall values, particularly for the 3-category models.

Figure 97 shows the random effect variance values for the level-1 covariate effects across all continuous model conditions. Again, the values tend to be underestimated (except when the population value is 0), with the lowest values tending to be in the symmetric/normal distribution categories. There is not really an influence of timepoints or sample size for these values as in the categorical model results (see Figure 88).

Power/Type-I Error Rate for Fixed Effects. The power to detect a significant autoregressive/actor or partner effect across all continuous model conditions is shown in Figure 98, and to detect a significant level-1 or level-2 covariate effect is shown in Figure 99. Overall, Figure 98 again shows that the autoregressive/actor effects have adequate power across all simulation conditions (top two rows), while the partner effect is below the 80% threshold when N = 50 and T = 14 timepoints, and still for the 3-category models when T = 28 timepoints for the normal and skewed outcome distribution conditions (gray line, bottom row of Figure 98). These results follow the categorical model results very closely (see Figure 89 for comparison). The power for covariate effects shown in Figure 99 also match very closely with the results discussed previously in the categorical section (see Figure 90), with all effects having adequate power when N = 200, but either never crossing the 80% threshold at N = 50 (i.e., univariate level-2 covariate, bottom row of Figure 99, first three columns) or only reaching adequate power when T = 56 timepoints.

The Type-I error rates for the actor and partner effects are shown in Figure 100 and for the level-1 and level-2 covariate effects are shown in Figure 101 for all continuous model conditions. When N = 200, the Type-I error rates for the actor and partner effects are higher than the acceptable range across the distribution and category conditions, reaching up to around 35% for the actor effect's skewed distribution and 3-category conditions (top row, fifth column, solid gray line). Generally, the 3-category rates are higher than 5-category rates, and the skewed distribution have higher rates overall for the actor effect, even getting above 10% when N = 50 for the 3-category conditions (top row, second column, solid gray line). In Figure 101, however, the covariate effects are all right within the range of 0.025 to 0.075 for the most part, showing adequate Type-I error rates across all conditions for those parameters.

All Continuous Results Summary. Overall, these collated continuous model results were summarized here, though none of the four research questions directly asked about a comparison between 3- and 5-category outcome conditions across simulation factors for these models; research question 3 (how does the number of categories on the outcome affect estimates?) was intended to compare across categorical DSEM models. However, I think it is important to look at these results together to highlight general trends for how a continuous DSEM performs when the outcome is categorical. The key takeaways from the results for these conditions can be summarized as follows:

- The bias for level-1 fixed effects and random effect variances (autoregressive/actor, partner, covariate) tends to be larger when the outcome has either a symmetric/normal or skewed distribution of responses
- Number of timepoints primarily affected the autoregressive/actor fixed effects and random effect variances, and the partner effect random effect variance, where T = 14 timepoints showed more bias than larger timepoints conditions; the skewed distribution conditions also showed the steepest change in bias across timepoints for these effects
- Sample size had the greatest effect on power and Type-I error rate, with N = 50 sample size conditions showing lower power across all parameters, and N = 200 sample size conditions showing higher Type-I error rates for the actor and partner

effects; Type-I error rates were also higher for skewed distribution conditions for the actor effect

- As in the categorical models, when there were differences between 3- and 5-category models, it was because the 3-category models performed worse (larger bias, lower power, higher Type-I error rates); this was primarily for level-1 fixed effects and random effect variances more than for the level-2 effects

CHAPTER 7

DISCUSSION

Dynamic structural equation models are uniquely suited to model intensive longitudinal data collected from dyads or individuals, for both categorical and continuous outcomes. To this point, however, most assessments of the performance of DSEM models have focused on univariate models with continuous outcomes. For this dissertation project, I examined the behavior and properties of categorical DSEM with both multilevel lag-1 autoregressive and multilevel lag-1 vector autoregressive models (univariate and multivariate models, respectively) across different sample sizes, timepoints, categories on the outcome, and proportion of responses in categories of the outcome. Additionally, I investigated how categorical outcomes perform when they are used in a continuous DSEM, to investigate whether previous work suggesting 5 categories are "enough" to ignore categorical methods can also be argued for in these intensive longitudinal designs. Below, I discuss the results in terms of the four research questions and accompanying hypotheses, then cover limitations and future directions.

Overview and Implications of Results

The first research question I had for this project was: How do number of clusters and number of timepoints affect estimates? I hypothesized that sample size (number of individuals or dyads for univariate and multivariate models, respectively) will have a large impact on the performance of these models, and that number of timepoints will have an impact on the recovery of the actor/autoregressive parameters in particular. Additionally, I hypothesized that the number of clusters and number of timepoints needed to recover unbiased parameters will be larger when the outcome is categorical than the previous studies' assessments of continuous outcomes.

Overall, these hypotheses were supported by my simulation results. When considering power results, both sample size and number of timepoints affected all fixed effect parameters, particularly the partner and covariate effects; the N = 50 sample size and T = 14 and 28 timepoint conditions often failed to reach 80% power for these parameters. The autoregressive/actor parameters overall, however, were impacted by number of timepoints more so than sample size, where T = 14 timepoint conditions showed more bias than T = 28 and T = 56 timepoint conditions across categorical DSEM model results. This was generally true for both the fixed and random effect variances for these parameters and supports Schultzberg & Muthén's (2018) assertions that the autoregressive effects in particular are sensitive to sample size.

Sample size and number of timepoints impacted the bias results for the partner and both the level-1 and level-2 covariate effects (both fixed and random effect variances, when applicable), with the smaller N = 50 sample size and T = 14 having more bias relative to the N = 200 sample size conditions; importantly, however, bias was still high when there were 28 timepoints in the smaller sample size conditions, particularly for the random effect variances.

Overall, this simulation study's results indicate that Gistelinck et al. (2021), Asparouhov et al. (2018), and Asparouhov & Muthén's (2019) claims of how many clusters and how many timepoints are enough to reduce bias / improve the performance of the model may not apply to categorical DSEM in the same way. Specifically, the thresholds for number of timepoints appear to definitely be larger than 10 and 20 as suggested by Gistelinck et al. (2021) and Asparouhov et al. (2018), respectively (even when N = 200), and likely are larger than 30 as well (as suggested by Asparouhov & Muthén, 2019) for random effect variances in particular and when the outcome is binary. However, it is impossible from this simulation study to determine to what extent the increased bias shown in my results above were related to the default M*plus* prior distributions that Asparouhov et al. (2018) cite as being partially responsible for showing bias when timepoints are fewer than 30.

The second research question was: How does the proportion of responses on the outcome affect estimates? I hypothesized that a larger number of clusters and more timepoints were necessary to adequately recover effects from models with skewed and normal/symmetric response distributions. My simulation's results showed more differences amongst the outcome distribution conditions for the partner and covariate effects than the actor/autoregressive effects, and for the random effect variances more than the fixed effects for these parameters. However, within the categorical DSEM results these differences tended to be small (e.g., a slight increase in bias for the skewed and normal conditions when sample size and number of timepoints were small, relative to uniform conditions). I believe that because the probit model specification in M*plus* is able to take these different distributions and keep them outside of the within- and between-level model equations (by setting the intercept to zero and calculating the thresholds as part of the latent decomposition), they do not impact the results as much as they would in a continuous DSEM.

The third research question was: How does the number of categories on the outcome affect estimates? I hypothesized that 5-category model results would be more precise (less bias, higher power) relative to the 2- and 3-category model conditions. This generally held up in the results for the autoregressive/actor fixed effects, as well as for the partner and level-1 covariate random effect variances, particularly when sample size was N = 50 and/or number of timepoints was T = 14, in which the 2-category model conditions had higher bias and lower power relative to the 3- and 5-category models. In many cases, the 3- and 5-category results looked more similar to each other than to the 2-category results, indicating that binary outcomes in particular require higher sample sizes and number of timepoints than outcomes with 3+ categories.

The fourth and final research question was: How does treating the 3- and 5-category outcomes as continuous compare to using a categorical model? I hypothesized that treating the outcome as categorical will be more precise and show less bias relative to treating the outcome as continuous for both the 3- and 5-category conditions, especially for the skewed distributions with large sample sizes. In my results, this was overwhelmingly the case for the non-zero fixed and random effect variance parameters: the continuous DSEM consistently underestimated these effects, even when the categorical model overestimated them.

For the autoregressive/actor fixed effects, the discrepancy between the categorical and continuous bias results were greatest in the normal/symmetric distribution conditions for both the 3- and 5-category outcome models. For the level-1 and level-2 covariate fixed effects, the largest discrepancy in bias results were in the skewed conditions for both the 3- and 5-category models. For the partner fixed effects, however, the largest discrepancy in bias results were in the normal distribution conditions for the 3-category models, but the skewed distribution conditions for the 5-category models. For the random effect variance parameters, estimates in the continuous models were often as low or lower than half of the population value (e.g., lower than 0.05 for effect with a population value of 0.10), and were particularly underestimated in the normal distribution conditions.

In addition to the bias results, the continuous models also showed much higher Type-I error rates for the actor and partner effects. While the categorical models had Type-I error rates that all fell within the range of acceptable values, when the continuous models had a large sample size (N = 200), they consistently fell outside that range, particularly in the 3-category models. The autoregressive effect also showed higher Type-I error rates at the smaller sample size (N = 50) when the outcome had a skewed distribution.

Finally, while not an intended evaluation criterion, creating large datasets with the intent of calculating within- and between-level variances for the continuous and categorical conditions (to later standardize some effects) allowed me to look at trends in explained variance values across different simulation conditions. These are shown in Table 7, and show that in general, the within-level explained variance is around 5-10% lower in the continuous models than in the categorical models. While this difference appears to decrease as number of categories increases, it highlights another way in which researchers might make different judgements about their model if they ignore the categorical nature of the outcome variable. Especially considering that the primary focus of intensive longitudinal models (and actor-partner interdependence models in general) are the within-level effects, these differences seemed appreciable enough to include here.

Limitations and Future Directions

As with all simulation studies, I sought to select conditions that could provide information about a wide variety of categorical DSEM models, this dissertation project has limitations to its design. Within the simulation conditions themselves, efficacy was the least useful evaluation criteria. It is dependent on number of replications, and seemed to also be influenced by the M*plus* default prior distributions and, for the random effect variance parameters, by the bias in the fixed effect parameters. These factors made it difficult to determine when efficacy results indicated a truly problematic overestimation of variability, and would require a more careful consideration of simulation details to be useful in the future (e.g., more replications, maybe using only for fixed effects or parameters whose population values meet a certain threshold of magnitude).

Another limitation within the simulation conditions for this project is that looking at the values/bias for the small random effect variances can be difficult to judge in terms of practical effects. For example, when the partner random effect variance with a population value of 0.05 is estimated to be 0.075 (a 50% relative bias!) – is this enough to make a practical difference for in a study's results? The fixed effect for that partner effect (assuming no bias in the fixed effect) is 0.15, implying that 95% of individuals or dyads have a value for that effect between [-0.39, 0.69] when the variance is overestimated as 0.075, and between [-0.29, 0.59] when the variance is accurately estimated. Are these intervals different enough to affect the conclusions or recommendations from a study? In some situations that might be true, but in other cases it could be a small enough absolute difference that, even though the relative difference is large, it does not substantively impact results.

Looking outside of the simulation factors examined here, there are several factors that could have been manipulated that were not, which would provide useful results in a future simulation. One category of factors that should be examined more closely in the future are the different M*plus* Bayesian estimation features, such the prior distributions. For this simulation, my aim was to show how these models perform using as many default features as possible, as this is presumably how many users would conduct a DSEM on their own categorical data. However, I believe that the results of this simulation were consistently influenced by the default prior distributions, in particular the random effect variance parameters. McNeish (2019) discusses how, particularly with small sample sizes, the default $\Gamma^{-1}(-1, 0)$ prior distribution that is meant to be diffuse can be unintentionally informative. To remedy this, the prior can be replaced with one that is *intentionally* informative and provides a mode for the prior that falls within an acceptable range of possible values the parameter could be (as opposed to being "infinity" in the default prior distribution). Other Bayesian estimation features that could be changed would be the number of chains and minimum number of iterations to run before the model could converge; I used the default number of chains (2) and specified the minimum number of iterations to be 2,000 as M*plus* does not have a default minimum for iterations. Increasing these two optimization criteria would increase the number of iterations used to construct the posterior distributions, making them more robust and able to provide a more accurate assessment of model performance (such as via efficacy).

Another factor to examine in the future is to vary the population values themselves. In this study, the population values remained the same across all conditions, but I consistently found in the results that the magnitude of the values likely impacted the performance of the model. This was apparent in both the bias results (e.g., smaller random effect variances showing *less* underestimation of effects in the continuous model, because values were bounded at 0), and in the power results (where power was generally lower for the population values of lower magnitude – partner effects relative to actor effects, for example – when sample sizes were small). Altering the population values would allow for the investigation of how different patterns of effects influence results – for example, if the partner effects were equal in magnitude to the actor effects, or if the actor effects were large and small for the two outcomes instead of having one zero and one non-zero effect. There are a multitude of ways to expand a future simulation study in this way to develop a clearer picture of how these models perform both within the categorical probit models and in the comparison between categorical and continuous DSEM.

The final main category of limitations to this study that could be improved in future simulations is related to the outcome variables themselves. Namely, the number of categories on the outcome and the form of the underlying distribution that manifests as the categorical observed variables. With respect to the number of categories on the outcome, this study had a maximum of 5 categories, which is at the lower end of the range that has been argued to perform adequately as a continuous model (e.g., Rhemtulla et al., 2012). These results suggest that 5 categories are likely still too small to be considered continuous for DSEM models; the logical next step, then, is to determine at what point a categorical outcome shows similar performance between the categorical and continuous DSEM. Additionally, all the observed responses were generated based on a continuous underlying normal distribution. Flora and Curran (2004) looked at the difference between categorical and continuous model performance as a function of various levels of non-normality in the underlying continuous distribution. They found that categorical models were robust to *moderate* levels of underlying non-normality, but Rhemtulla et al. (2012) found that more extreme violations of normality impacted categorical methods more than continuous ones. Both of these studies evaluated CFA models, so similar specifications can be used to see to what extent non-normality impacts categorical and continuous DSEM models when the outcomes are ordinal.

Conclusions

This dissertation project showed how the performance of categorical DSEM models vary across number of clusters, timepoints, and categories on the outcome, and across distribution of responses on the outcome. Taken altogether, the results indicated that:

- Categorical DSEM models require a larger number of clusters and timepoints to produce unbiased results relative to a continuous DSEM with a continuous outcome variable
- The distribution of responses on the outcome did not have as much of an impact on the categorical model results as hypothesized (but did matter in the continuous DSEM models)
- The 5-category outcomes showed the least bias and higher power on average in the category models, and none of the parameters had issues in Type-I error rates.
- Treating the outcome as continuous, DSEM consistently underestimated effects across parameters and conditions, and showed large Type-I error rates for the actor and partner effects when sample size was large

Overall, this project was motivated by a real daily diary study with Likert-scale items. I was interested in the extent to which I could trust the estimates produced in the motivating data example, which had a sample size of 54 dyads assessed across 21 days, to which I fit a categorical DSEM for a 5-category outcome variable and a continuous covariate in an L-APIM analysis. Results from the simulation study indicated that these estimates might have a slight positive bias for the autoregressive/actor effects, and be underpowered to detect partner effects and both level-1 and level-2 covariate effects. It does appear, however, that ignoring the categorical nature of the outcome would have resulted in larger bias across more parameters, along with higher Type-I error rates for the actor and partner effects. Future studies should assess the extent to which these results (and the results of the simulation study's conditions in general) persist when number of categories on the outcome is larger, the default prior distributions are altered, or population values are altered, along with considering other categorical outcome situations not covered in this dissertation project.

REFERENCES

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

Allison, P. D. (2012, April). Handling missing data by maximum likelihood. Paper presented at the SAS Global Forum, Orlando, FL. Retrieved from http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling*, *25*(3), 359-388.

Asparouhov, T. & Muthén, B.O. (2015). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. In J.R. Harring, L.M. Stapleton & S.N. Beretvas (Eds.), *Advances in Multilevel Modeling for Educational Research* (pp. 163-192). Information Age Publishing.

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

Asparouhov, T. & Muthén, B. (2020). Comparison of models for the analysis of intensive longitudinal data. *Structural Equation Modeling*, *27*(2), 275-297.

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

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*, 135-167.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. *Psychological Methods*, *11*, 142-163.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*(1), 597-616.

Bolger, N. & Laurenceau, J.-P. (2013). *Intensive Longitudinal Methods.* Guilford Press.

Bolger, N. & Shrout, P. (2007). Accounting for statistical dependencies in longitudinal data on dyads. In T. Little, J. Bovaird, & N. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 149-182). Lawrence Erlbaum Associates.

Bringmann, L. F., Ferrer, E., Hamaker, E. L., Borsboom, D., & Tuerlinckx, F. (2018). Modeling nonstationary emotion dynamics in dyads using a time-varying vector-autoregressive model. *Multivariate Behavioral Research*, *53*, 293-314.

Chatfield, C. (2003). *The analysis of time series: An introduction*. CRC Press.

Chen, M., Chow, S. M., Hammal, Z., Messinger, D. S., & Cohn, J. F. (2020). A person- and time-varying vector autoregressive model to capture interactive infant-mother head movement dynamics. *Multivariate Behavioral Research*, advance online publication.

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

Cook, K. F., Cella, D., Boespflug, E. L., & Amtmann, D. (2010). Is less more? A preliminary investigation of the number of response categories in self-reported pain. *Patient Related Outcome Measures, 1*, 9-18.

Cook, W. L., & Kenny, D. A. (2005). The actor–partner interdependence model: A model of bidirectional effects in developmental studies. *International Journal of Behavioral Development*, *29,* 101-109.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum.

Curran, P. J. & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology, 62,* 583-619.

Curran, P. J., Howard, A. L., Bainter, S. A., Lane, S. T., & McGinley, J. S. (2014). The separation of between-person and within-person components of individual change over time: a latent curve model with structured residuals. *Journal of Consulting and Clinical Psychology*, *82,* 879-894.

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

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford University Press.

Durbin, J. & Koopman, S. J. (2012). *Time series analysis by state space methods* (2nd ed.). Oxford University Press.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

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

Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods*, *23,* 3105-3119.

Gistelinck, F., & Loeys, T. (2019). The actor–partner interdependence model for longitudinal dyadic data: An implementation in the SEM framework. *Structural Equation Modeling*, *26*, 329-347.

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

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development*, *58*, 80–92.

Gonzalez, R. & Griffin, D. (2012). Dyadic data analysis. In H. Cooper (Ed.), *APA handbook of research methods in psychology: Vol. 3. Data analysis and research publication* (pp. 439-450). American Psychological Association.

Gunthert, K. C., & Wenze, S. J. (2014). Daily diary methods. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of Research Methods for Studying Daily Life* (pp. 144-159). Guilford Press.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. *Multivariate Behavioral Research*, *53*, 820-841.

Hamaker, E. L., Asparouhov, T., & Muthén, B. (2022). Dynamic structural equation modeling as a combination of time series modeling, multilevel modeling, and structural equation modeling. To be published as Chapter 31 in R. H. Hoyle (Ed.), *The Handbook of Structural Equation Modeling (2nd Edition)*. Guilford Press.

Hamaker, E. L., & Grasman, R. P. P. P. (2012). Regime switching state-space model applied to psychological processes: Handling missing data and making inferences. *Psychometrika*, *77*, 400–422.

Hamaker, E. L., & Grasman, R. P. P. P. (2015). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology, 5,* 1492.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, *64*, 627-634.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine, 31*, 3328-3336.

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

Iida, M., Shrout, P. E., Laurenceau, J.-P., & Bolger, N. (2012). Using diary methods in psychological research. In H. Cooper (Ed.), *APA handbook of research methods in psychology: Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 277-305). American Psychological Association.

Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in Psychology*, *6*, 727.

Jongerling, J., Laurenceau, J. P., & Hamaker, E. L. (2015). A multilevel AR (1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research*, *50*, 334-349.

Kashy, D. A. & Kenny, D. A. (2000). The analysis of data from dyads and groups. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 451-477). Cambridge University Press.

Kashy, D. A., & Snyder, D. K. (1995). Measurement and data analytic issues in couples research. *Psychological Assessment*, *7*, 338-348.

Kenny, D. A. (1996). Models of non-independence in dyadic research. *Journal of Social and Personal Relationships*, *13*, 279-294.

Kenny, D. A., & Cook, W. (1999). Partner effects in relationship research: Conceptual issues, analytic difficulties, and illustrations. *Personal Relationships*, *6*, 433-448.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*(3), 422–431.

Kenny, D. A. & Kashy, D. A. (2011). Dyadic data analysis using multilevel modeling. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 335-370). Taylor & Francis.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic Data Analysis*. Guilford Press.

Kim, C.J., & Nelson, C.R. (1999). *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*. MIT Press.

Kuiper, R. M., & Ryan, O. (2018). Drawing conclusions from cross-lagged relationships: Re-Considering the role of the time-interval. *Structural Equation Modeling*, *25*, 809-823.

Kop, W. J., Verdino, R. J., Gottdiener, J. S., O'Leary, S. T., Bairey Merz, C. N., & Krantz, D. S. (2001). Changes in heart rate and heart rate variability before ambulatory ischemic events. *Journal of the American College of Cardiology, 38*(3), 742-749.

Laurenceau, J.-P. & Bolger, N. (2005). Using diary methods to study marital and family processes. *Journal of Family Psychology, 19*(1), 86-97.

Laurenceau, J.-P., & Bolger, N. (2012). Analyzing diary and intensive longitudinal data from dyads. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 407-422). Guilford Press.

Ledermann, T., & Kenny, D. A. (2017). Analyzing dyadic data with multilevel modeling versus structural equation modeling: A tale of two methods. *Journal of Family Psychology, 31*, 442-452.

Ledermann, T., & Kenny, D. A. (2015). A toolbox with programs to restructure and describe dyadic data. *Journal of Social and Personal Relationships, 32*, 997-1011.

Ledermann, T., & Macho, S. (2014). Analyzing change at the dyadic level: The common fate growth model. *Journal of Family Psychology*, *28*, 204-2013.

Long, S. J. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Sage Publications.

Lüdtke, O., Marsh, H. W., Robitsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*, 203-229.

McNeish, D. (2019). Two-level dynamic structural equation models with small samples. *Structural Equation Modeling, 26*(6), 948-966.

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods,* 25, 610-635.

McNeish, D. & Matta, T.H. (2020). Flexible treatment of time-varying covariates with time unstructured data. *Structural Equation Modeling, 27*, 298-317.

McNeish, D., Mackinnon, D.P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data. *Structural Equation Modeling, 28*(5), 807-822.

Muthén, B. (1997). 10. Latent variable modeling of longitudinal and multilevel data. *Sociological Methodology*, *27*, 453-480.

Muthén, L. K. & Muthén, B. O. (1998-2017). *Mplus User's Guide*. Eighth Edition. Muthén & Muthén.

Nezlek, J. B. (2014). Multilevel modeling analyses of diary-style data. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of Research Methods for Studying Daily Life* (pp. 357-383). The Guilford Press.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica, 49,* 1417-1426.

Planalp, E. M., Du, H., Braungart-Rieker, J. M., & Wang, L. (2017). Growth curve modeling to studying change: A comparison of approaches using longitudinal dyadic data with distinguishable dyads. *Structural Equation Modeling*, *24*, 129-147.

Rast, P., Hofer, S. M., & Sparks, C. (2012). Modeling individual differences in within-person variation of negative and positive affect in a mixed effects location scale model using BUGS/JAGS. *Multivariate Behavioral Research, 47*(2), 177-200.

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

Reis, T., Collins, W. A., & Berscheid, E. (2000). The relationship context of human behavior and development. *Psychological Bulletin, 126*(6), 844-872.

Rhemtulla, M., Brousseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354-373.

Rindskopf, D. (1984). Using phantom and imaginary latent variables to parameterize constraints in linear structural models. *Psychometrika, 49,* 37-47.

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education, 5*, 1-7.

Ryan, O., Kuiper, R.M., & Hamaker, E.L. (2018). A continuous time approach to intensive longitudinal data: What, Why and How? In K.V. Montfort, J.H.L. Oud, & M.C. Voelkle (Eds.), *Continuous time modeling in the behavioral and related sciences* (pp. 27-54). New York: Springer.

SAS Institute Inc. (2018). *SAS/STAT® 15.1 User's Guide*. Chapter 81: The Mixed Procedure (pp. 6536-6729). Cary, NC: SAS Institute Inc.

Savord, A., McNeish, D., Iida, M., Quiroz, S., & Ha, T. (2022). Fitting the actor-partner interdependence model as a dynamic structural equation model in M*plus*. *Structural Equation Modeling,* 1-19.

Schiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4,* 1-32.

Schoemann, A. M. & Sakaluk, J. K. (2016, May). *Missing data in dyadic modeling: Issues and opportunities*. Paper presented at the 6th Modern Modeling Method Conference, University of Connecticut, Storrs, CT.

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

Schultzberg, M., & Muthén, B. (2018). Number of subjects and time points needed for multilevel time-series analysis: A simulation study of dynamic structural equation modeling. *Structural Equation Modeling*, *25*(4), 495-515.

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

Voelkle, M. C. & Oud, J. H. (2015). Relating latent change score and continuous time models. *Structural Equation Modeling, 22*(3), 366-381.

Voelkle, M. C., Oud, J. H., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: relating authoritarianism and anomia. *Psychological Methods*, *17*, 176-192.

Wang, L.P., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Multivariate Behavioral Research*, *20*, 63–83.

Wendorf, C. A. (2002). Comparisons of structural equation modeling and hierarchical linear modeling approaches to couples' data. *Structural Equation Modeling*, *9*, 126-140.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics-Simulation and Computation*, *22*(4), 1079-1106.

Zhang, Z., Hamaker, E. L., & Nesselroade, J. R. (2008). Comparisons of four methods for estimating a dynamic factor model. *Structural Equation Modeling*, *15*(3), 377-402.

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

APPENDIX A

TABLES

Table 1. Results from Motivating Data Example Fitting an L-APIM With Categorical Outcomes and Continuous Covariates as a Categorical DSEM

	Male Outcome SupportM			*Female Outcome SupportF*		
Effect	Notation	Posterior Median	95% Cred. Interval	Notation	Posterior Median	95% Cred. Interval
Fixed Effect Parameters						
Threshold 1	τM_1	-3.471	[-4.065, -2.869]	τF_1	-3.483	[-3.988, -3.047]
Threshold 2	τM_2	-2.500	[-2.956, -1.927]	τF_2	-2.533	[-2.944, -2.153]
Threshold 3	τM_3	-1.175	[-1.56, -0.673]	τF_3	-1.264	[-1.602, -0.928]
Threshold 4	τM_4	0.243	[-0.117, 0.682]	τF_4	0.052	[-0.267, 0.389]
Actor	γ_{30}	**0.245**	**[0.107, 0.403]**	γ_{40}	0.107	[-0.028, 0.249]
Partner	γ_{50}	0.018	[-0.134, 0.151]	γ_{60}	**0.141**	**[0.026, 0.290]**
Connect$^{(w)}$ Effect	γ_{70}	**0.138**	**[0.087, 0.198]**	γ_{80}	**0.150**	**[0.097, 0.207]**
Connect$^{(b)}$ Effect	γ_{11}	**0.304**	**[0.190, 0.425]**	γ_{21}	**0.354**	**[0.265, 0.450]**
Connect Latent Mean	γ_{90}	-0.212	[-1.260, 0.828]	$\gamma_{10,10}$	0.246	[-0.824, 1.307]
Variance Parameters						
Threshold/Intercept Variance	σ_{11}	**1.912**	**[0.914, 3.544]**	σ_{22}	**1.203**	**[0.682, 2.115]**
Autocorrelation Actor Variance	σ_{33}	**0.087**	**[0.020, 0.212]**	σ_{44}	**0.103**	**[0.033, 0.218]**
Partner Variance	σ_{55}	**0.097**	**[0.009, 0.304]**	σ_{66}	**0.043**	**[0.002, 0.199]**
Connect$^{(w)}$ Effect Variance	σ_{77}	**0.014**	**[0.003, 0.041]**	σ_{88}	**0.020**	**[0.008, 0.046]**
Connect$^{(b)}$ Variance	σ_{99}	**13.526**	**[9.065, 21.049]**	$\sigma_{10,10}$	**13.532**	**[8.965, 21.285]**
Connect$^{(w)}$ Variance	ω_1	**12.786**	**[11.518, 14.254]**	ω_2	**17.929**	**[16.295, 19.836]**
Covariance Parameters						
Within-level residual covariance	ρ_{21}	**0.169**	**[0.034, 0.300]**			
Between-level intercept covariance	σ_{21}	**1.050**	**[0.461, 1.977]**			

Notes. 95% Cred. Interval = 95% credible interval. Bolded entries indicate the 95% credible interval does not contain 0.

Table 2. Predicted Probability of Responding in Each Category of Support Outcome as a Function of Latent Person-Mean for Connect for Male and Female Partners

Latent Person Mean $Connect_d^b$	Predicted Probability of Responding in Category of Support				
	Not at all	A little	Moderately	Quite a bit	Extremely
Male Partners					
1	0.57	0.17	0.16	0.07	0.02
2	0.51	0.18	0.18	0.09	0.03
3	0.45	0.19	0.20	0.11	0.04
4	0.40	0.19	0.22	0.13	0.06
5	0.34	0.18	0.24	0.16	0.08
6	0.29	0.18	0.25	0.18	0.11
7	0.24	0.17	0.25	0.21	0.13
8	0.20	0.16	0.25	0.23	0.17
9	0.16	0.14	0.25	0.24	0.21
10	0.12	0.12	0.24	0.26	0.25
11	0.10	0.11	0.22	0.27	0.30
12	0.07	0.09	0.21	0.27	0.36
13	0.06	0.08	0.19	0.27	0.42
14	0.04	0.06	0.16	0.26	0.47
15	0.03	0.05	0.14	0.25	0.53
16	0.02	0.04	0.12	0.23	0.59
17	0.01	0.03	0.10	0.21	0.65
18	0.01	0.02	0.08	0.19	0.70
19	0.01	0.02	0.06	0.16	0.75
20	0.00	0.01	0.05	0.14	0.80
Female Partners					
1	0.64	0.22	0.12	0.02	0.00
2	0.54	0.25	0.17	0.04	0.00
3	0.43	0.27	0.23	0.06	0.01
4	0.33	0.28	0.28	0.10	0.01
5	0.24	0.26	0.33	0.14	0.03
6	0.17	0.23	0.36	0.20	0.05
7	0.11	0.19	0.36	0.25	0.08
8	0.07	0.15	0.35	0.31	0.12
9	0.04	0.11	0.31	0.35	0.19
10	0.02	0.07	0.26	0.37	0.27
11	0.01	0.05	0.21	0.37	0.36
12	0.01	0.03	0.15	0.35	0.46
13	0.00	0.02	0.11	0.31	0.57
14	0.00	0.01	0.07	0.25	0.67
15	0.00	0.00	0.04	0.20	0.76
16	0.00	0.00	0.02	0.14	0.83
17	0.00	0.00	0.01	0.10	0.89
18	0.00	0.00	0.01	0.06	0.93
19	0.00	0.00	0.00	0.04	0.96
20	0.00	0.00	0.00	0.02	0.98

Table 3. Approximate Percentage of Responses in Each Category of the Outcome For 2-, 3-, and 5-Category Models with Symmetric/Normal, Skewed, and Uniform Response Distributions

		Distribution		
	Response Category	NORMAL	SKEWED	UNIFORM
5-CAT	0	10	50	20
	1	20	20	20
	2	40	10	20
	3	20	10	20
	4	10	10	20
3-CAT	0	16	65	33
	1	68	25	34
	2	16	10	33
2-CAT	0	N/A	25	50
	1	N/A	75	50

Notes. CAT = categories. Values represent approximate percentage of total responses in each response category of the outcome.

Table 4. Population Values for All Non-Threshold Parameters for the Univariate Model Simulation Conditions

Effect	Notation	Value
Fixed Effect Parameters		
Autocorrelation	γ_{10}	0.25
Covariate1$^{(w)}$ Effect	γ_{20}	0.15
Covariate2$^{(w)}$ Effect	γ_{30}	0.00
Covariate1$^{(b)}$ Effect	γ_{01}	0.35
Covariate2$^{(b)}$ Effect	γ_{02}	0.00
Covariate1 Latent Mean	γ_{40}	0.00
Covariate2 Latent Mean	γ_{50}	0.00
Variance Parameters		
Threshold/Intercept Variance	σ_{00}	1.50
Autocorrelation Variance	σ_{11}	0.10
Covariate1$^{(w)}$ Effect Variance	σ_{22}	0.05
Covariate2$^{(w)}$ Effect Variance	σ_{33}	0.00
Covariate1$^{(b)}$ Variance	σ_{44}	1.00
Covariate2$^{(b)}$ Variance	σ_{55}	1.00
Covariate1$^{(w)}$ Variance	ω_1	1.00
Covariate2$^{(w)}$ Variance	ω_2	1.00

Table 5. Population Values for All Non-Threshold Parameters for the Multivariate Model Simulation Conditions

	Y1		*Y2*	
Effect	Notation	Value	Notation	Value
Fixed Effect Parameters				
Autocorrelation/Actor	γ_{30}	0.25	γ_{40}	0.00
Partner	γ_{50}	0.00	γ_{60}	0.15
Covariate$^{(w)}$ Effect	γ_{70}	0.15	γ_{80}	0.00
Covariate$^{(b)}$ Effect	γ_{11}	0.00	γ_{21}	0.35
Covariate Latent Mean	γ_{90}	0.00	$\gamma_{10,10}$	0.00
Variance Parameters				
Threshold/Intercept Variance	σ_{11}	2.00	σ_{22}	1.00
Autocorrelation/Actor Variance	σ_{33}	0.10	σ_{44}	0.05
Partner Variance	σ_{55}	0.10	σ_{66}	0.05
Covariate$^{(w)}$ Effect Variance	σ_{77}	0.05	σ_{88}	0.10
Covariate$^{(b)}$ Variance	σ_{99}	1.00	$\sigma_{10,10}$	1.00
Covariate$^{(w)}$ Variance	ω_1	1.00	ω_2	1.00
Covariance Parameters				
Within-level residual covariance	ρ_{21}	0.20		
Between-level intercept covariance	σ_{21}	1.00		

Table 6. Population Values for Threshold Parameters Across All Simulation Conditions for Multivariate and Univariate Models

		Multivariate						**Univariate**		
		Y1			*Y2*			*Y*		
	Threshold	Normal	Skewed	Uniform	Normal	Skewed	Uniform	Normal	Skewed	Uniform
5-Category	1	-2.417	0.000	-1.587	-2.017	0.000	-1.329	-2.222	0.000	-1.462
	2	-0.988	0.980	-0.477	-0.824	0.851	-0.399	-0.908	0.927	-0.439
	3	0.988	1.575	0.477	0.824	1.367	0.399	0.908	1.489	0.439
	4	2.417	2.398	1.587	2.017	2.081	1.329	2.222	2.267	1.462
	Intercepts	*2.00*	*1.10*	*2.00*	*2.00*	*1.10*	*2.00*	*2.00*	*1.10*	*2.00*
3-Category	1	-1.918	0.751	-0.826	-1.526	0.615	-0.661	-1.730	0.685	-0.748
	2	1.918	2.502	0.826	1.526	2.046	0.661	1.730	2.280	0.748
	Intercepts	*1.00*	*0.45*	*1.00*	*1.00*	*0.45*	*1.00*	*1.00*	*0.45*	*1.00*
2-Category	1	---	-1.272	0.000	---	-0.995	0.000	---	-1.137	0.000
	Intercepts	---	*0.75*	*0.50*	---	*0.75*	*0.50*	---	*0.75*	*0.50*

Table 7. Within- and Between-Level R-Squared Values across Different Distributions, Models, and Number of Categories for Univariate and Multivariate Simulation Conditions

		Multivariate								**Univariate**			
		Y1				**Y2**				**Y**			
		Within		Between		Within		Between		Within		Between	
Distribution	Cats	CAT	CON	CAT	CON	CAT	CON	CAT	CON	CAT	CON	CAT	CON
Normal	2	---	---	---	---	---	---	---	---	---	---	---	---
	3	0.277	0.130	0.056	0.052	0.206	0.100	0.000	0.000	0.207	0.109	0.077	0.073
	5	0.279	0.211	0.122	0.107	0.206	0.160	0.000	0.000	0.209	0.164	0.077	0.075
Skewed	2	0.275	---	0.053	---	0.211	---	0.000	---	0.203	---	0.078	---
	3	0.276	0.168	0.054	0.044	0.207	0.120	0.000	0.000	0.211	0.142	0.076	0.064
	5	0.278	0.208	0.111	0.099	0.206	0.148	0.000	0.000	0.210	0.167	0.076	0.067
Uniform	2	0.272	---	0.053	---	0.204	---	0.000	---	0.206	---	0.075	---
	3	0.276	0.173	0.052	0.050	0.206	0.125	0.000	0.000	0.207	0.139	0.075	0.072
	5	0.277	0.222	0.110	0.108	0.205	0.155	0.000	0.000	0.207	0.169	0.076	0.074

Notes. CAT = categorical DSEM. CON = continuous DSEM. Cats = Categories on the outcome(s).

APPENDIX B

FIGURES

Figure 1. Path Diagram for a Multilevel Lag-1 Autoregressive DSEM with Continuous Outcome and Continuous Covariate

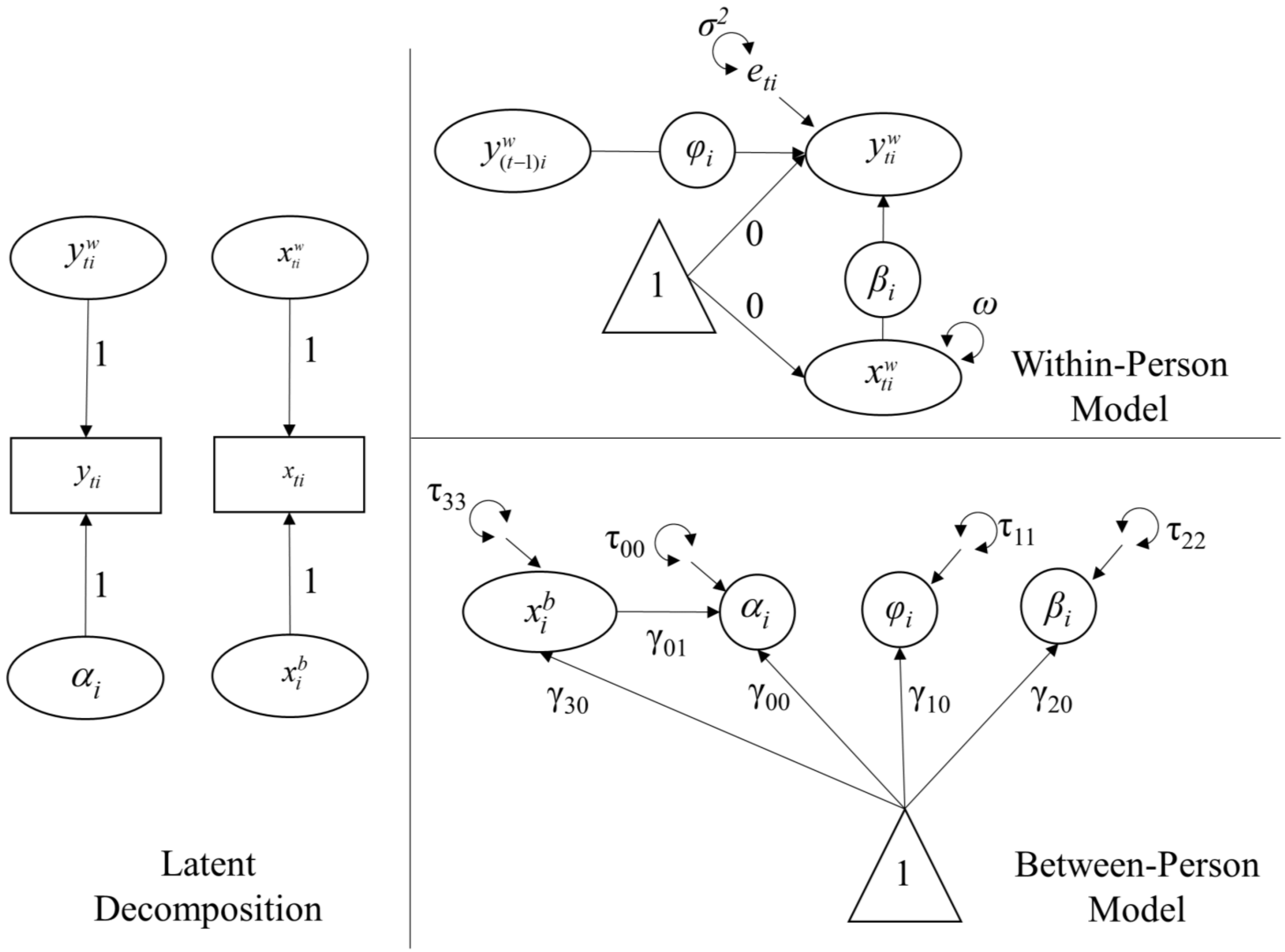

Figure 2. Conceptualization of Longitudinal Dyad Data as (a) a 3-Level Model, and (b) a 2-Level Model

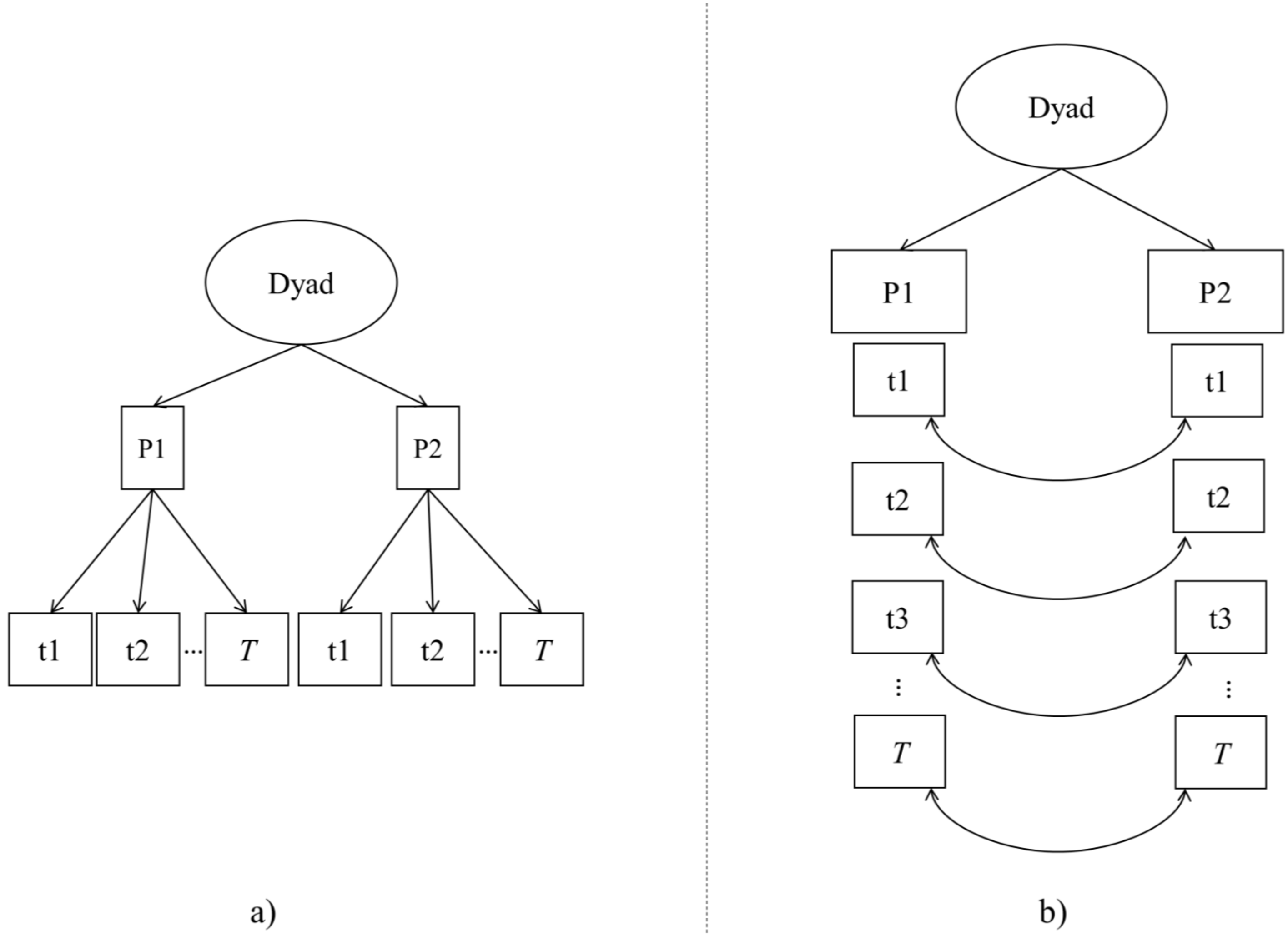

Figure 3. Conceptual Path Diagram for the APIM

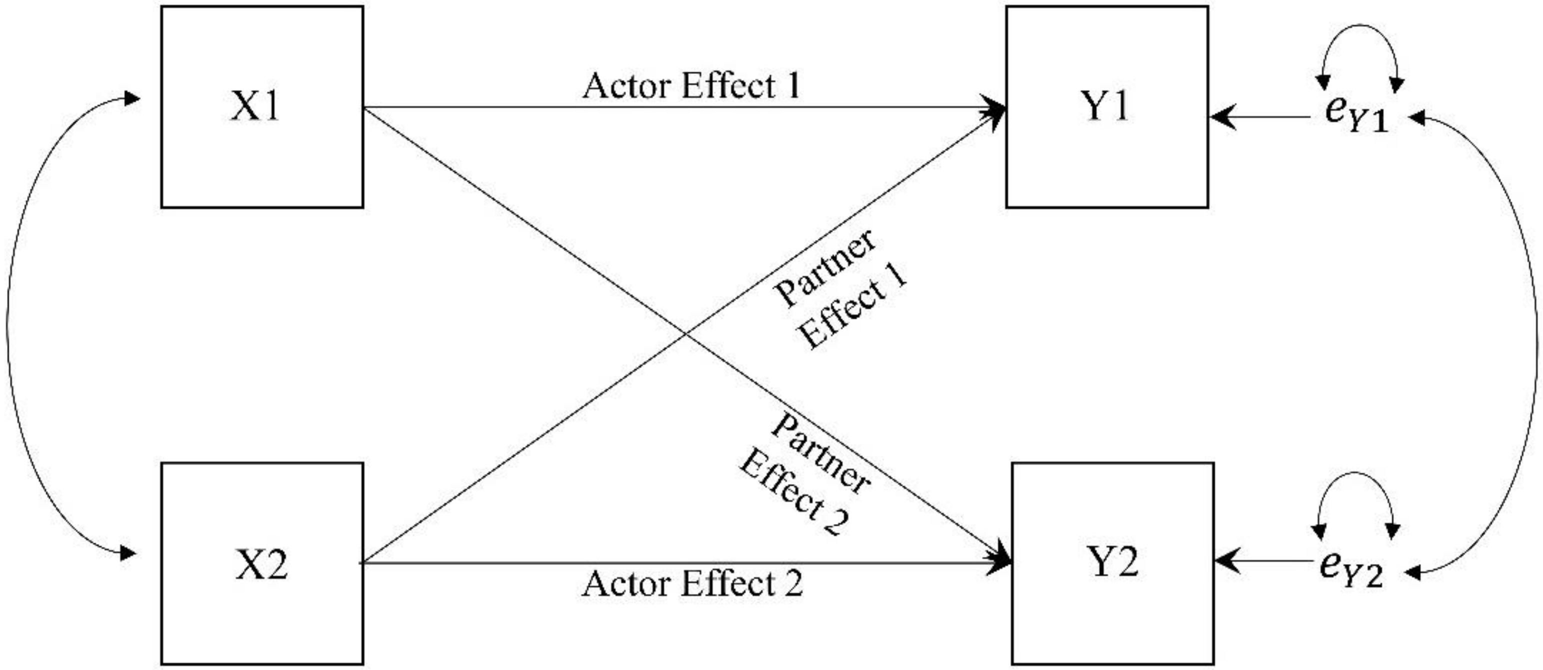

Figure 4. L-APIM as a Structural Equation Model with 3 Repeated Measures

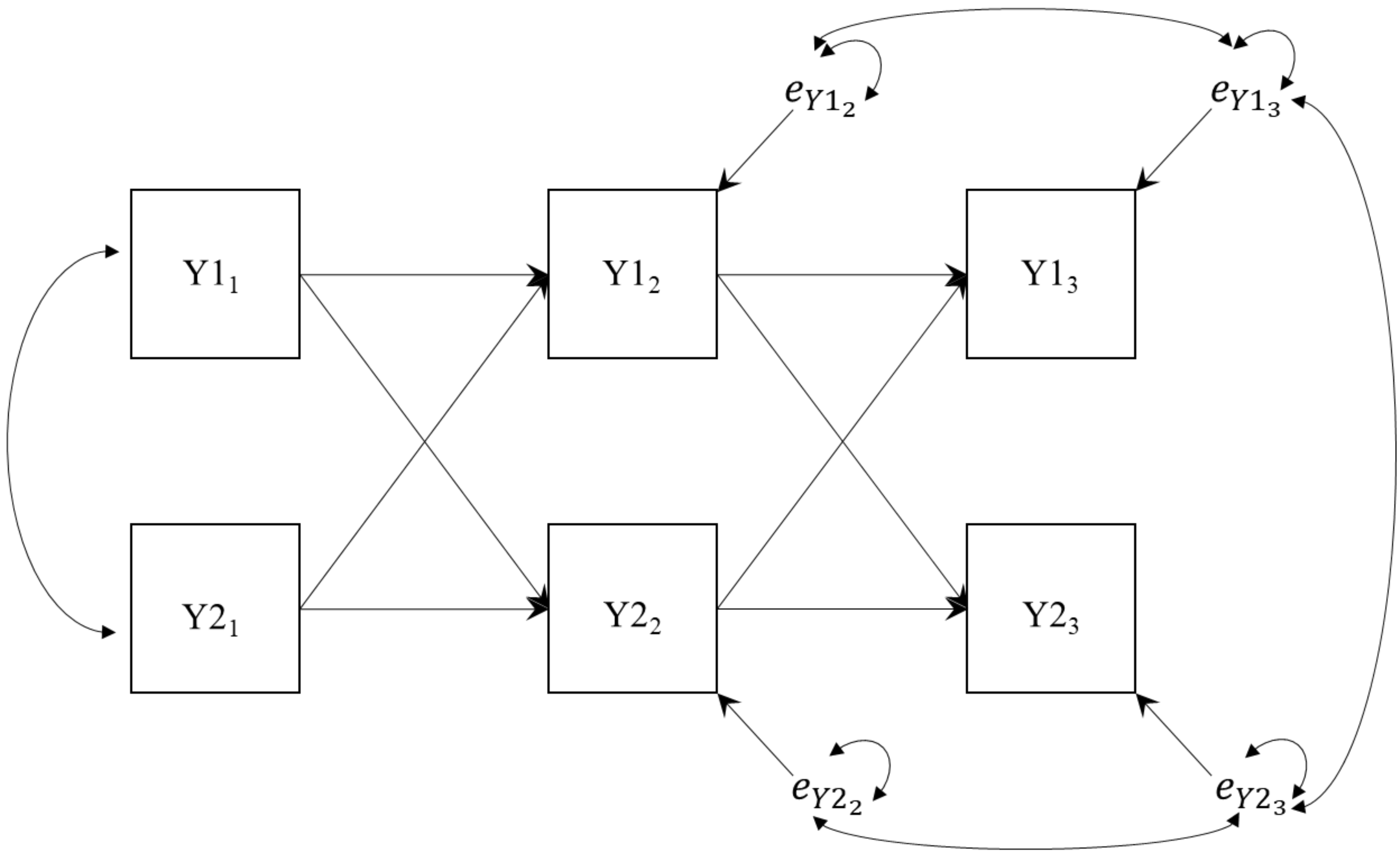

Figure 5. L-APIM Latent Decomposition, Within-Level, and Between-Level Specification with Random Intercepts and Slopes in the DSEM Framework

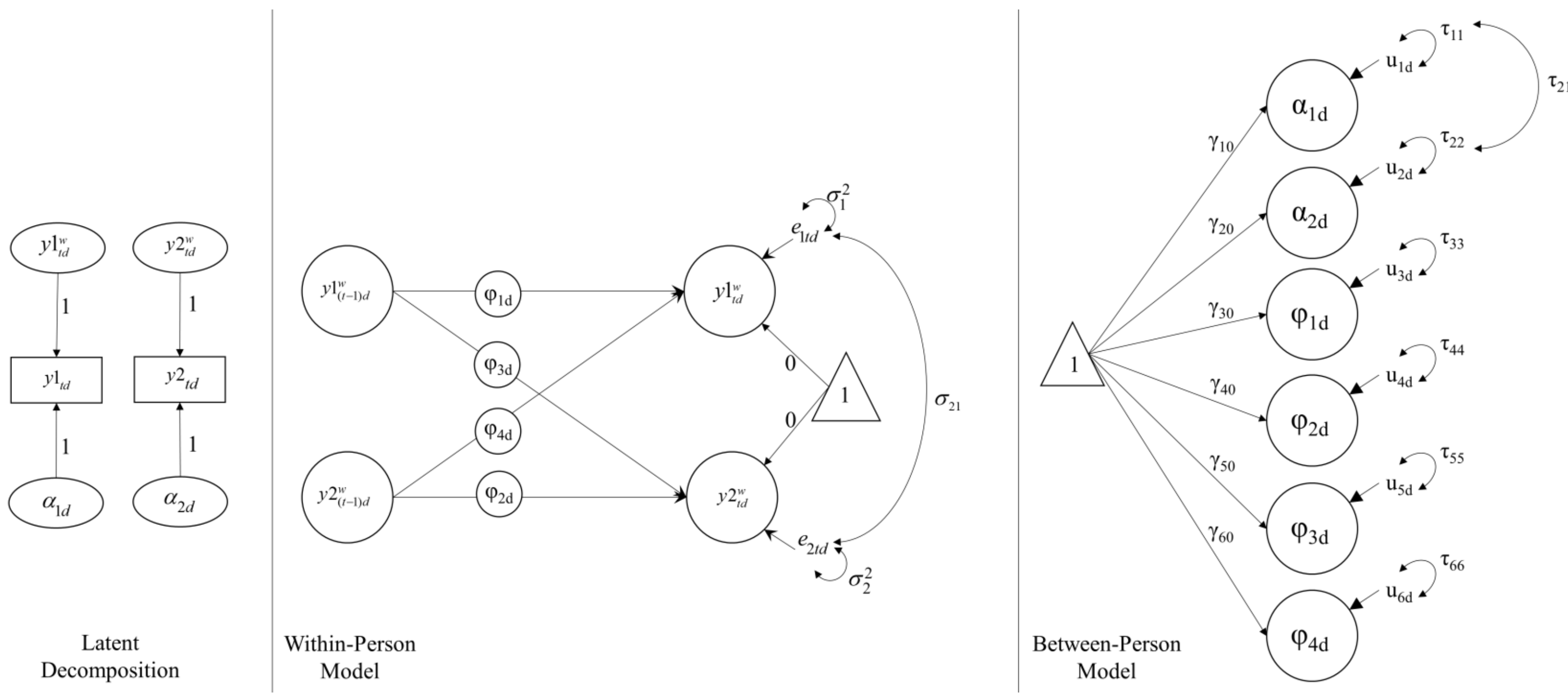

Figure 6. Categorical DSEM for a Binary Outcome and Continuous Covariate

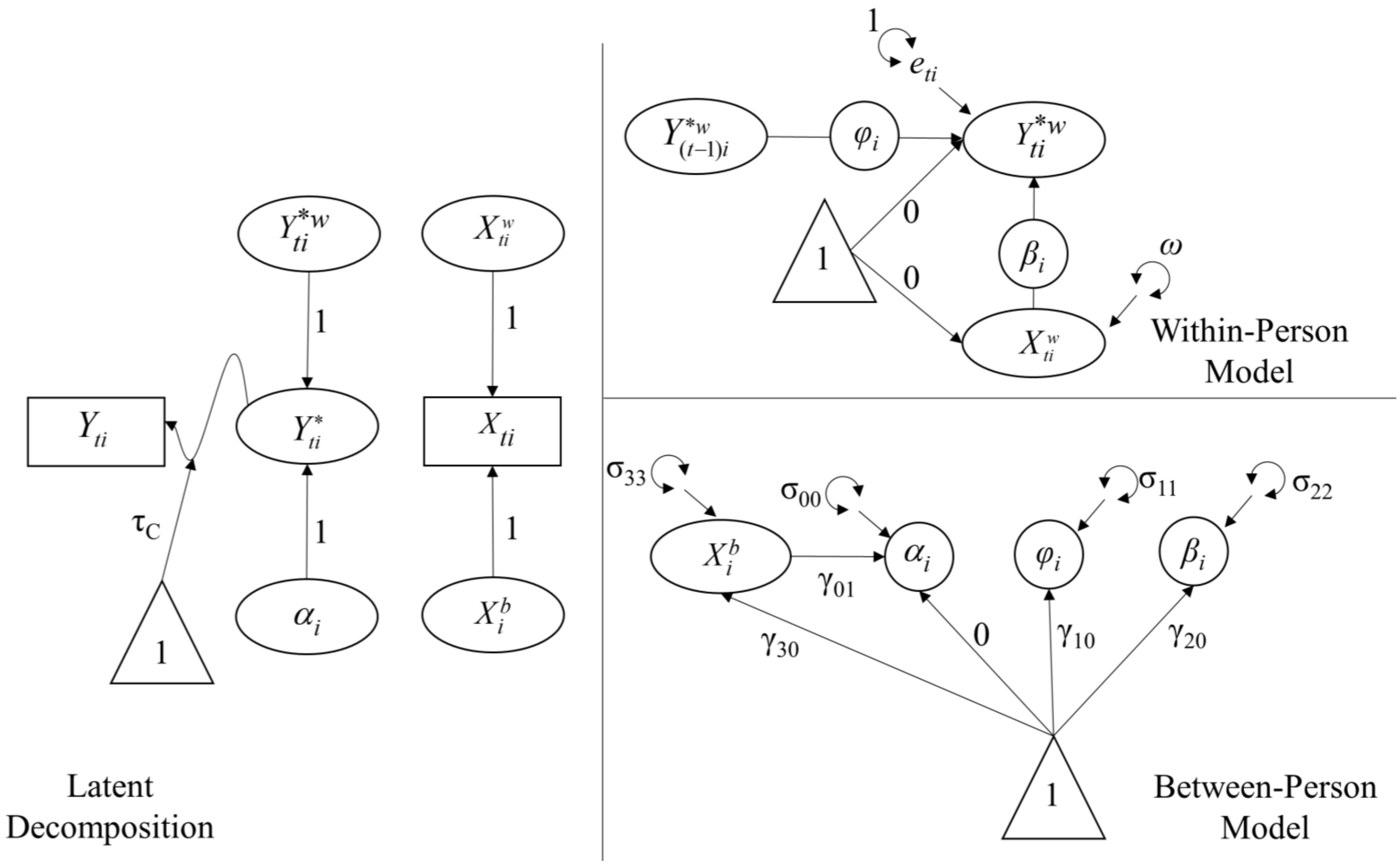

Figure 7. Hypothetical Underlying Normal Distribution of Stress Scores with Thresholds Denoting Boundaries Between Manifest Categories on the Observed Outcome Variable

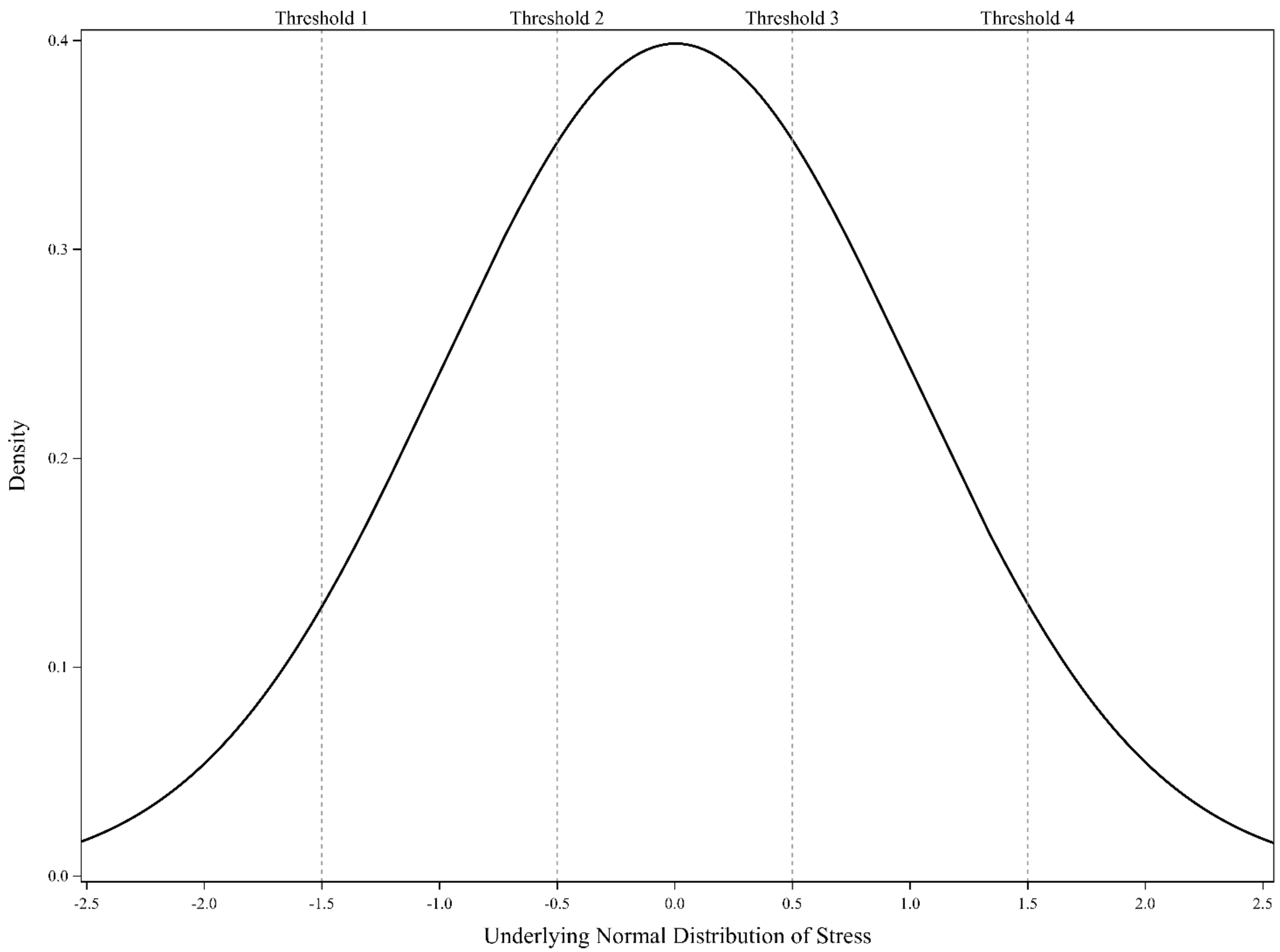

Figure 8. L-APIM Latent Decomposition, Within-Level, and Between-Level Specifications for Motivating Data Example with 5-Category Outcomes and Continuous Covariates

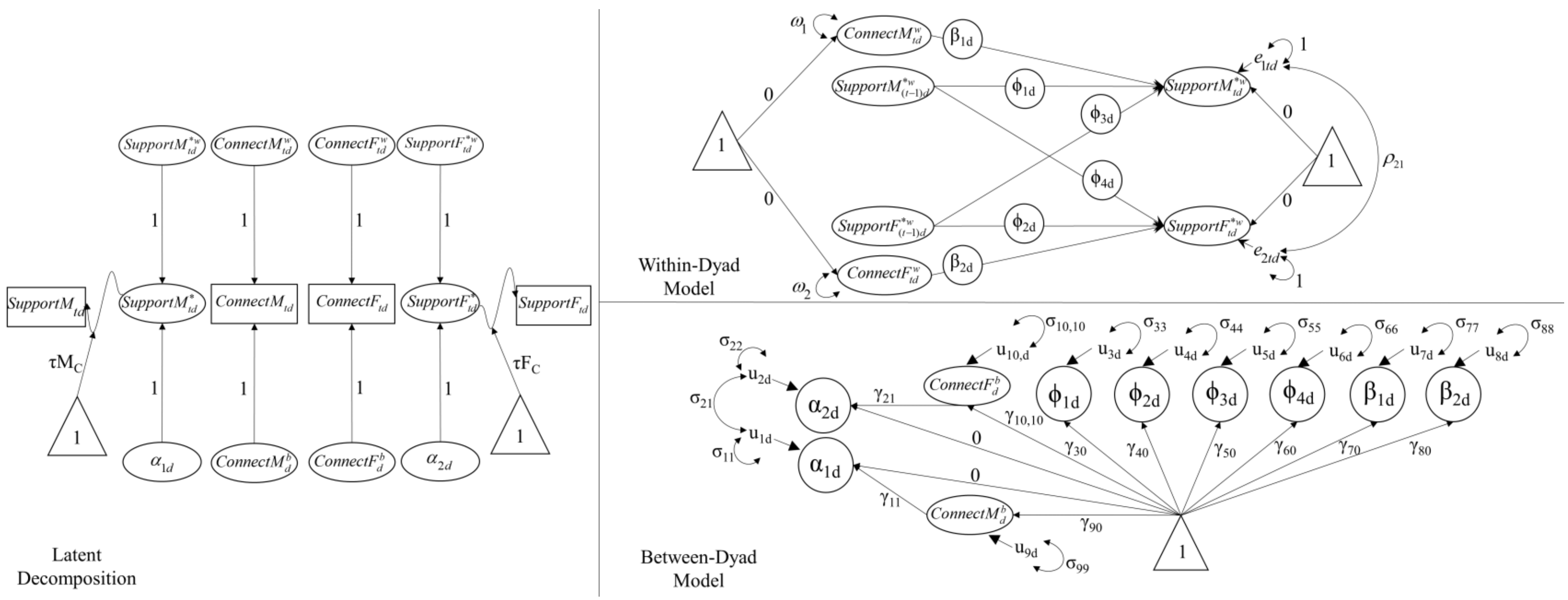

Figure 9. Threshold Locations at Male Latent Person-Mean for Feeling Supported (Center), as well as for 1 Standard Deviation Below and Above (Top and Bottom, Respectively) the Male Latent Person-Mean Average

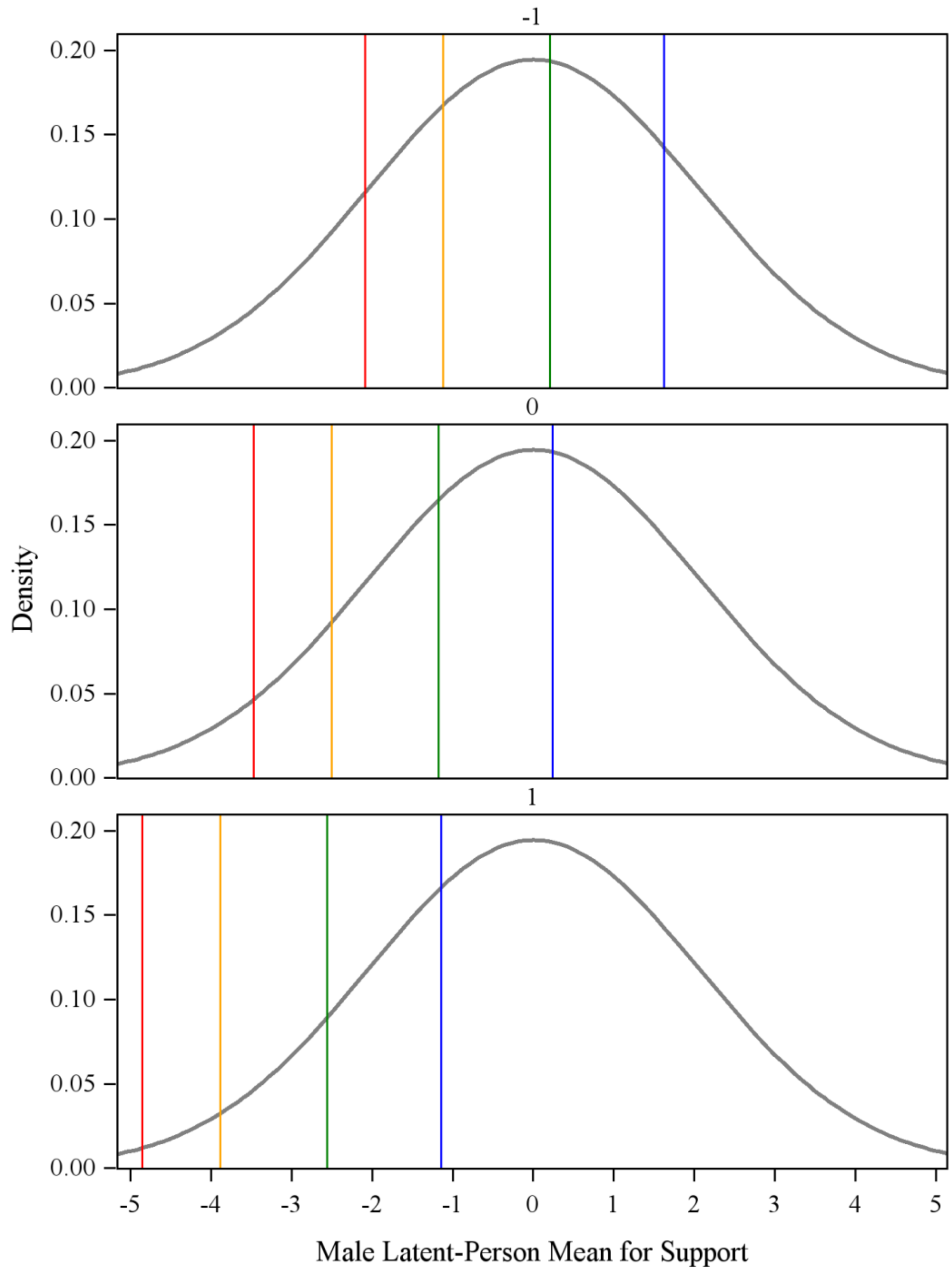

Figure 10. The Generating Model for All Univariate DSEM Models with a Categorical Outcome and Two Continuous Time-Varying Covariates

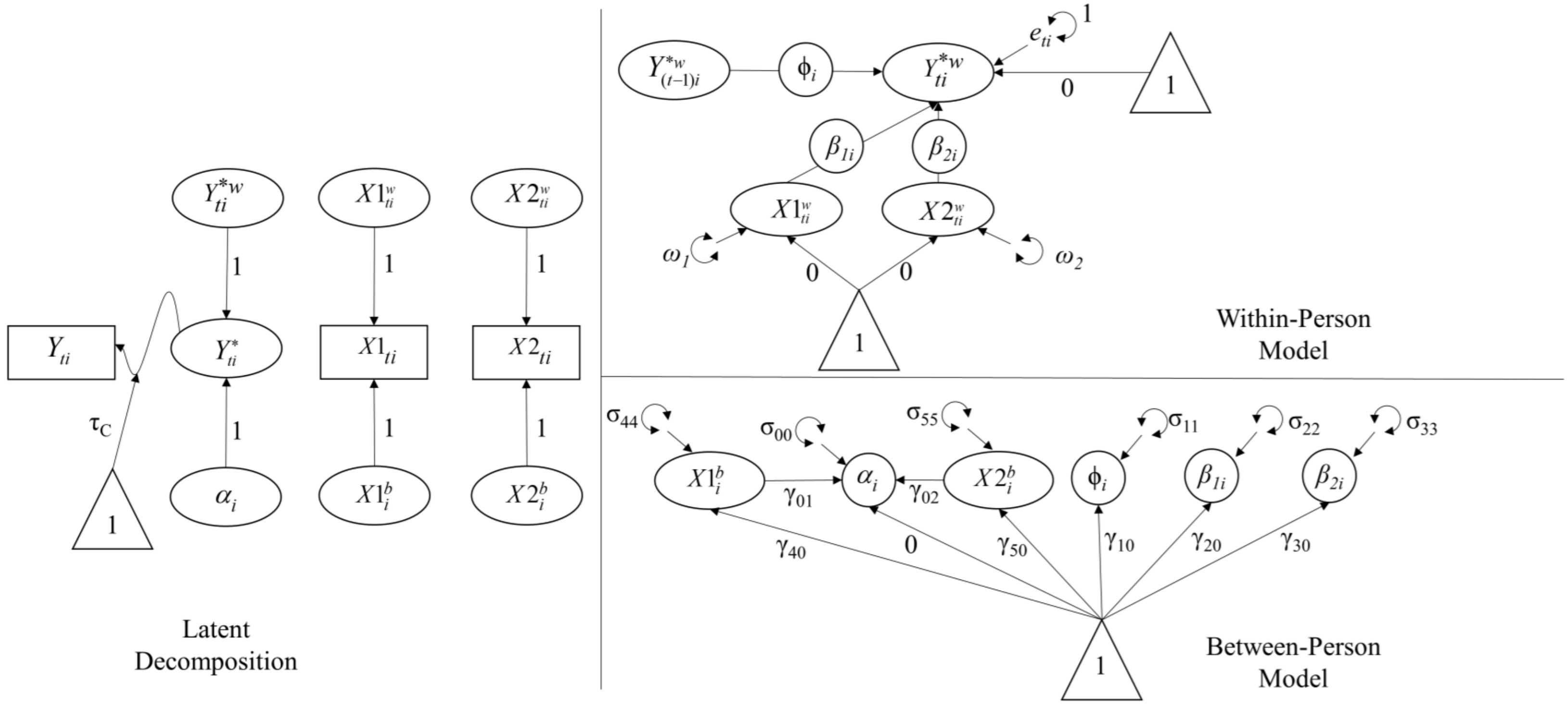

Figure 11. The Generating Model for All Multivariate DSEM Models with Two Categorical Outcomes and Two Time-Varying Covariates; Represents L-APIM

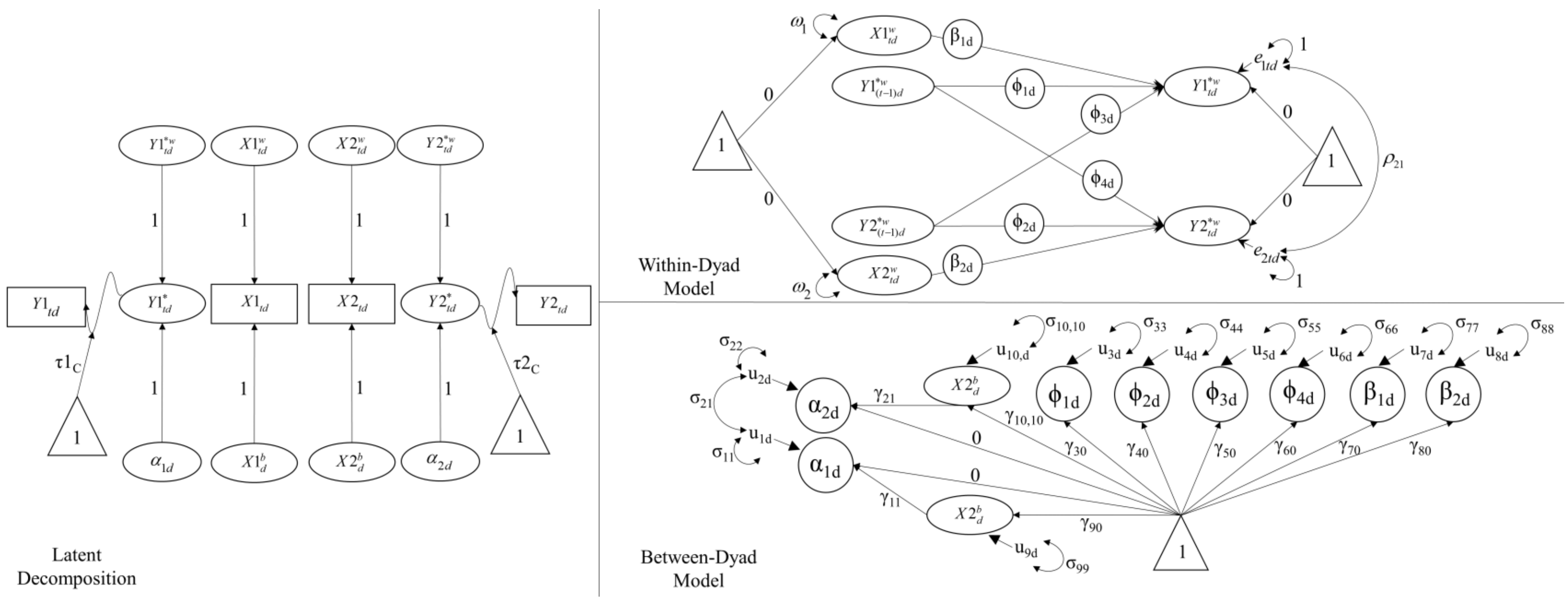

Figure 12. Average Bias for Actor and Partner Effects across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

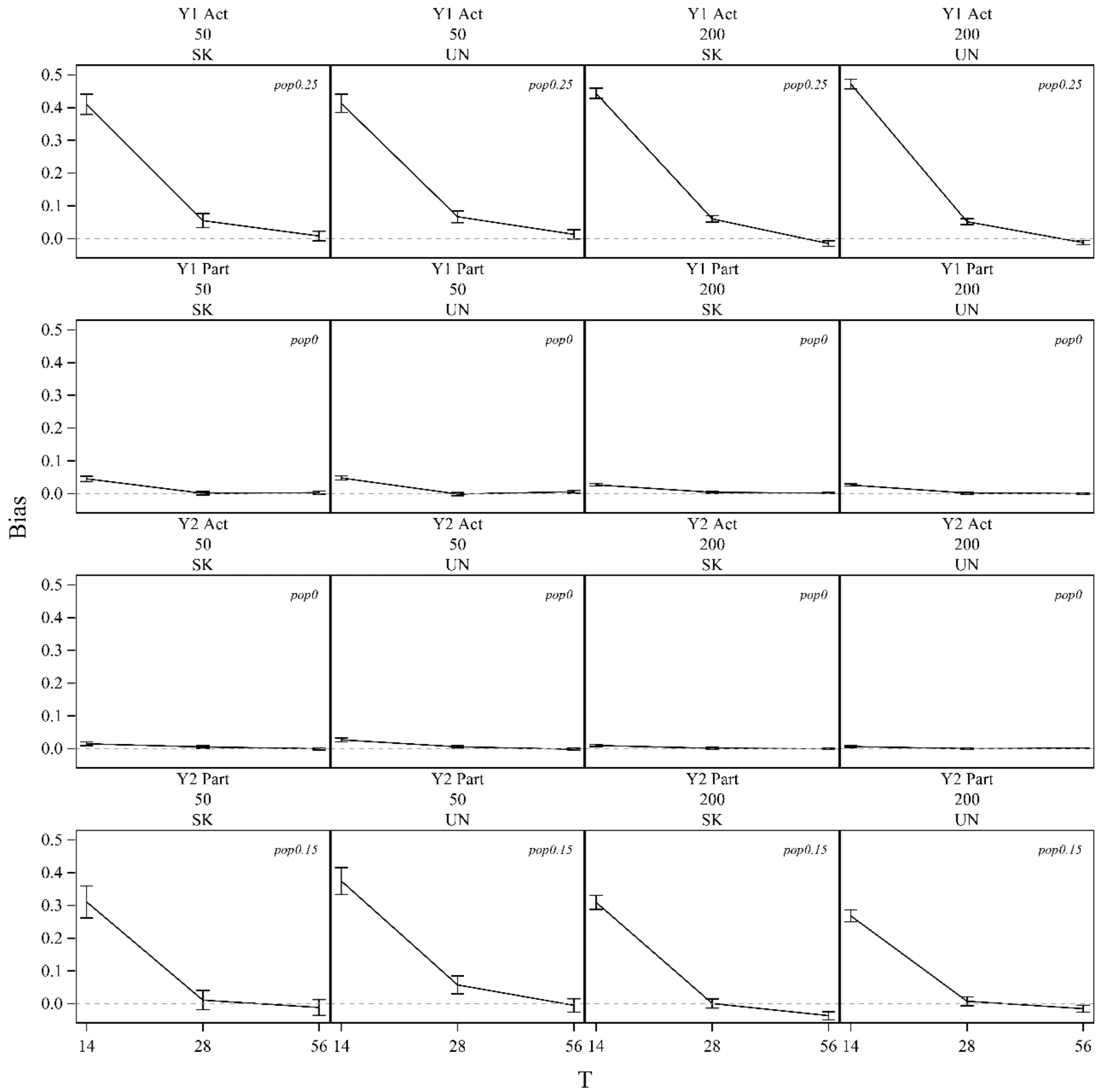

Figure 13. Average Bias for Autoregressive Effects across Simulation Conditions for Categorical Univariate DSEM Models with 2-Category Outcomes

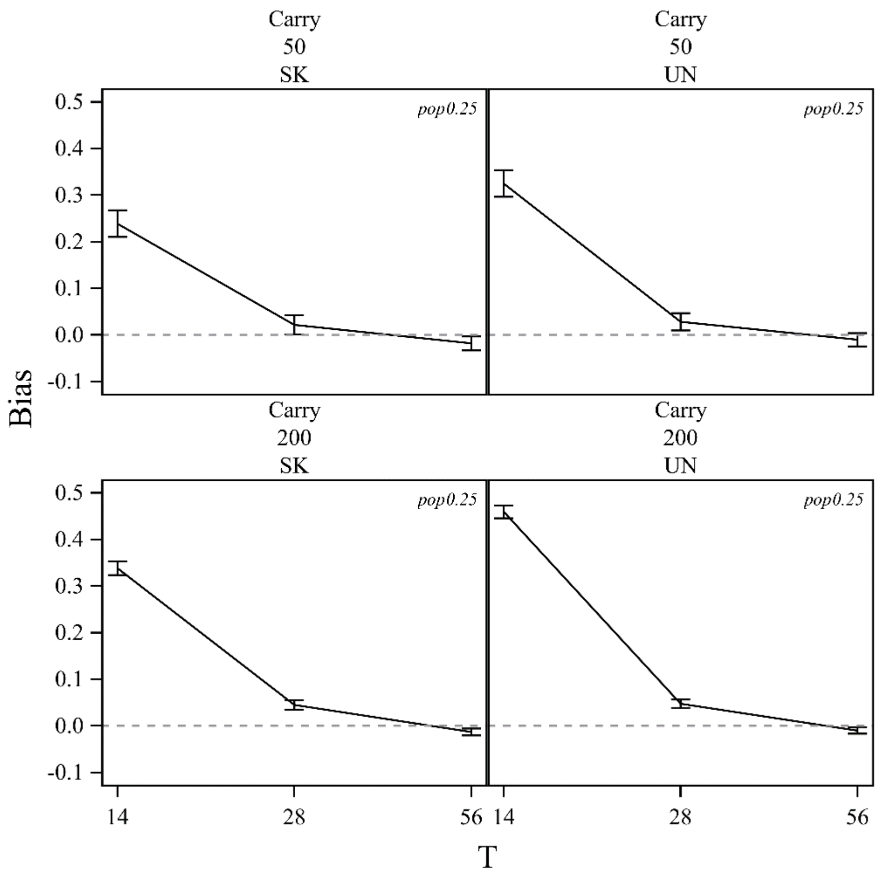

Figure 14. Average Bias for Actor and Partner Random Effect Variances across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

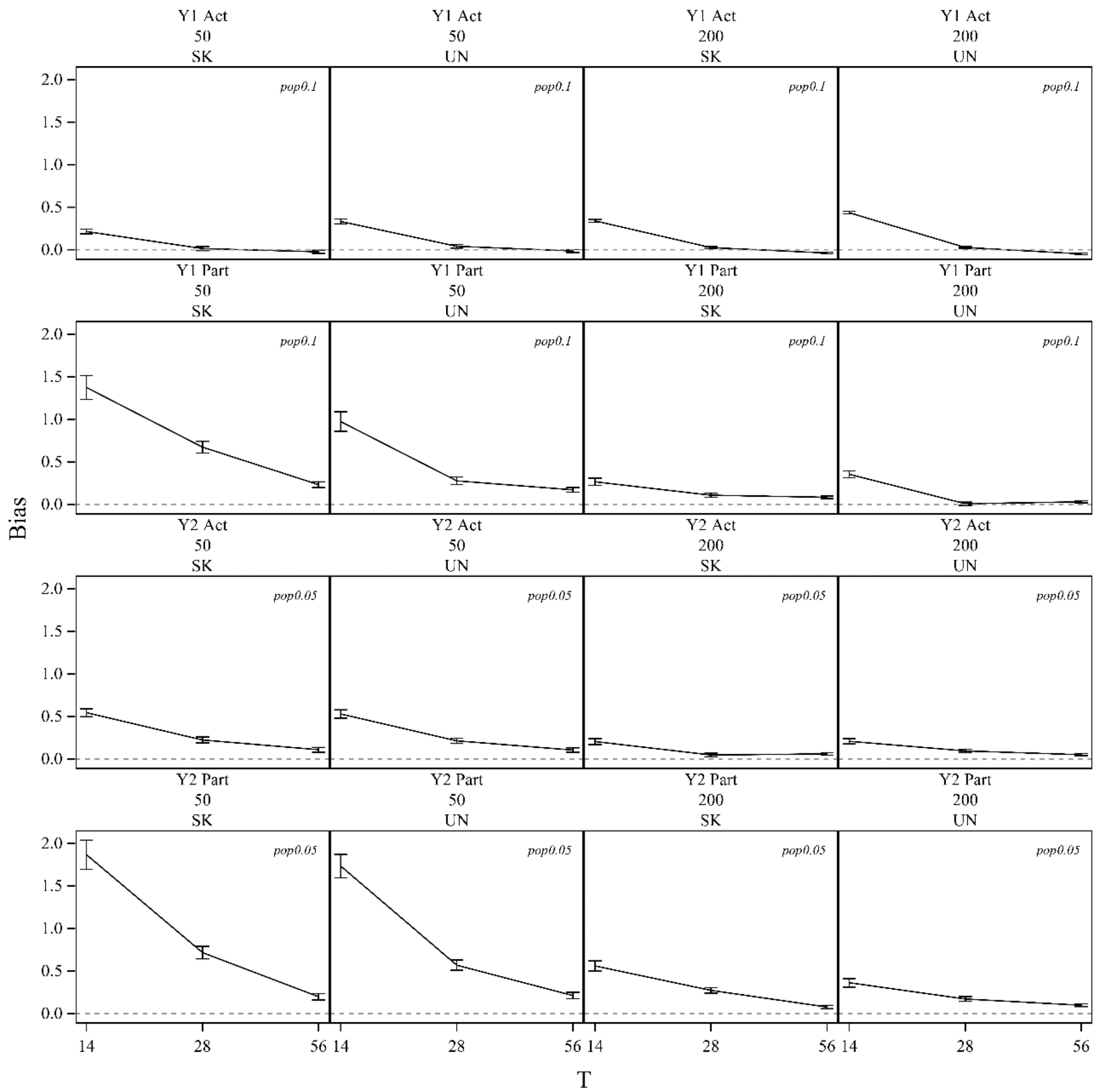

Figure 15. Average Bias for Autoregressive Random Effect Variances across Simulation Conditions for Univariate Categorical DSEM Models with 2-Category Outcomes

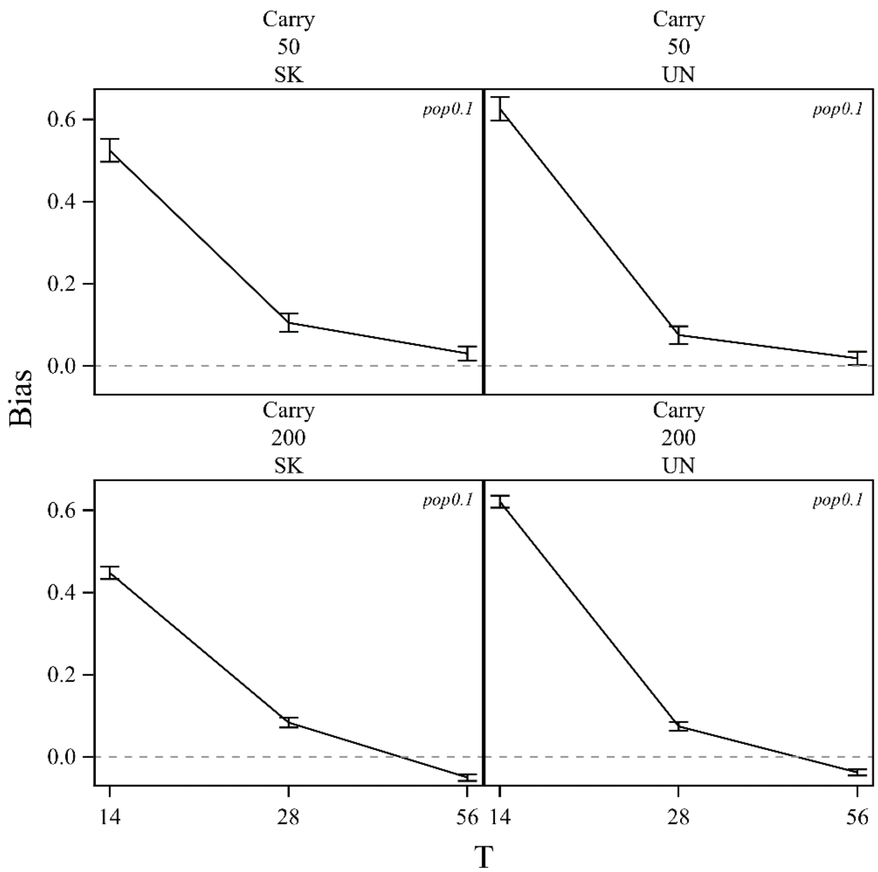

Figure 16. Average Bias for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

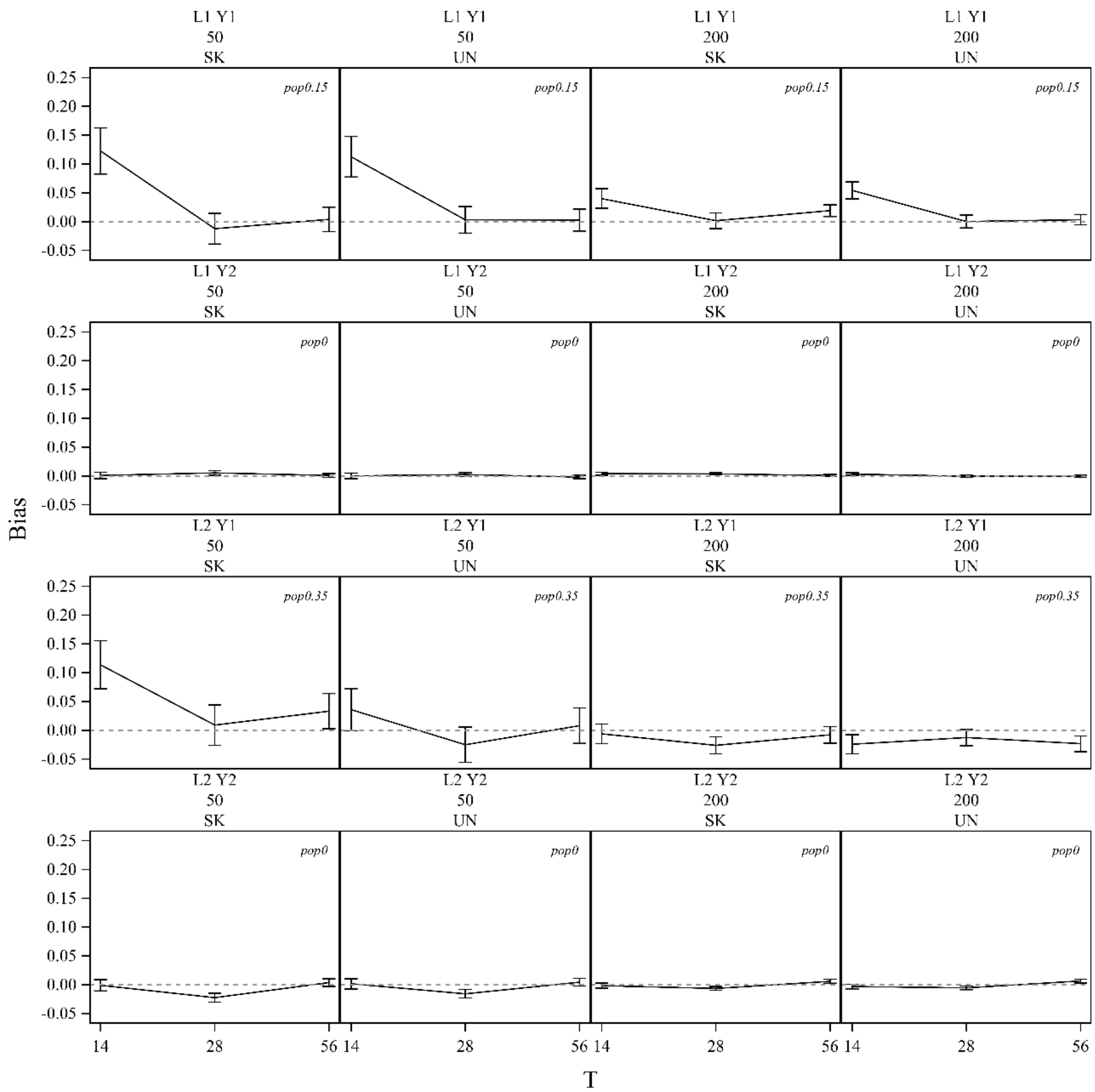

Figure 17. Average Bias for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical DSEM Models with 2-Category Outcomes

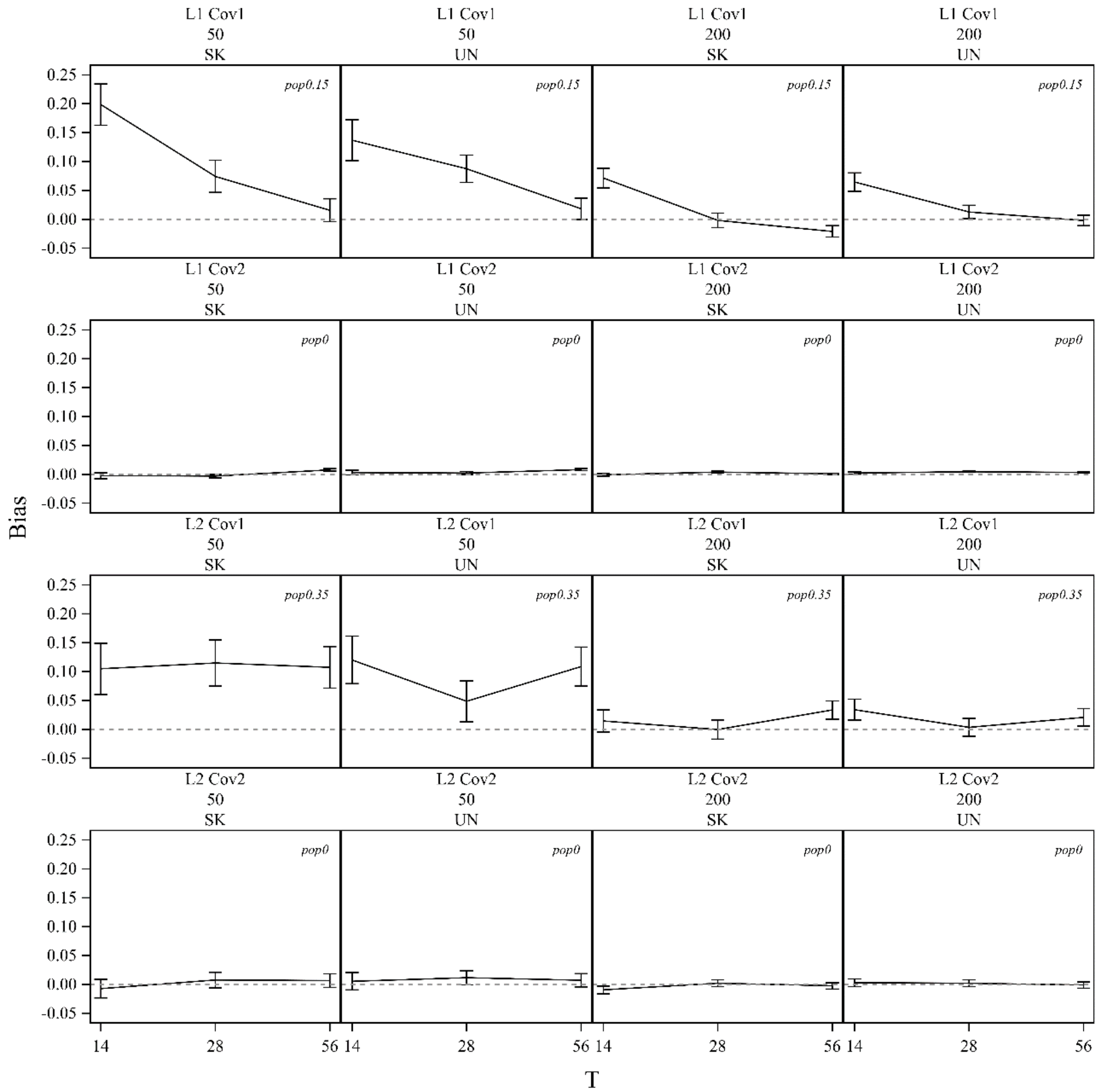

Figure 18. Average Bias for Level-1 Covariate Random Effect Variances across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

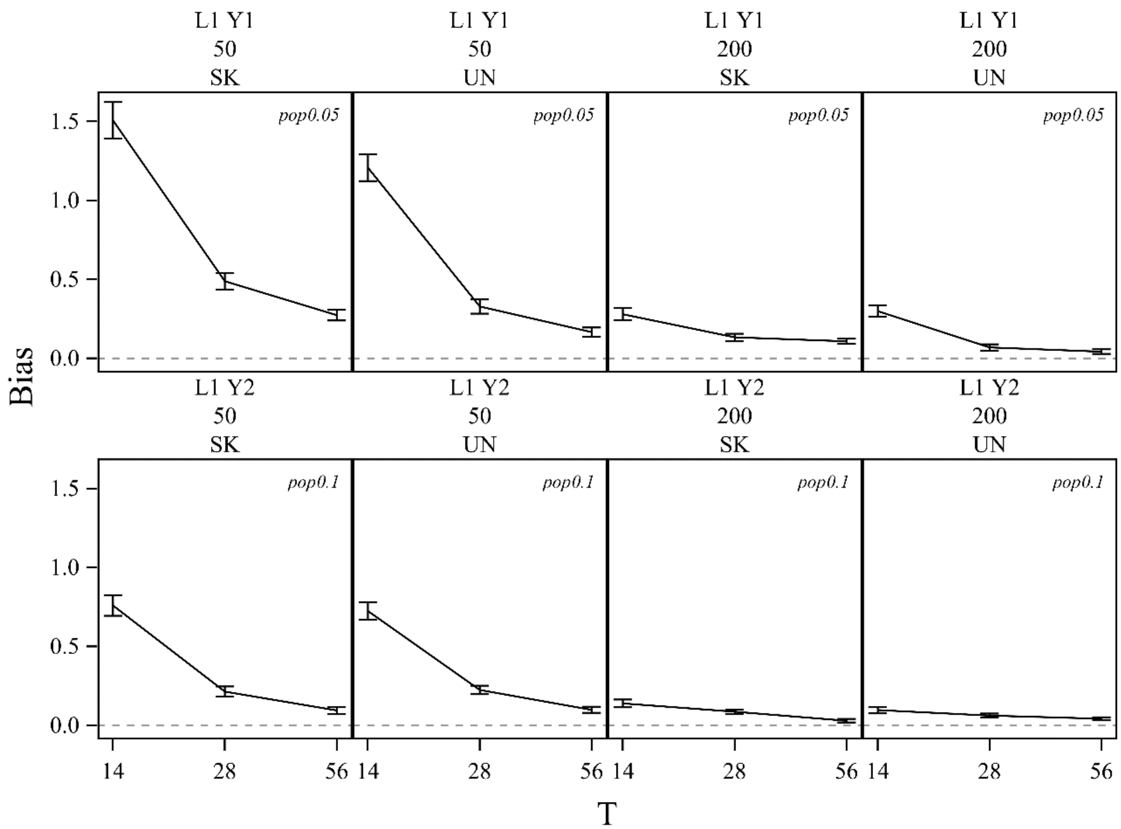

Figure 19. Average Bias for Level-1 Covariate Random Effect Variances across Simulation Conditions for Univariate Categorical DSEM Models with 2-Category Outcomes

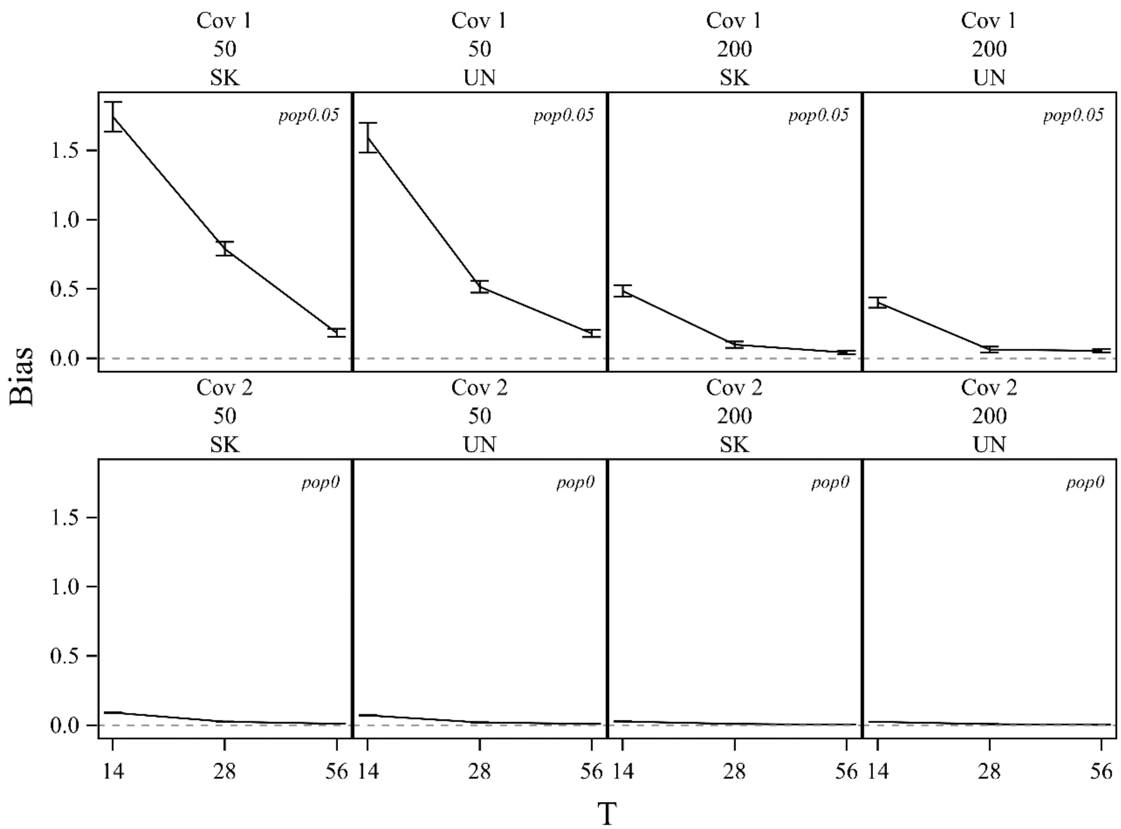

Figure 20. Efficacy Results for Actor and Partner Effects across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

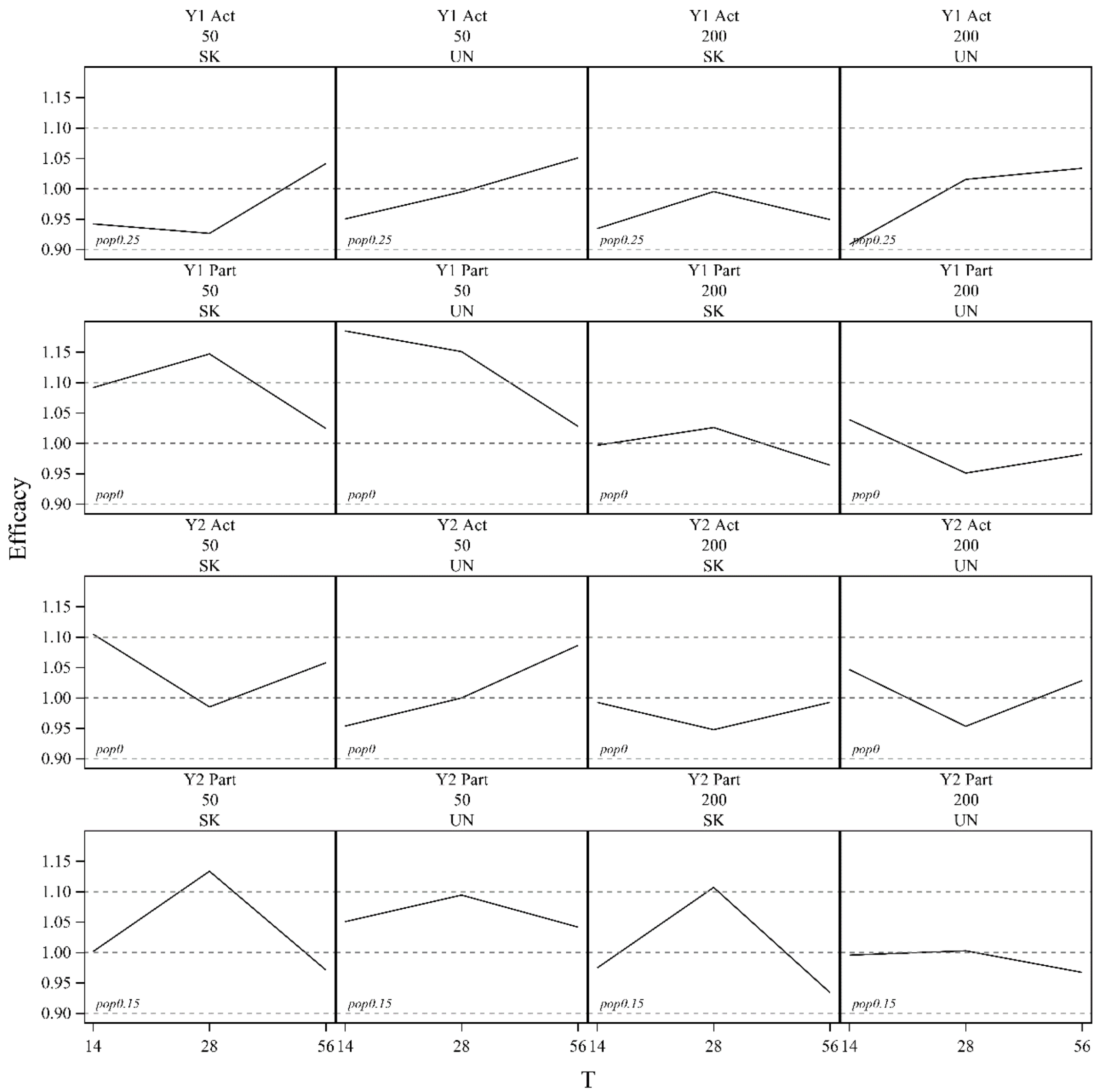

Figure 21. Efficacy Results for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

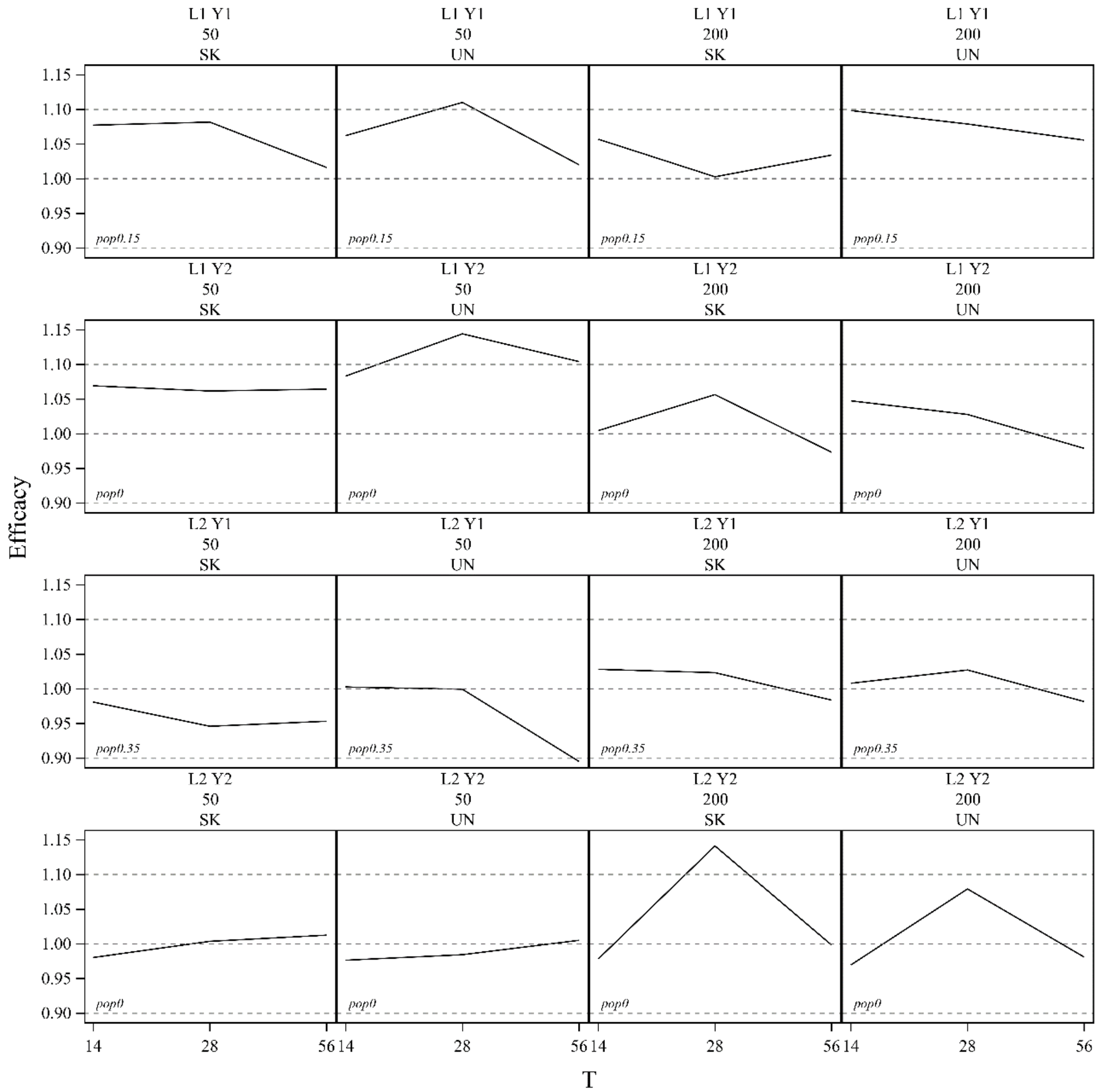

Figure 22. Efficacy Results for Autoregressive Effects across Simulation Conditions for Univariate Categorical DSEM Models with 2-Category Outcomes

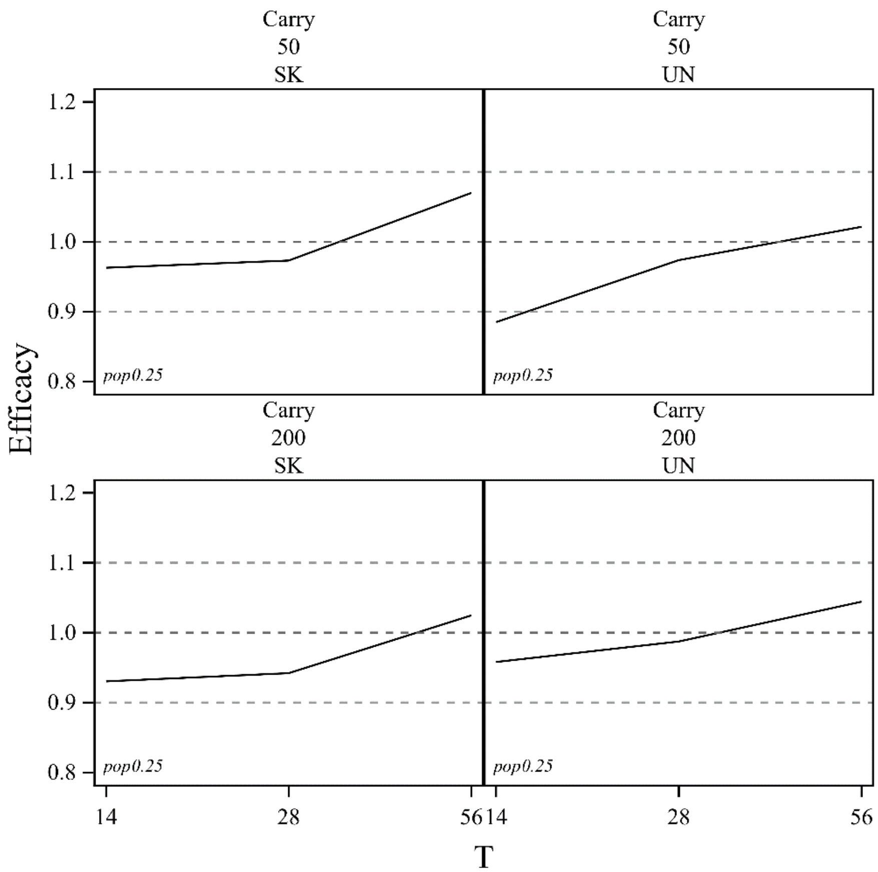

Figure 23. Efficacy Results for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical DSEM Models with 2-Category Outcomes

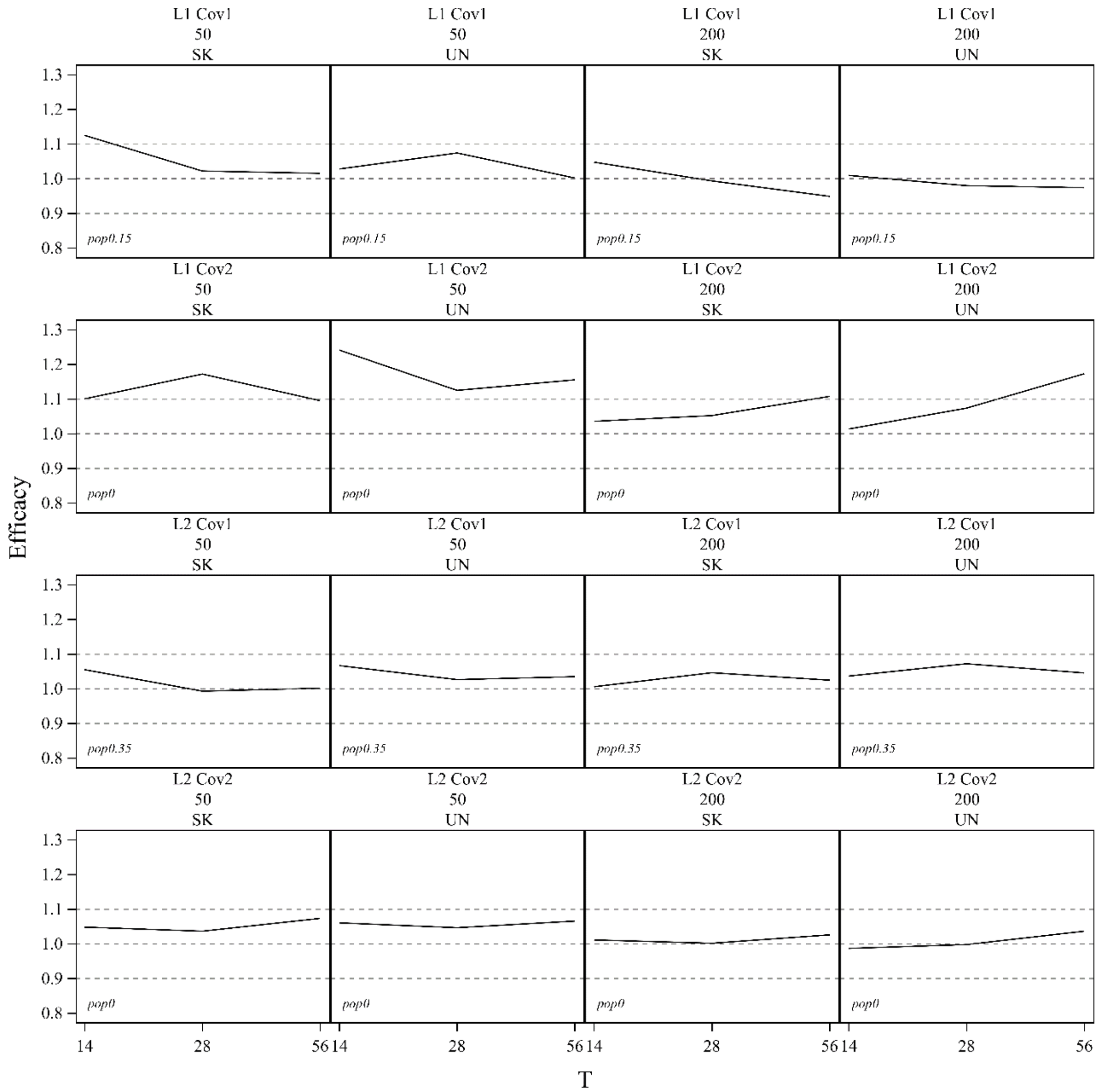

Figure 24. Efficacy Results for Actor and Partner Random Effect Variances across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

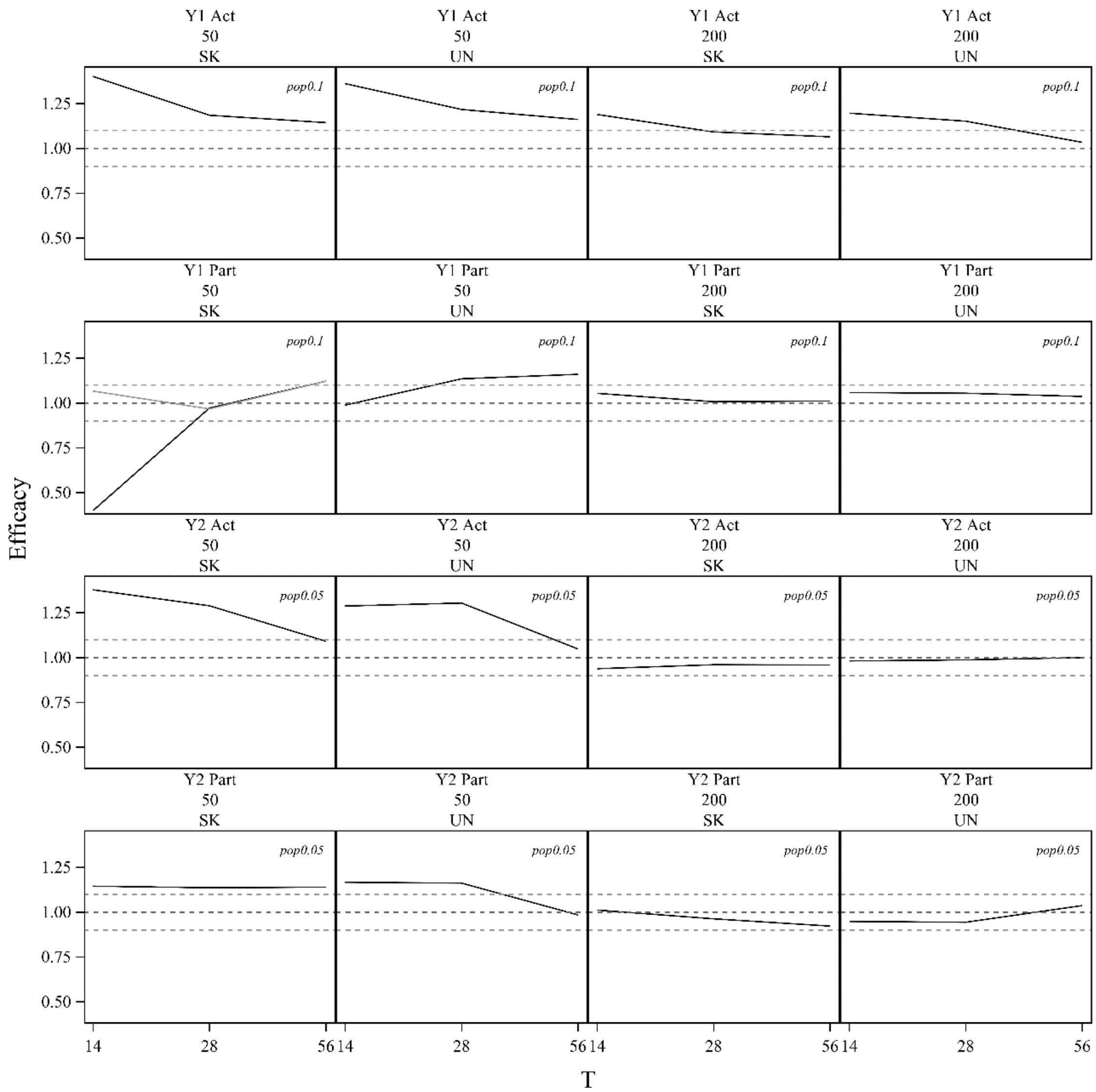

Figure 25. Efficacy Results for Level-1 Covariate Random Effect Variances across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

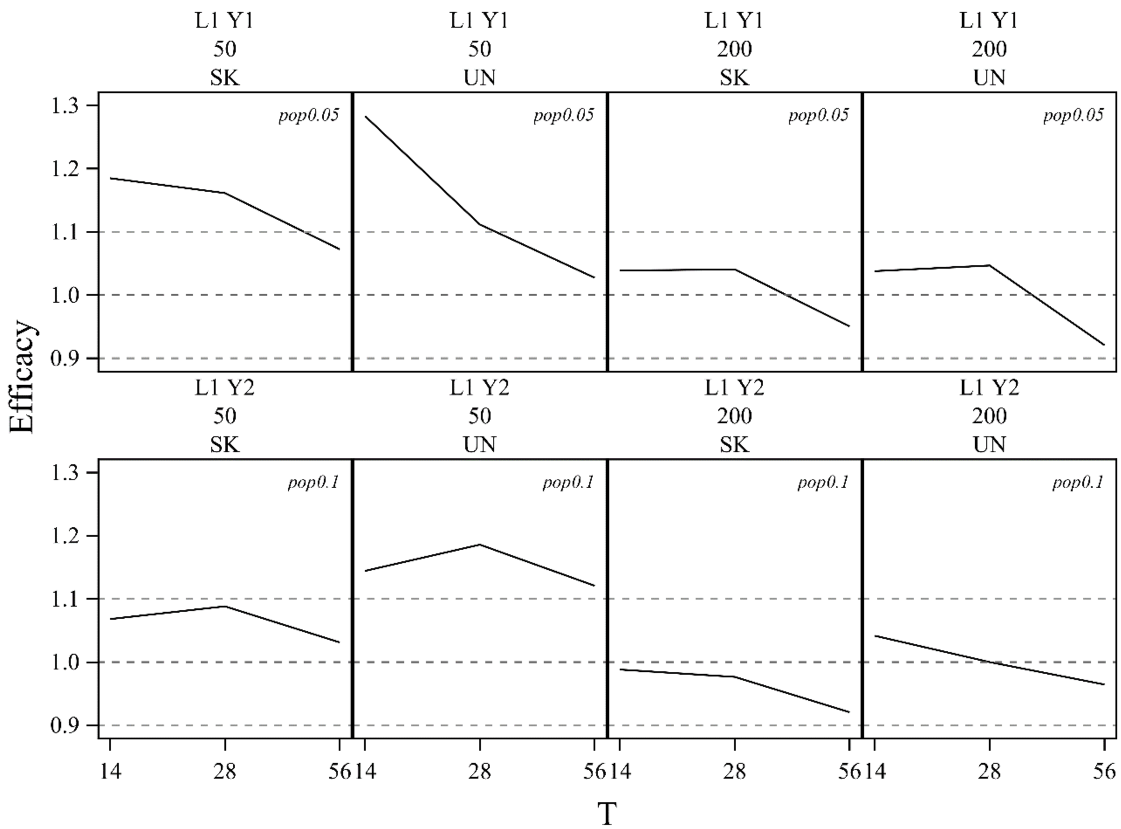

Figure 26. Efficacy Results for Autoregressive Random Effect Variances across Simulation Conditions for Univariate Categorical DSEM Models with 2-Category Outcomes

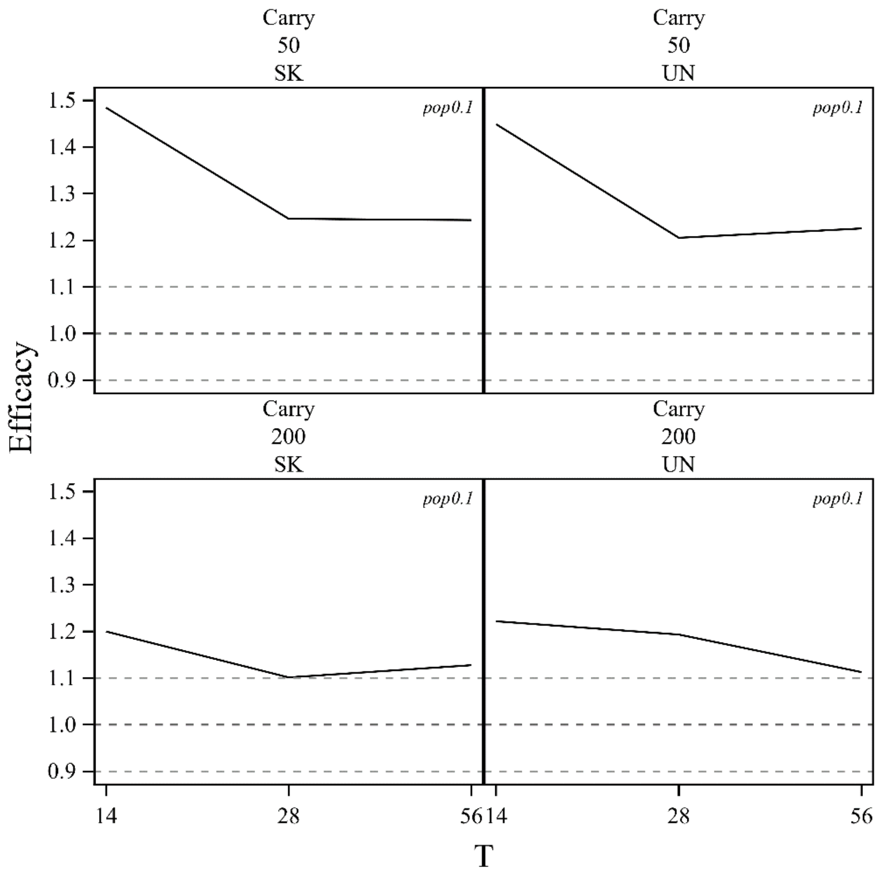

Figure 27. Efficacy Results for Level-1 Covariate Random Effect Variances across Simulation Conditions for Univariate Categorical DSEM Models with 2-Category Outcomes

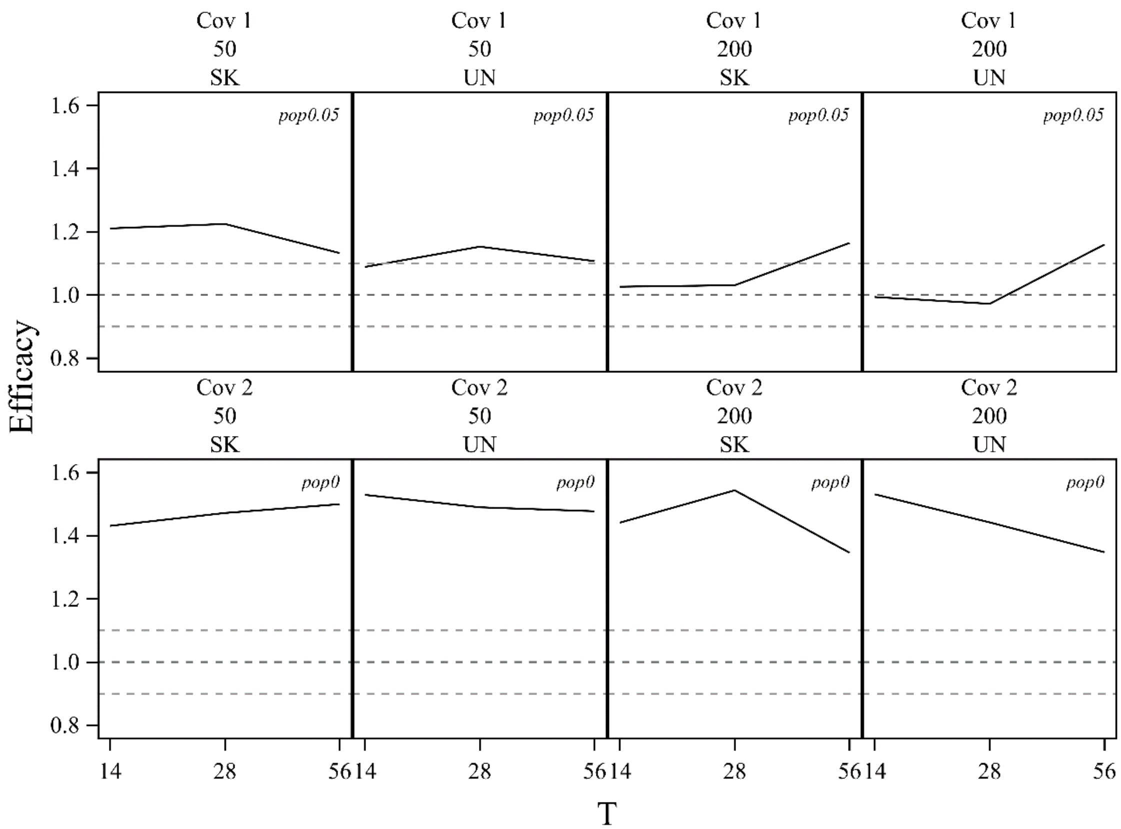

Figure 28. Power to Detect Significant Actor and Partner Effects across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

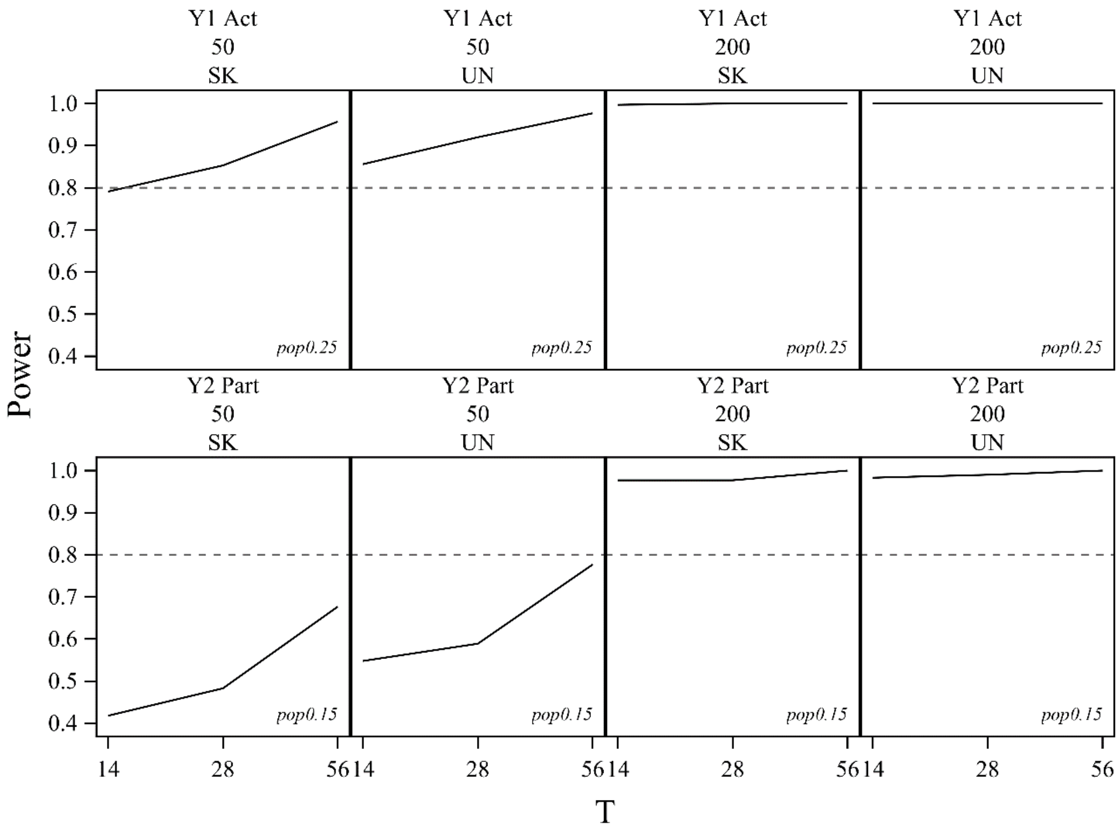

Figure 29. Type-I Error Rates for Actor and Partner Effects across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

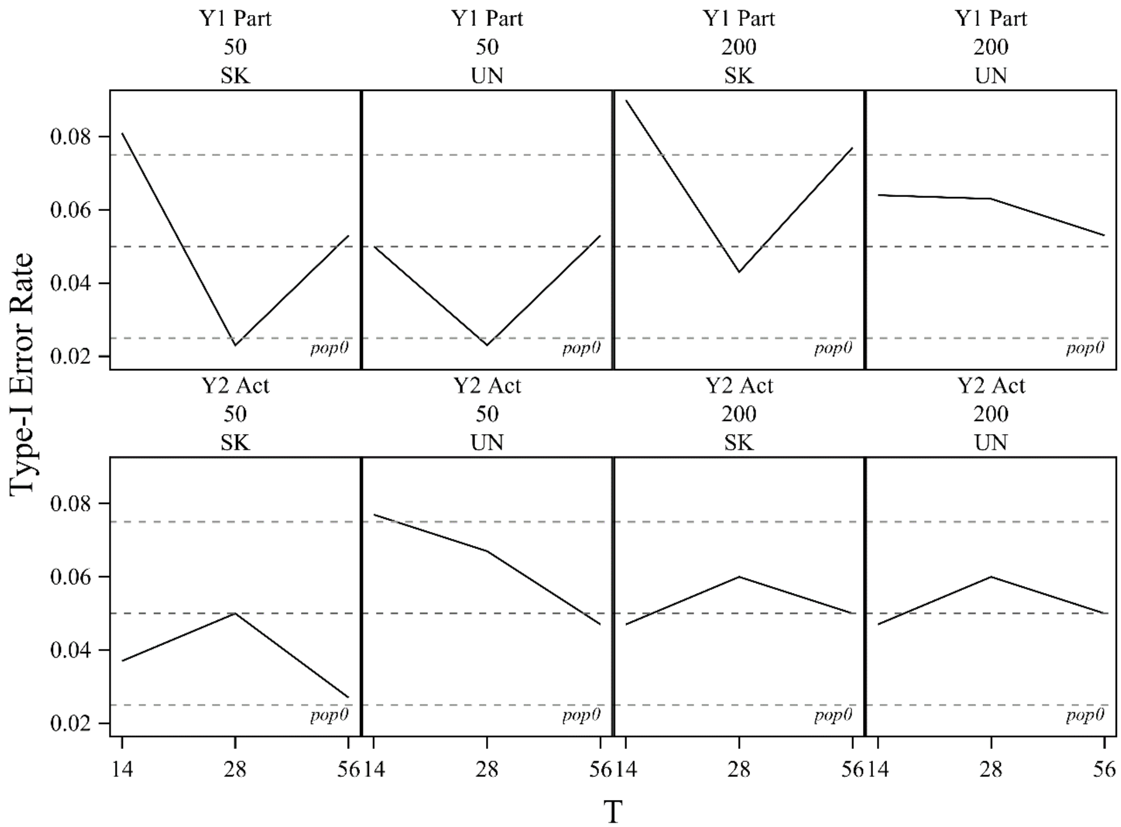

Figure 30. Power to Detect Significant Autoregressive Effects across Simulation Conditions for Categorical Univariate DSEM Models with 2-Category Outcomes

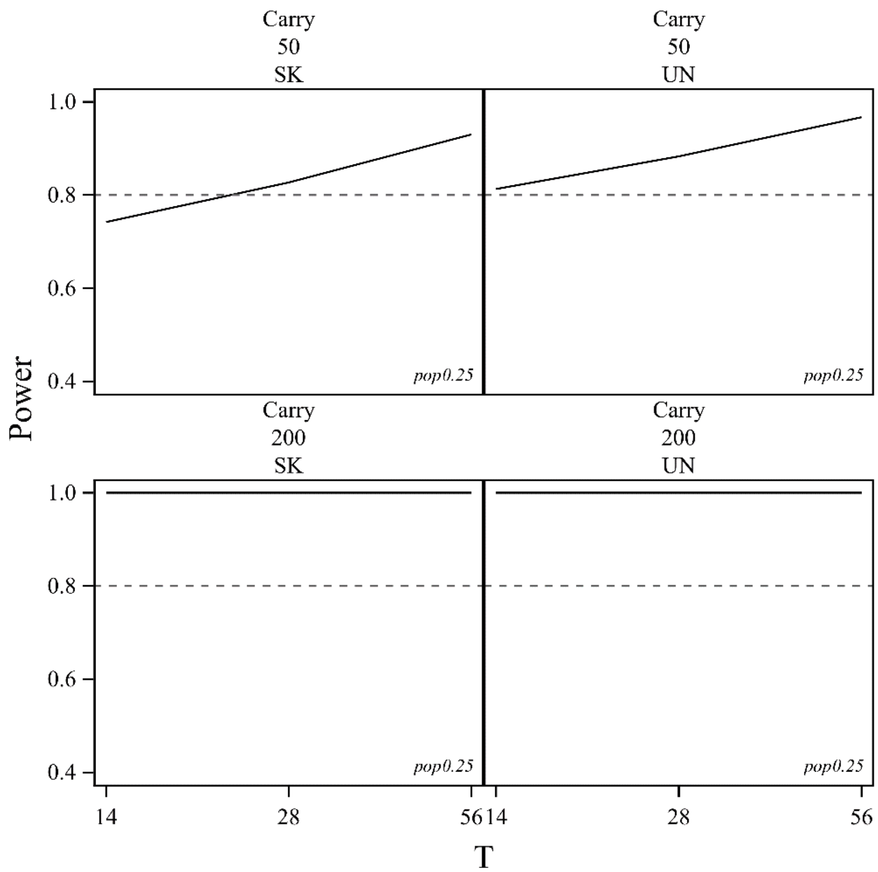

Figure 31. Estimated Type-I Error Rates for Autoregressive Effects across Simulation Conditions for Categorical Univariate DSEM Models with 2-Category Outcomes

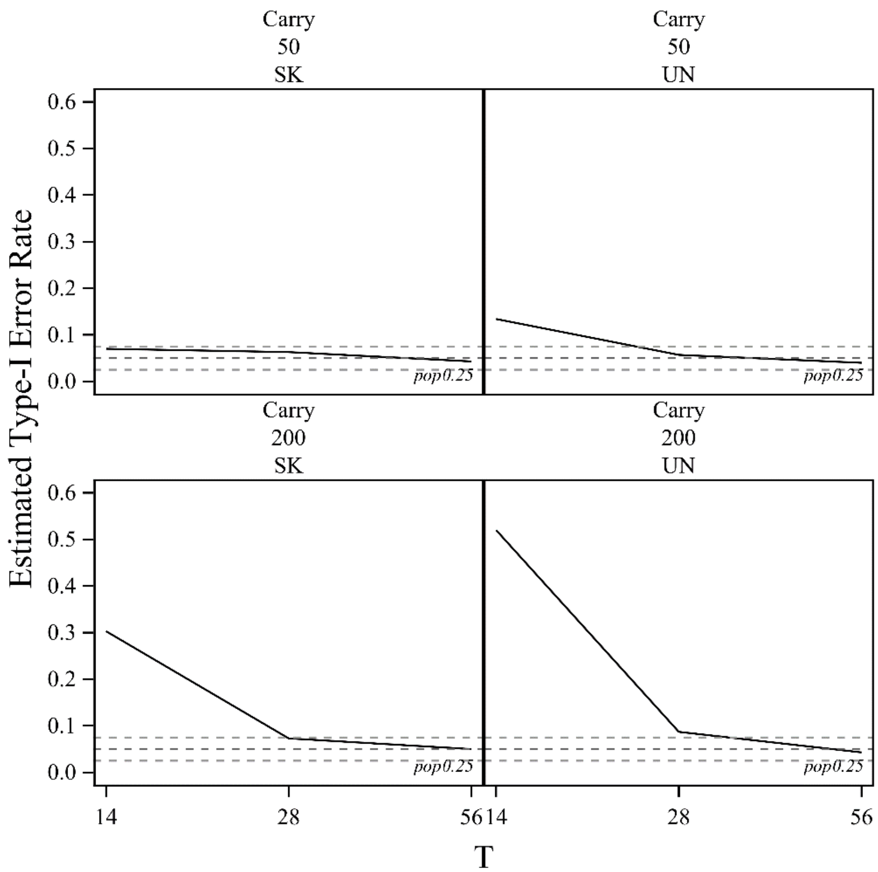

Figure 32. Power to Detect Significant Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

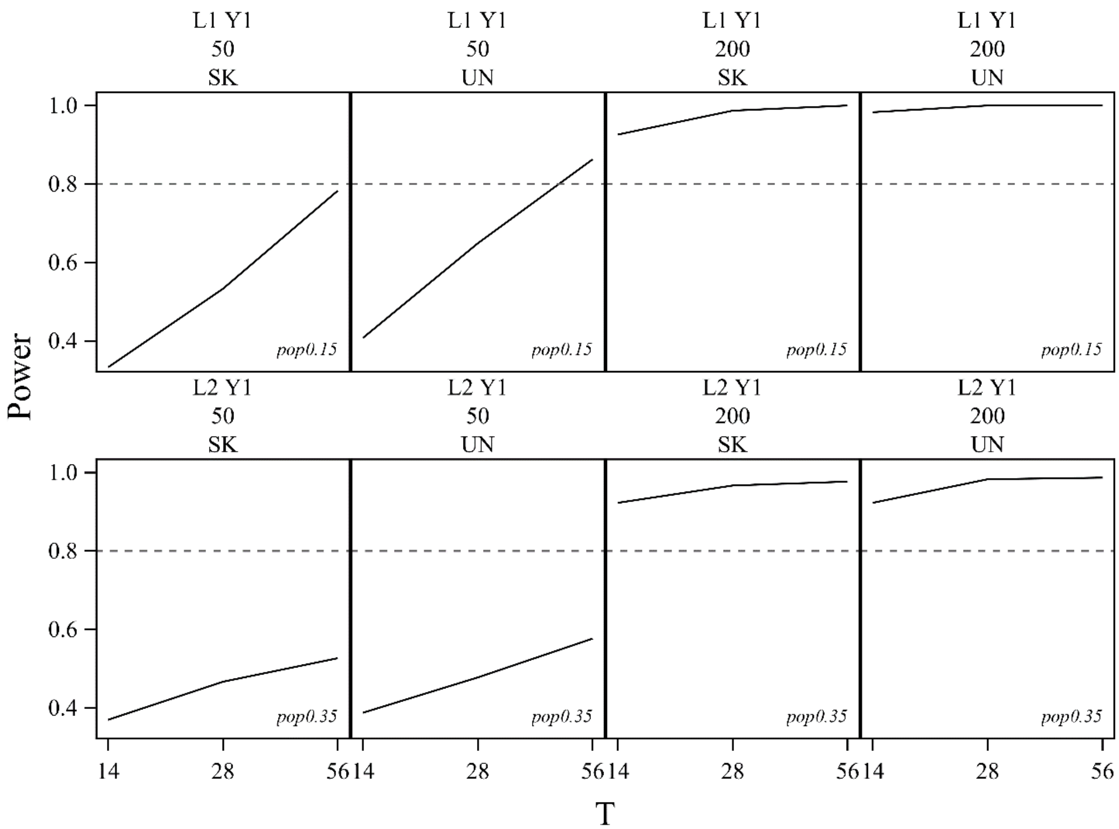

Figure 33. Type-I Error Rates for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical DSEM Models with 2-Category Outcomes

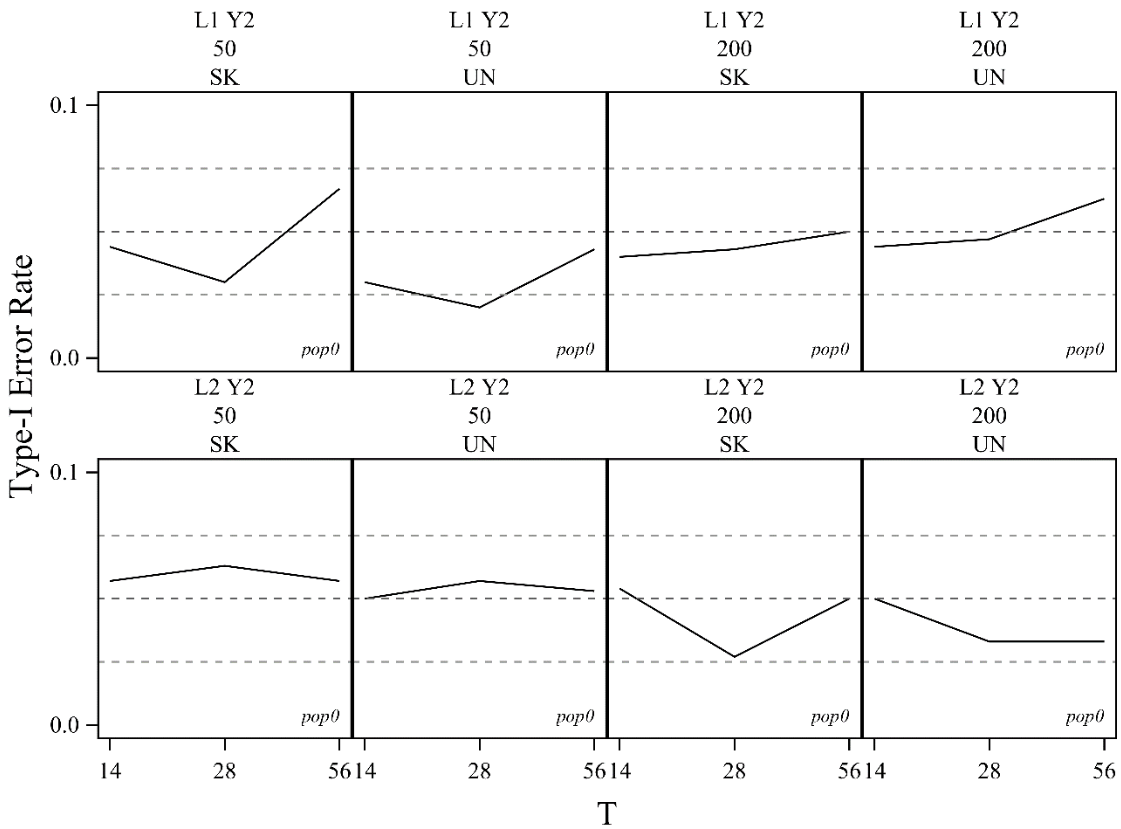

Figure 34. Power to Detect Significant Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical DSEM Models with 2-Category Outcomes

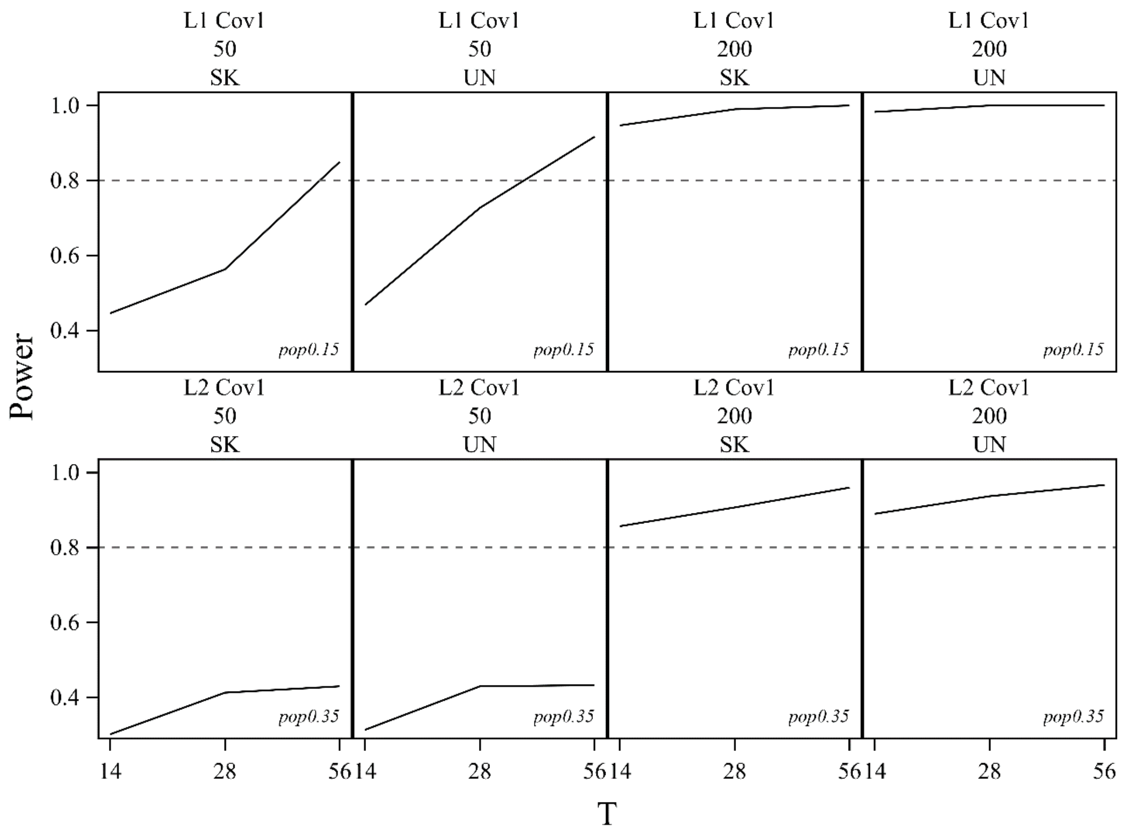

Figure 35. Type-I Error Rates for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical DSEM Models with 2-Category Outcomes

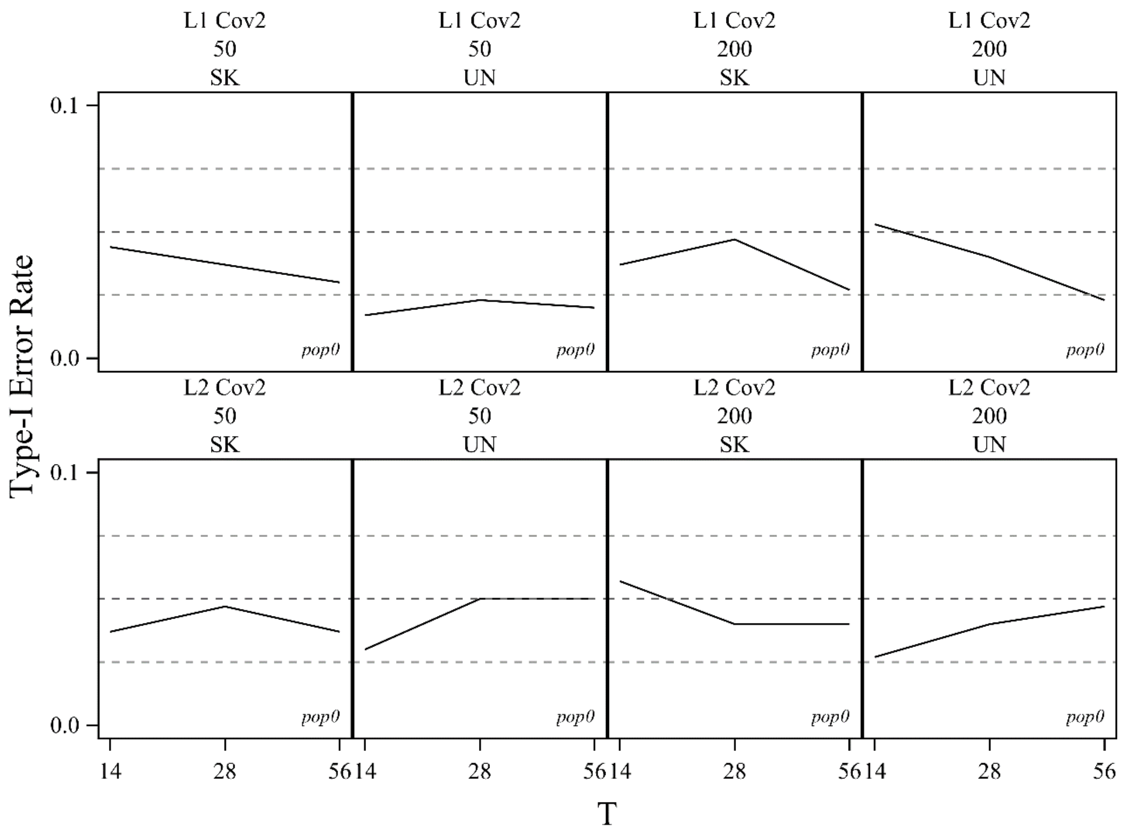

Figure 36. Average Bias for Actor and Partner Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

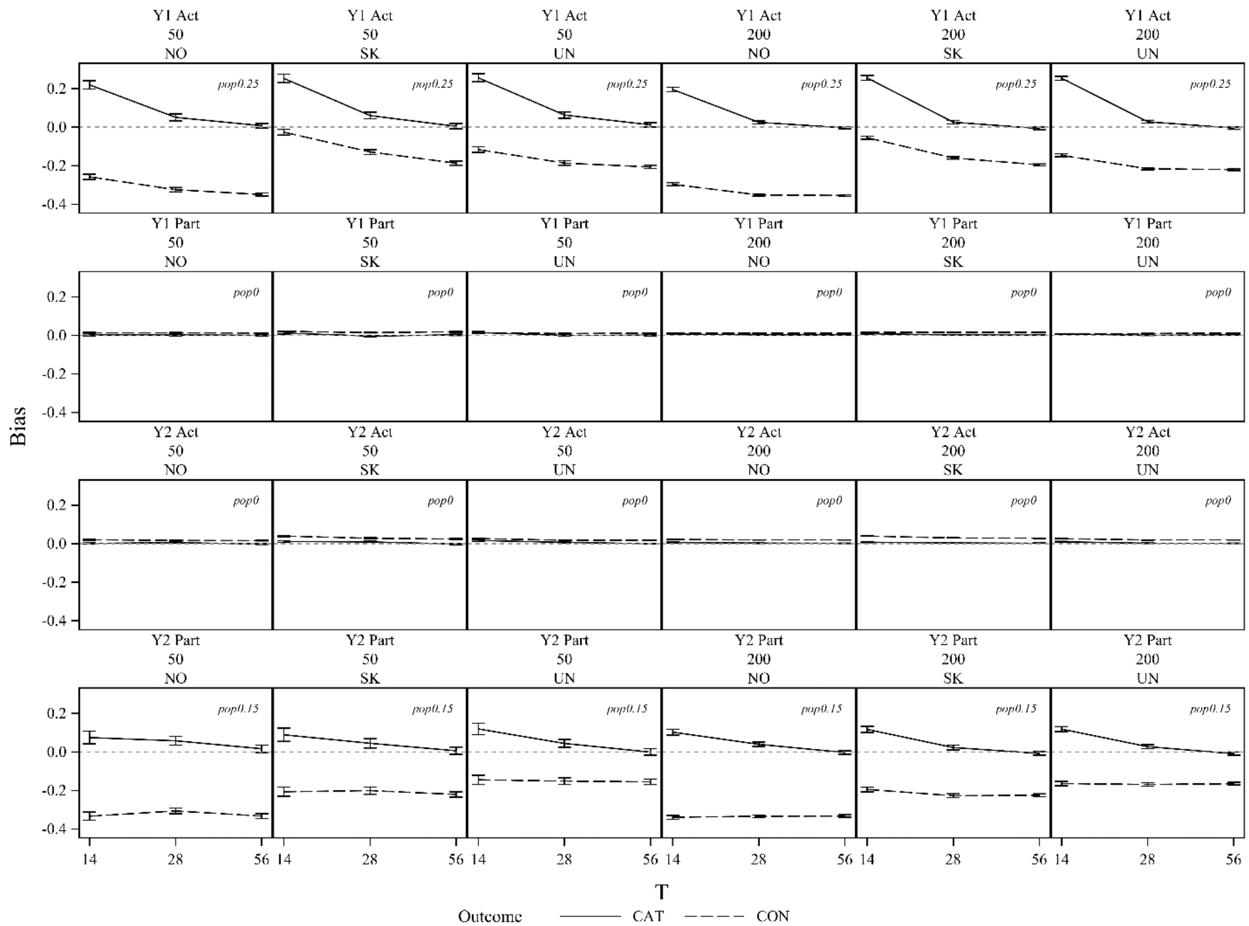

Figure 37. Average Bias for Autoregressive Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

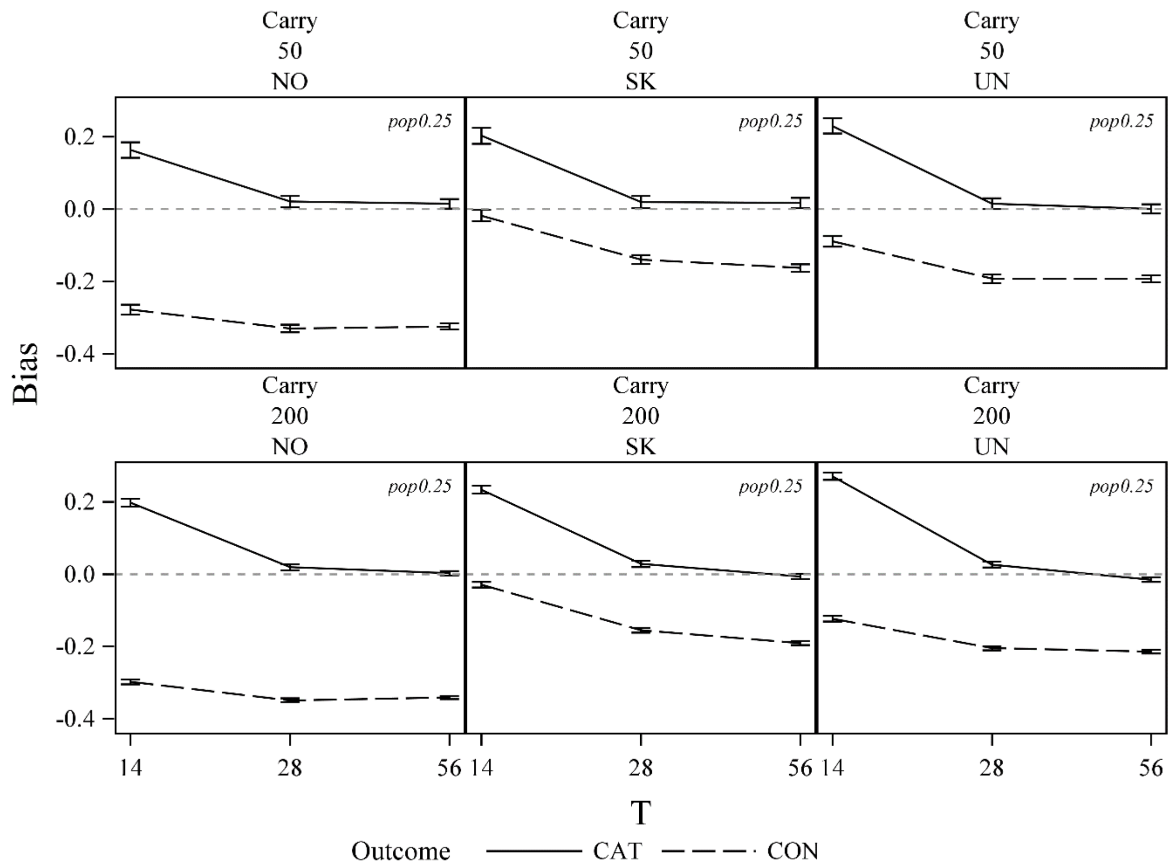

Figure 38. Average Bias for Standardized Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

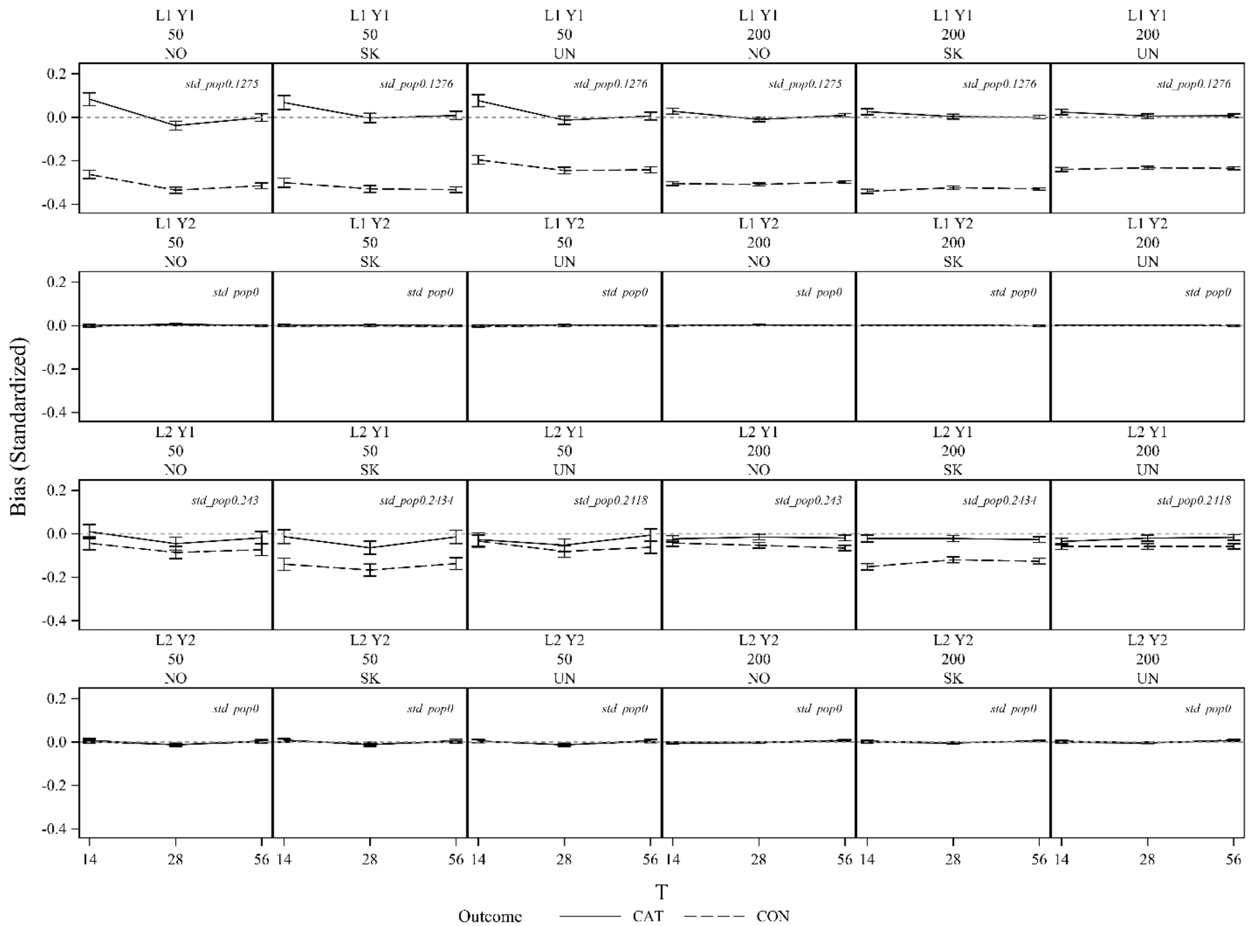

Figure 39. Average Bias for Standardized Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

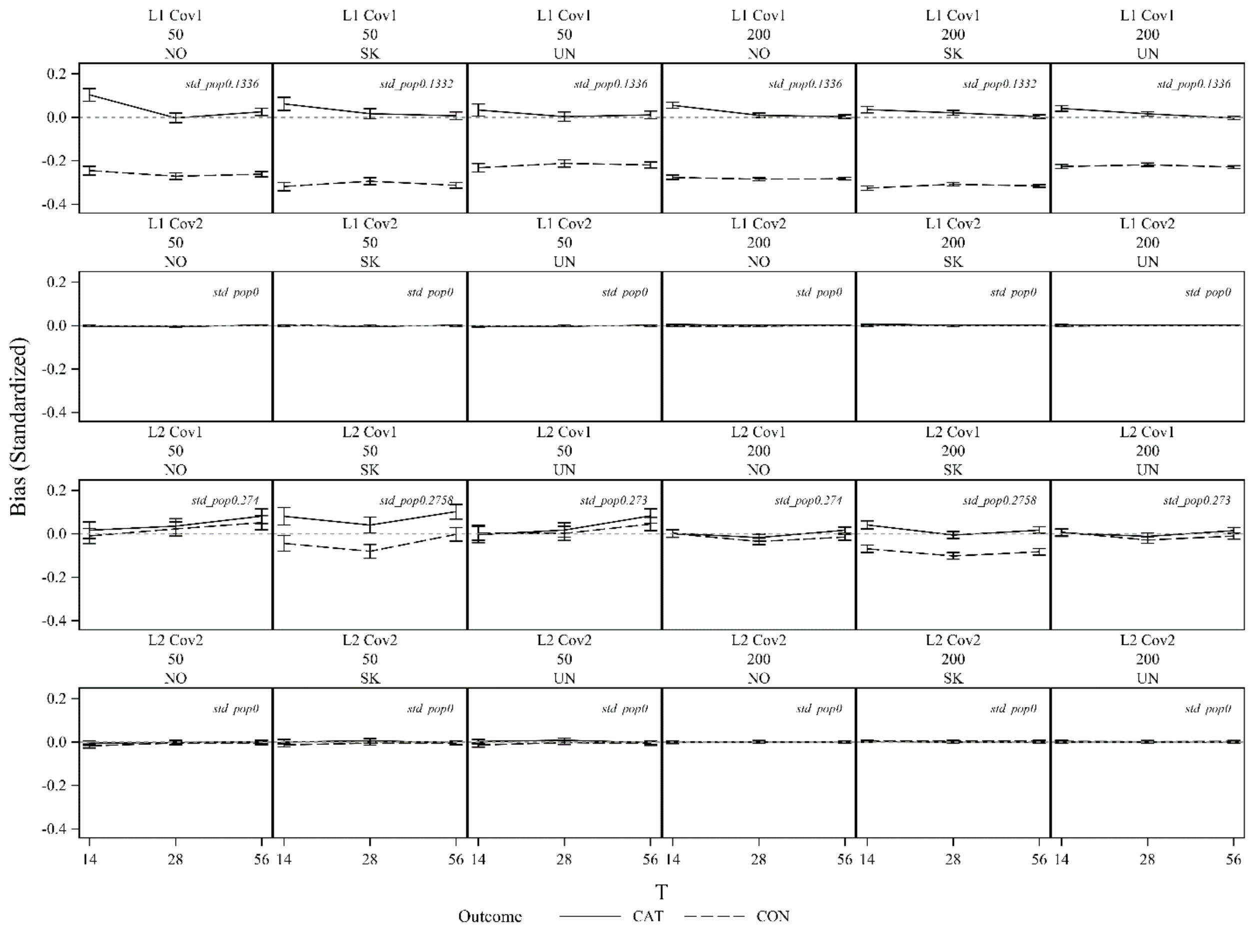

Figure 40. Average Median Value of the Posterior Distributions for Actor and Partner Random Effect Variances across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

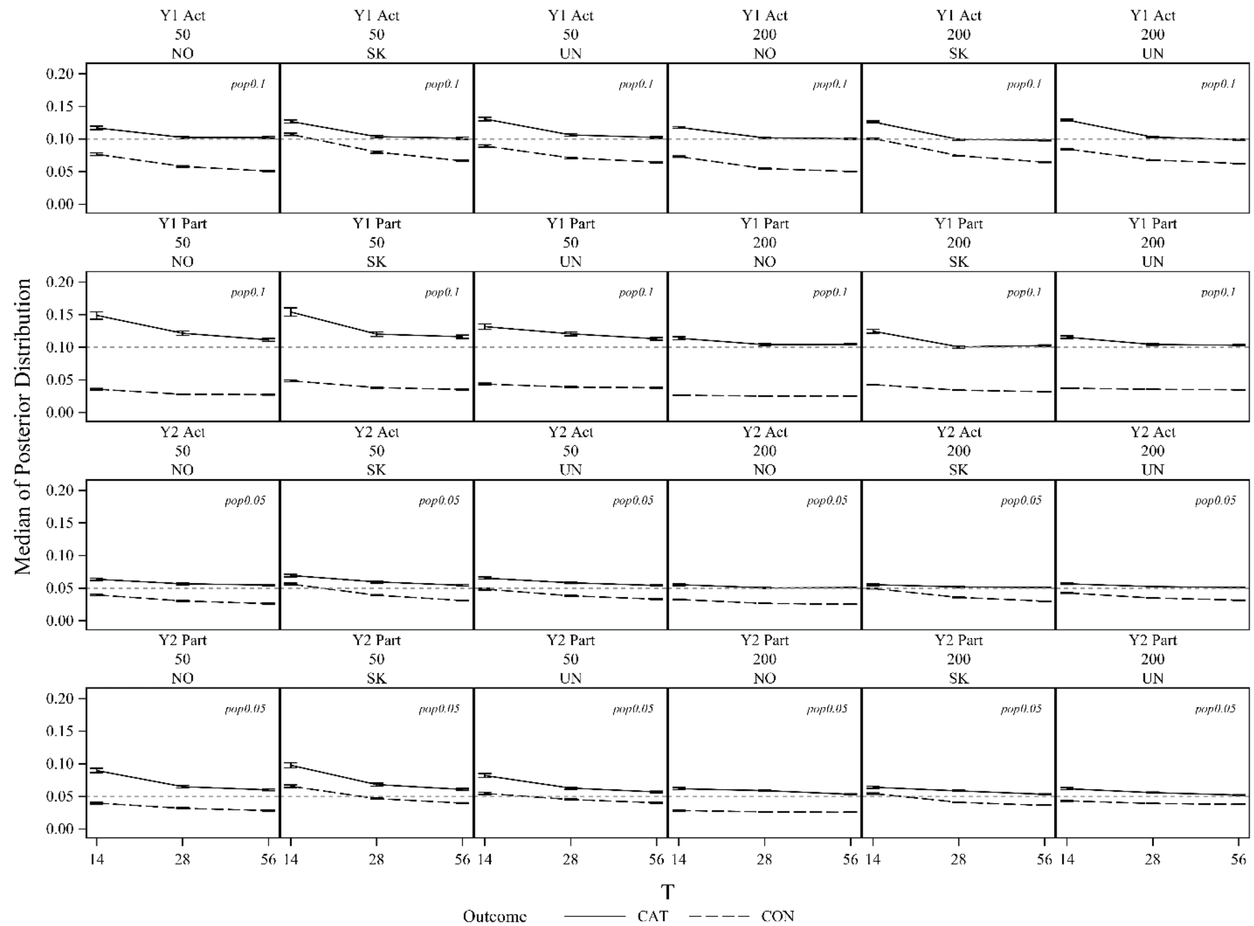

Figure 41. Average Median Value of the Posterior Distributions for Autoregressive Random Effect Variances across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

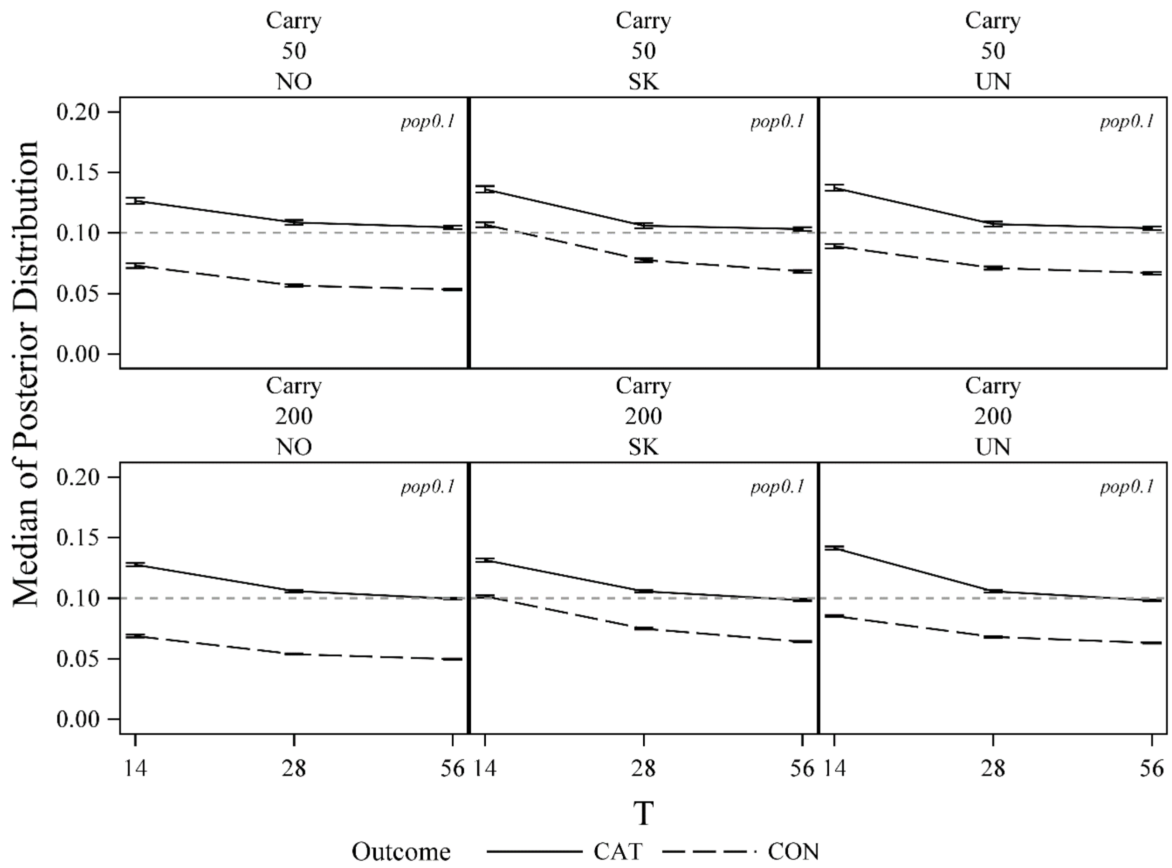

Figure 42. Average Median Value of the Posterior Distributions for Level-1 Covariate Random Effect Variances across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

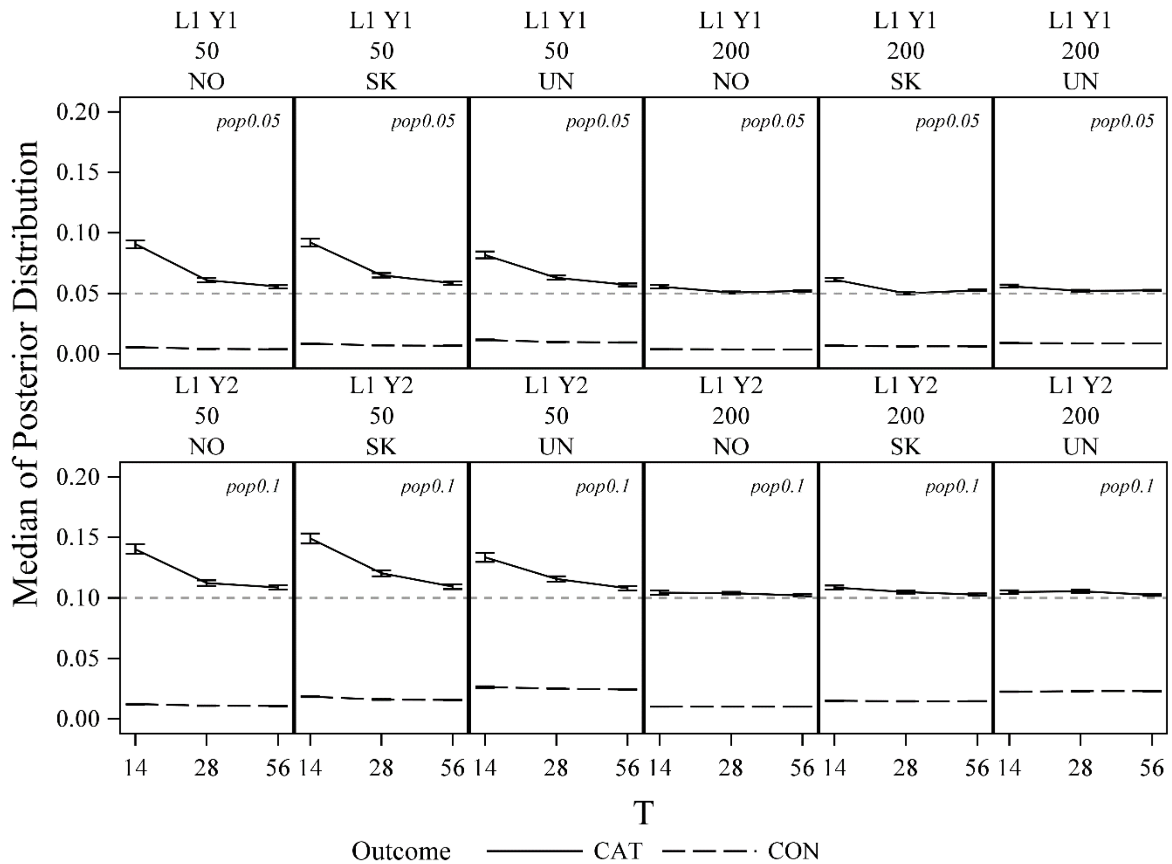

Figure 43. Average Median Value of the Posterior Distributions for Level-1 Covariate Random Effect Variances across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

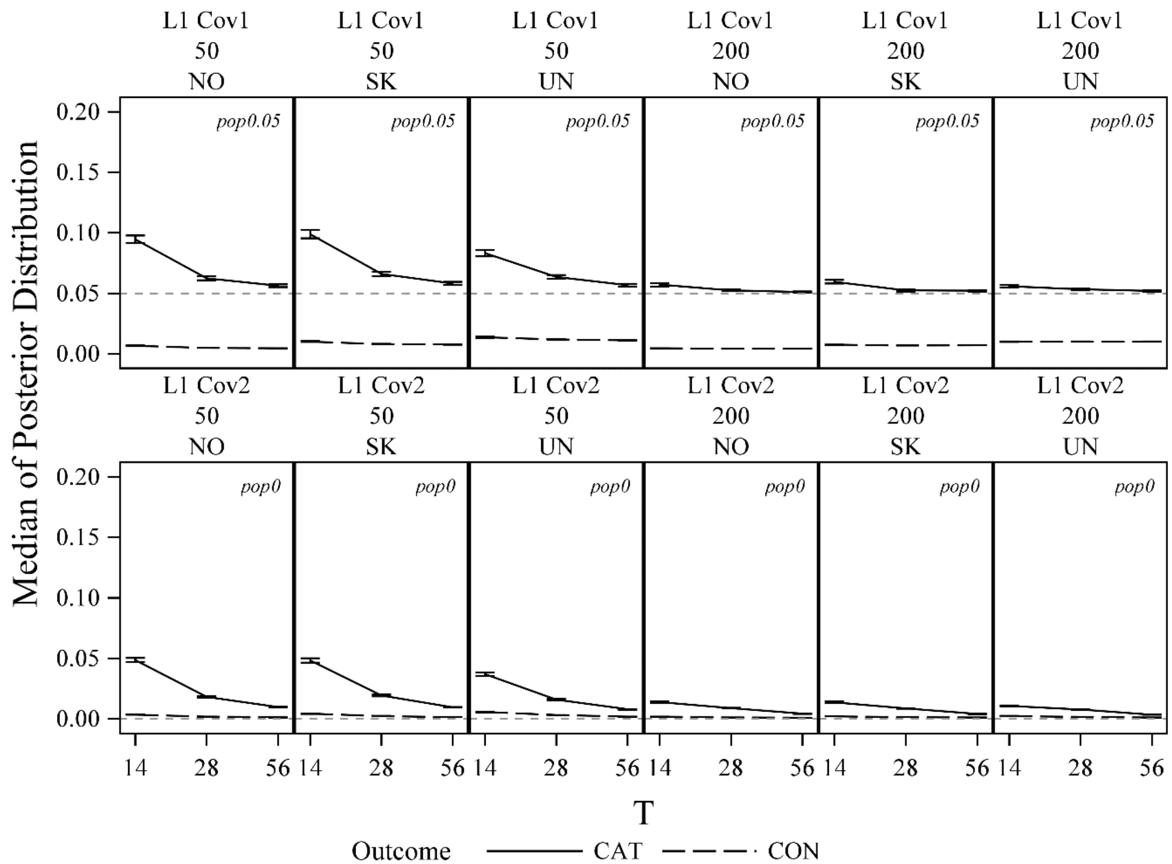

Figure 44. Efficacy Results for Actor and Partner Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

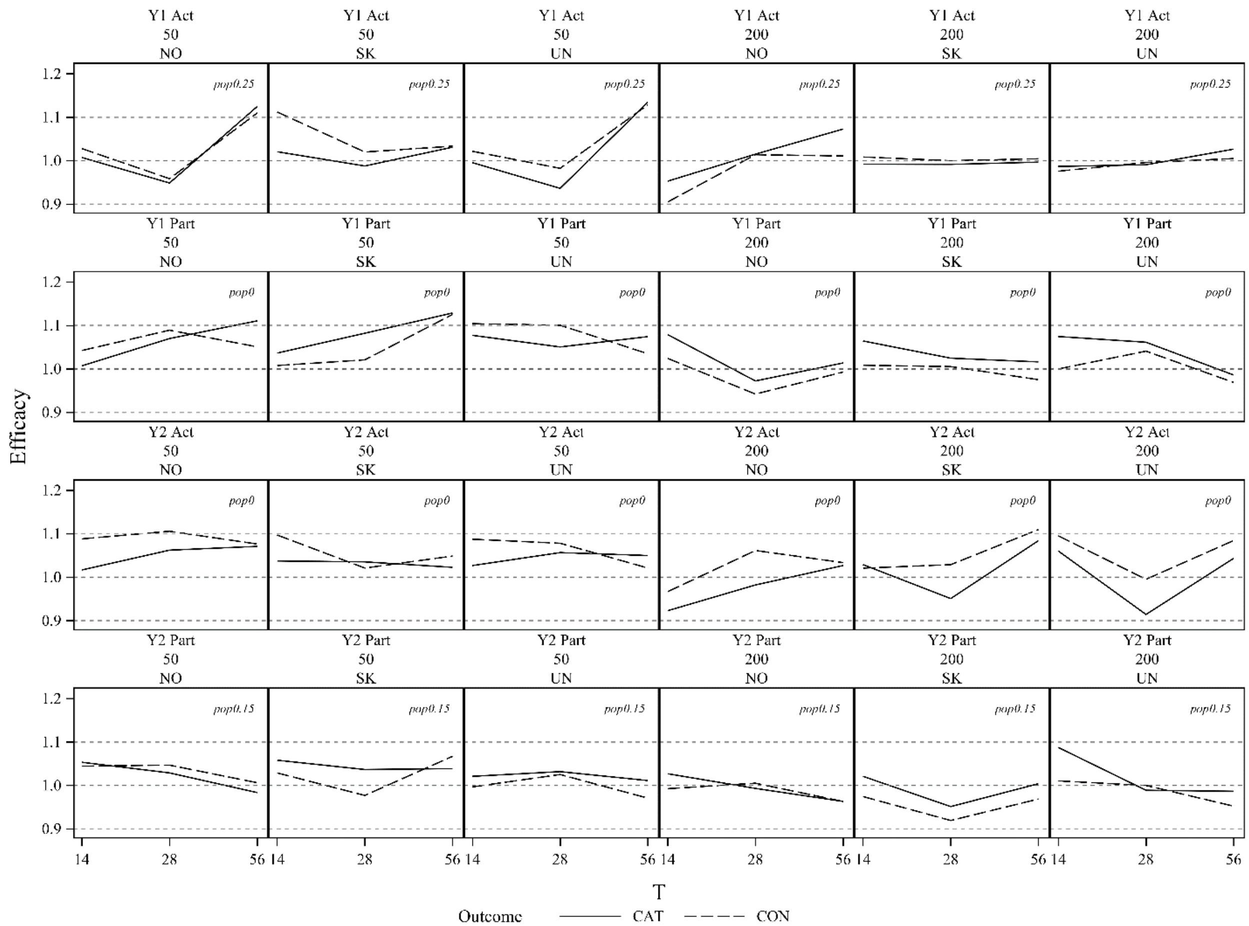

Figure 45. Efficacy Results for Autoregressive Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

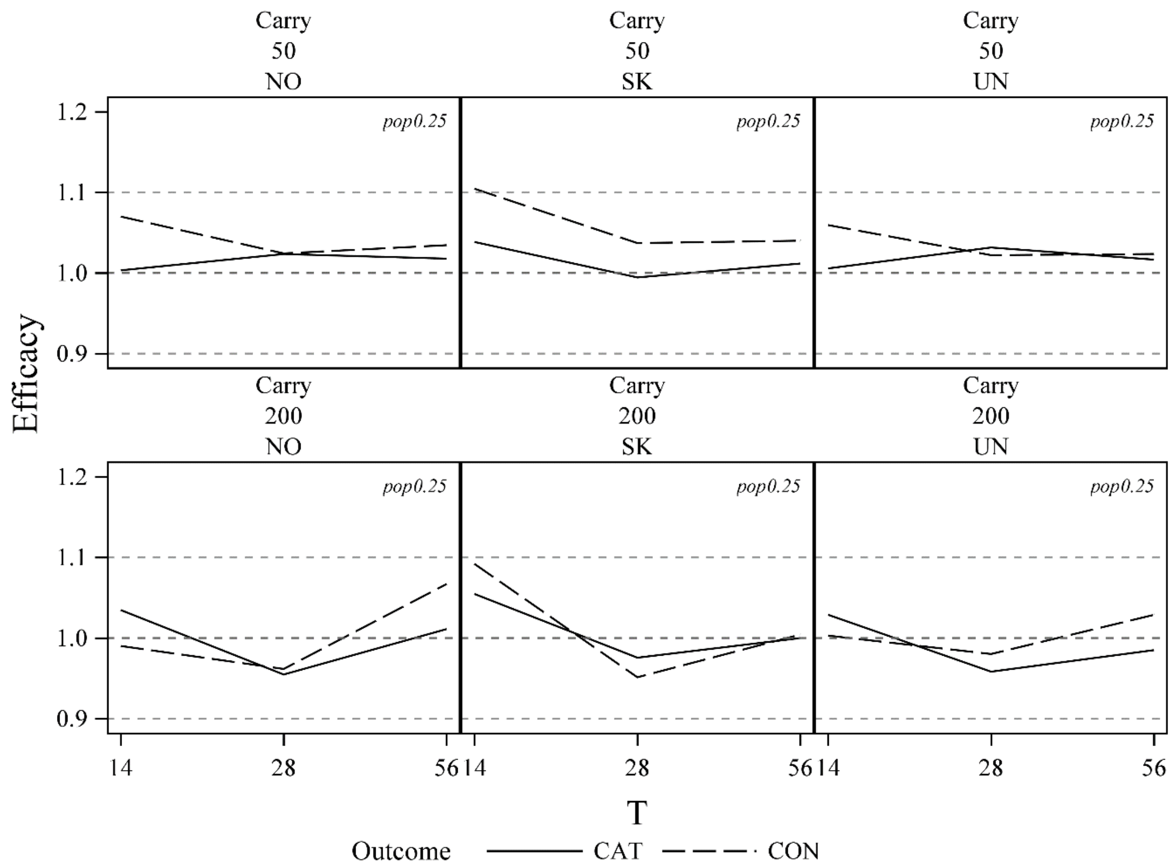

Figure 46. Efficacy Results for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

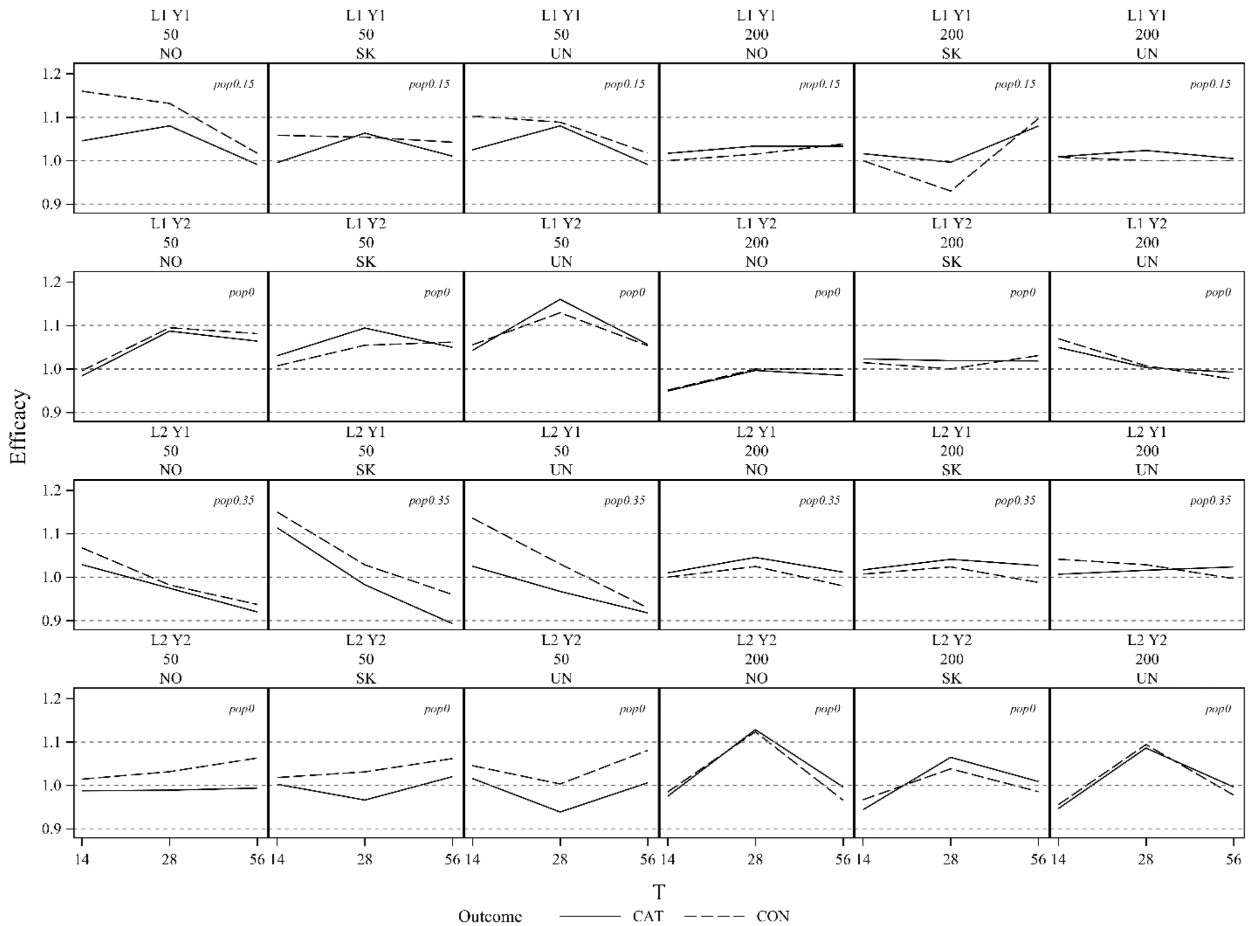

Figure 47. Efficacy Results for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

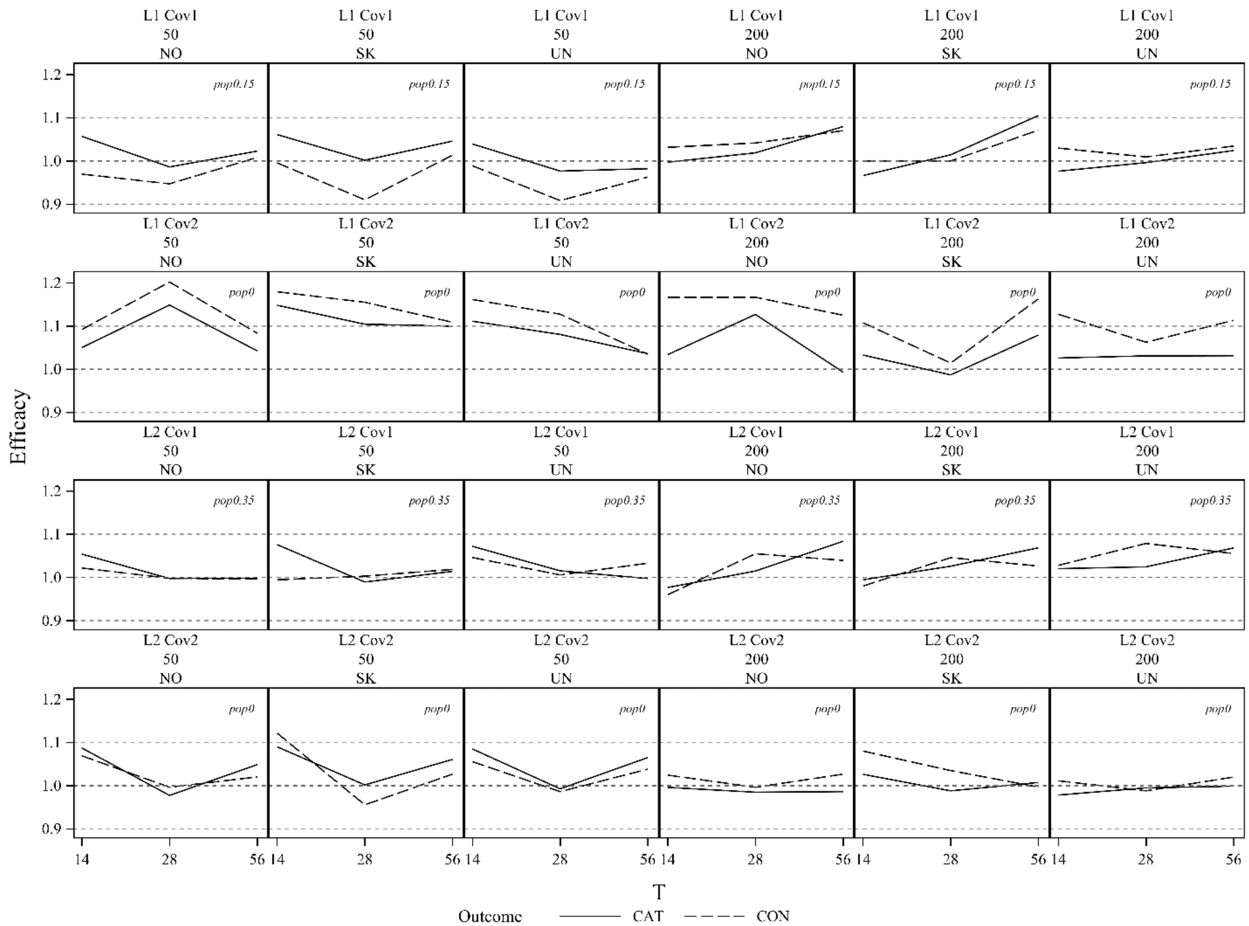

Figure 48. Efficacy Results for Actor and Partner Random Effect Variances across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

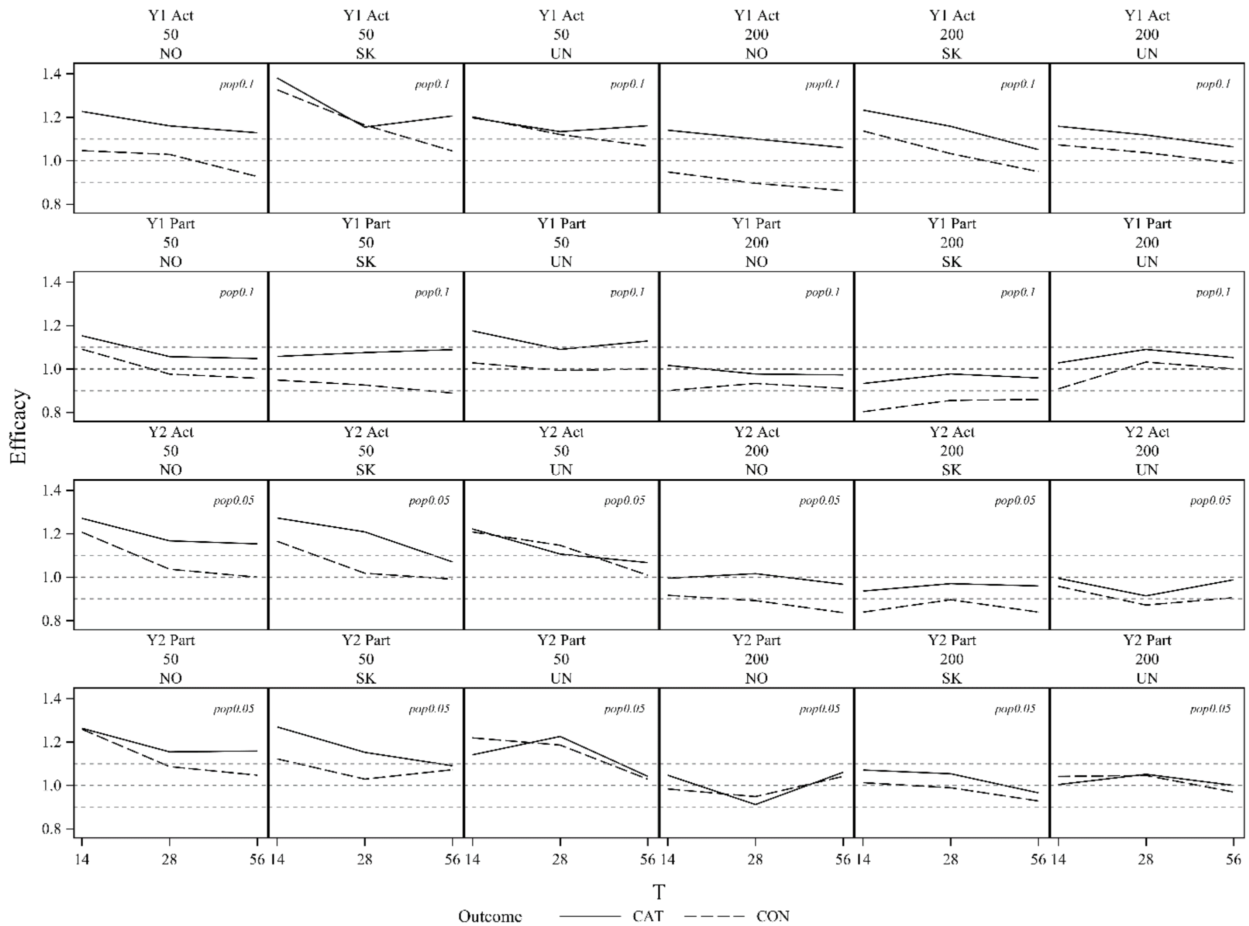

Figure 49. Efficacy Results for Autoregressive Random Effect Variances across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

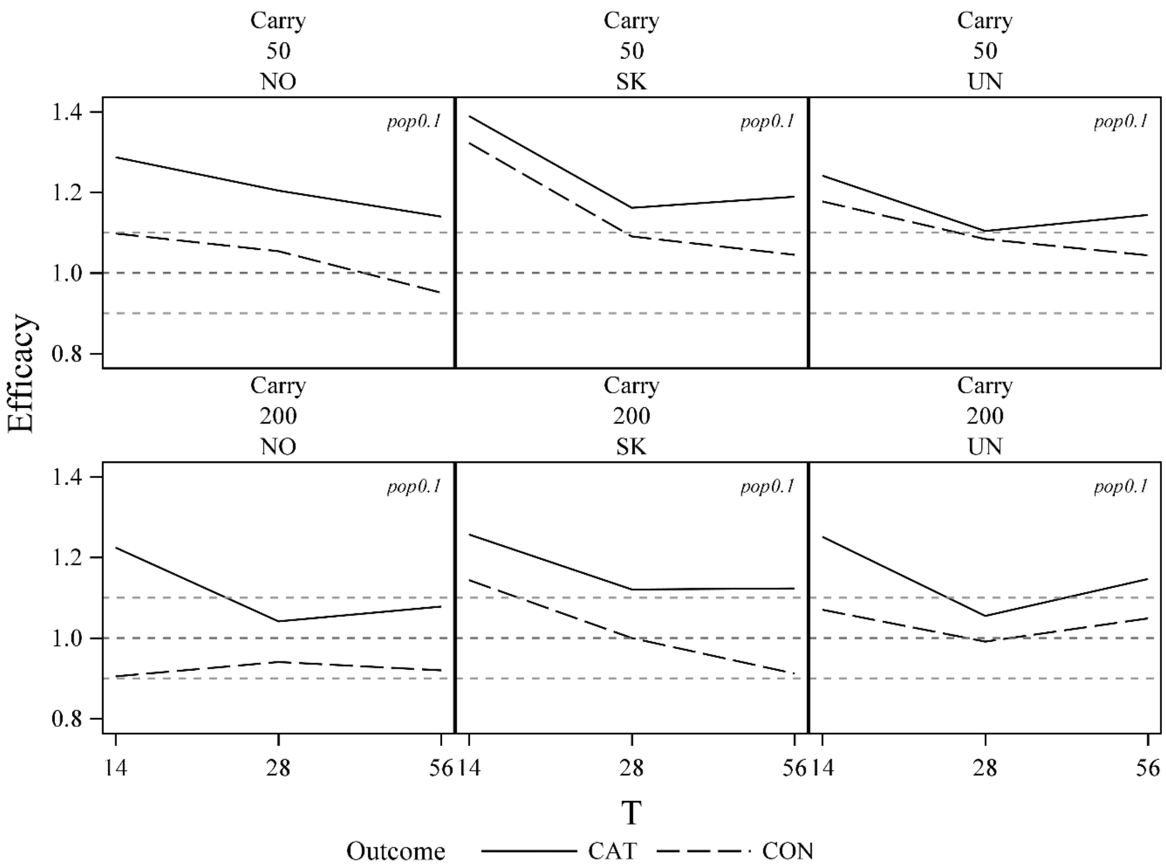

Figure 50. Efficacy Results for Level-1 Covariate Random Effect Variances across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

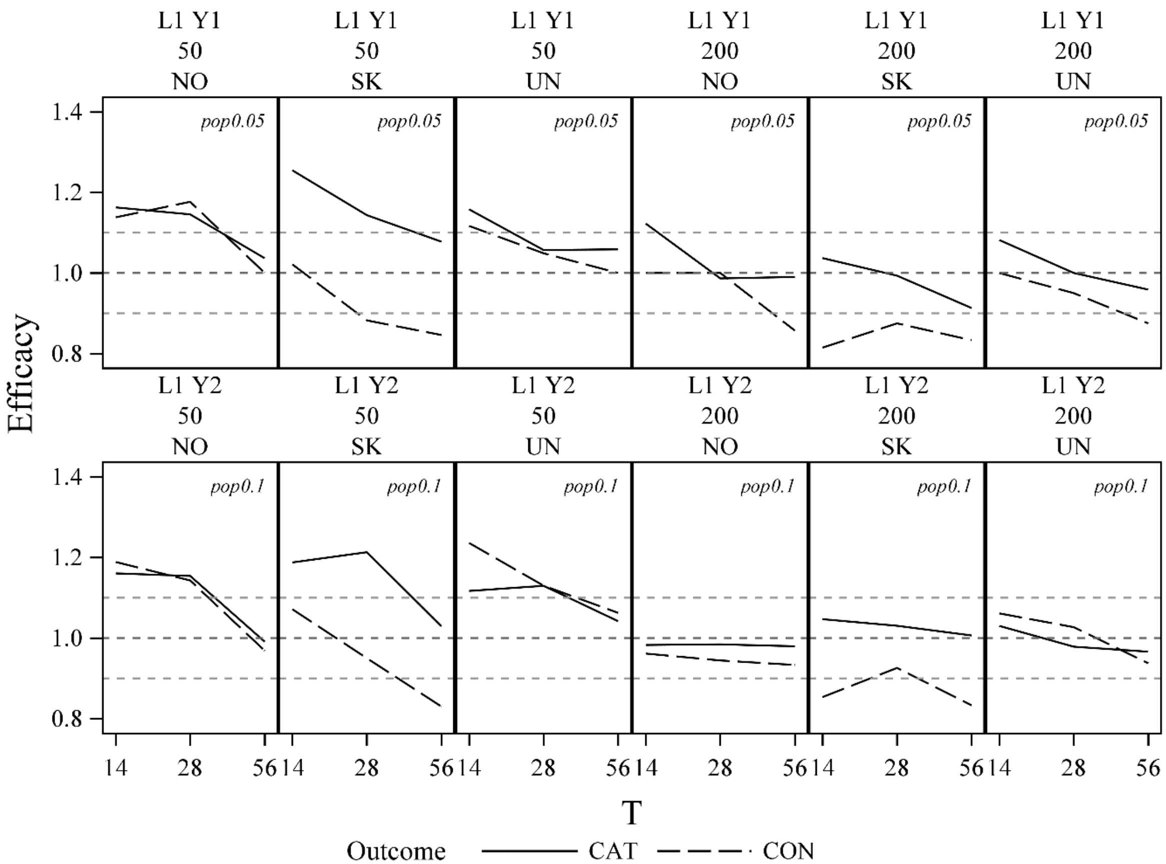

Figure 51. Efficacy Results for Level-1 Covariate Random Effect Variances across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

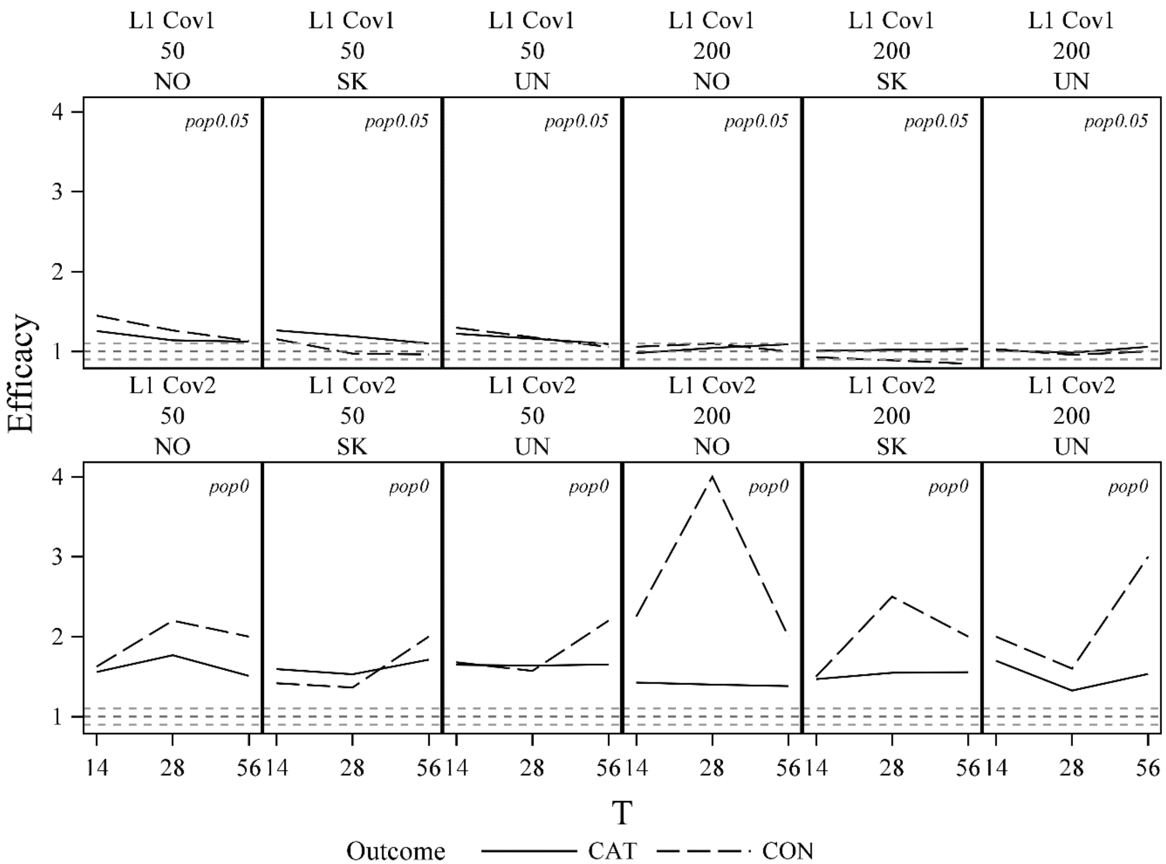

Figure 52. Power to Detect Significant Actor and Partner Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

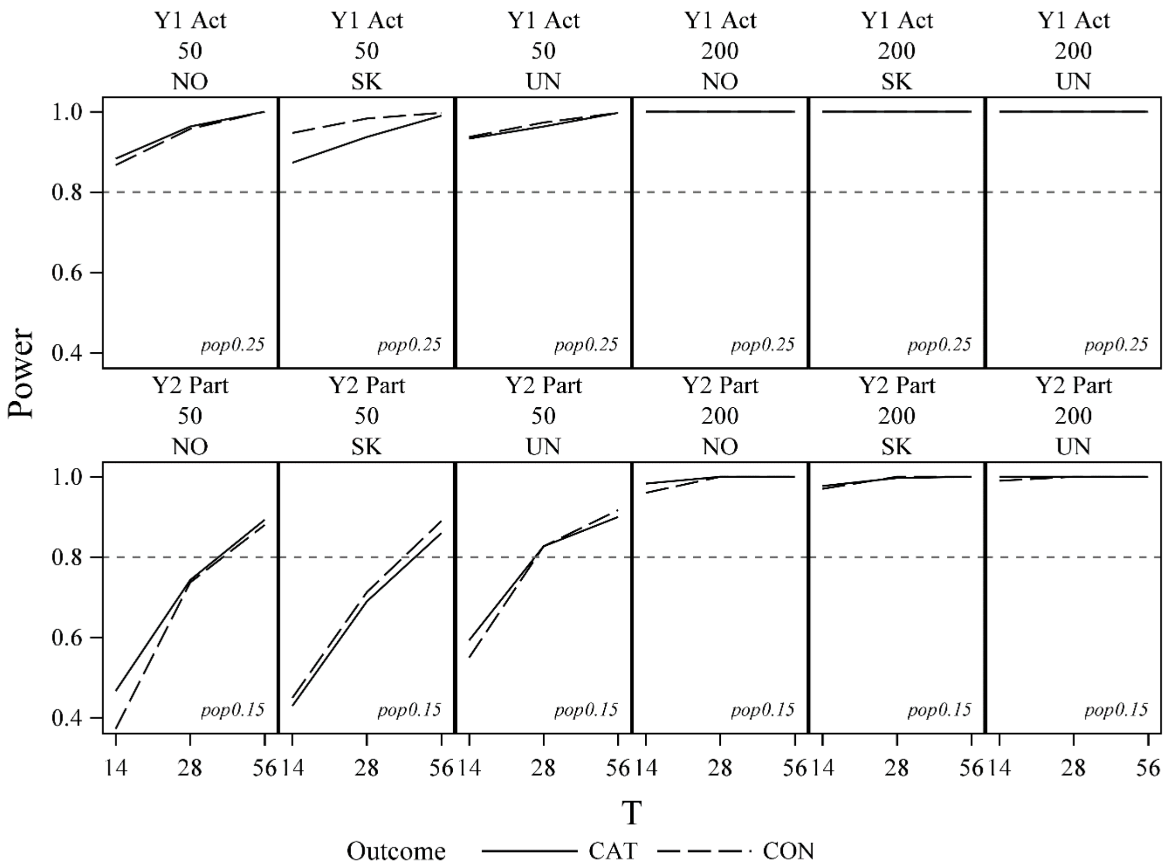

Figure 53. Power to Detect Autoregressive Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

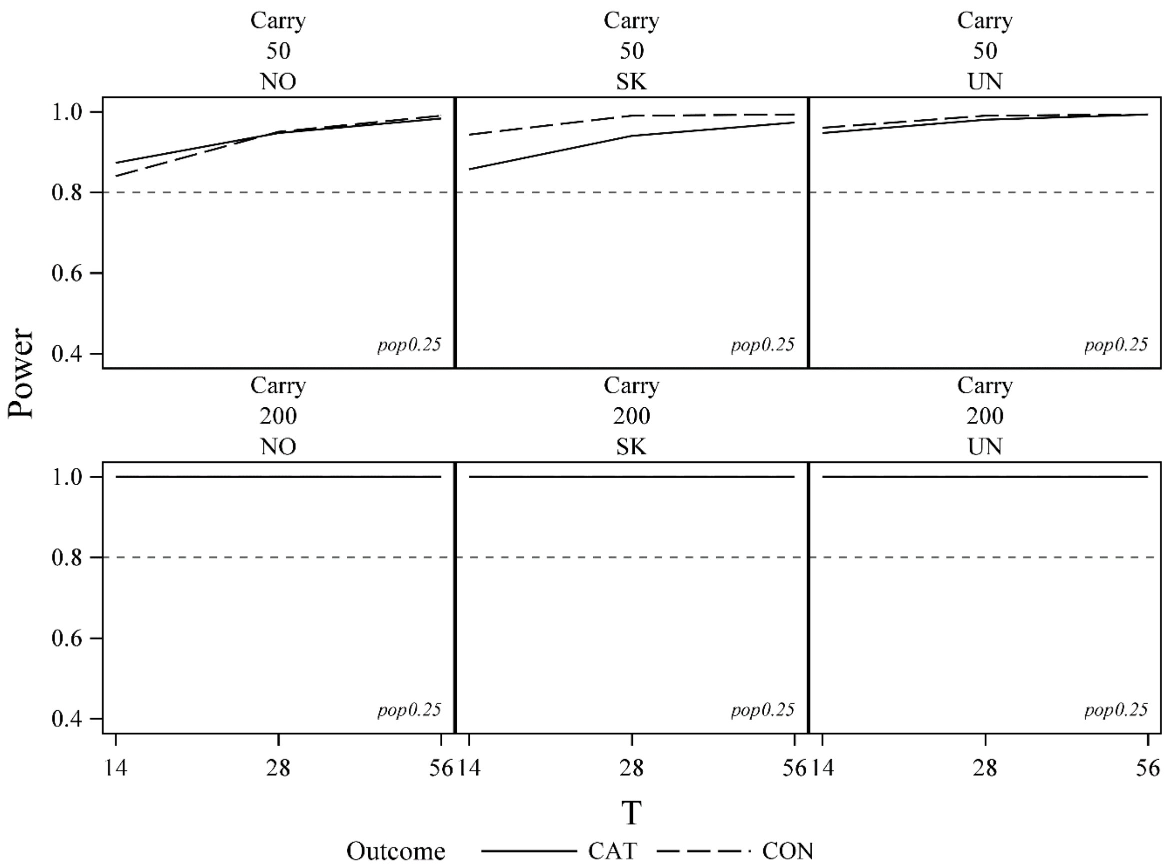

Figure 54. Power to Detect Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

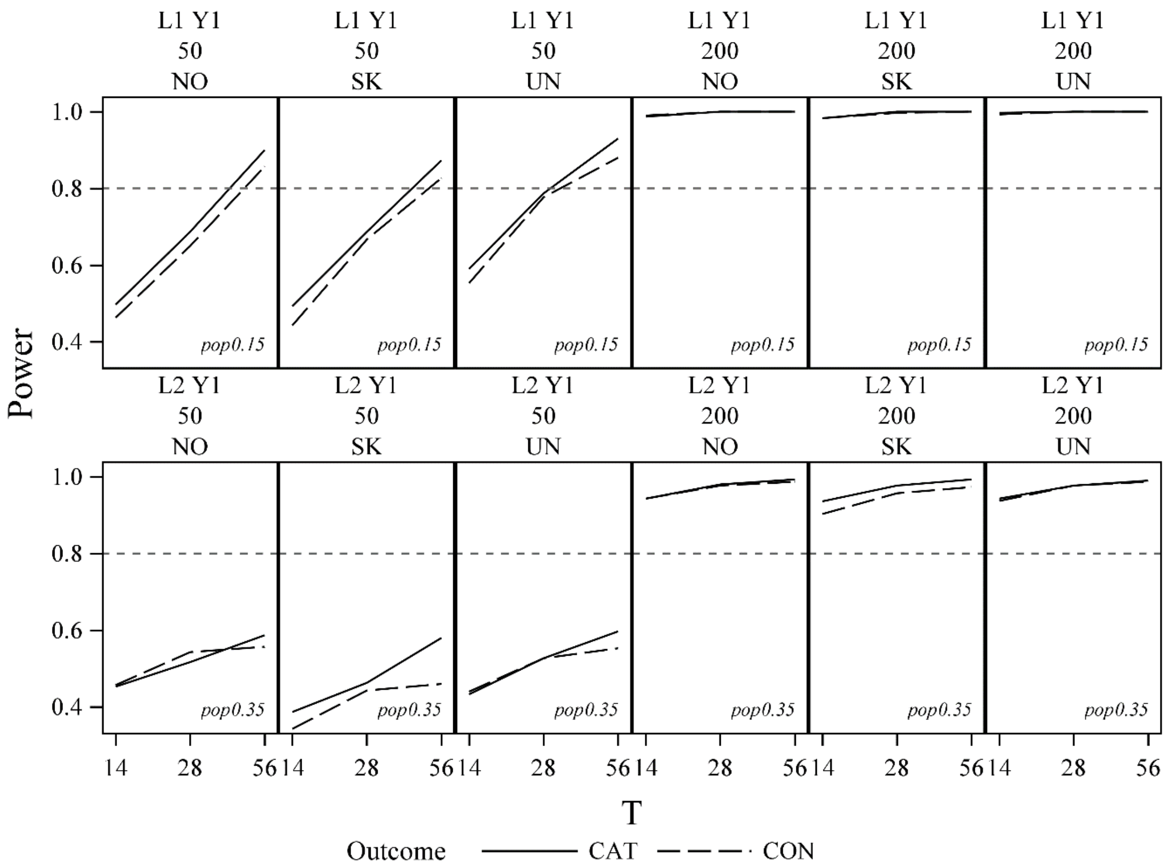

Figure 55. Power to Detect Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

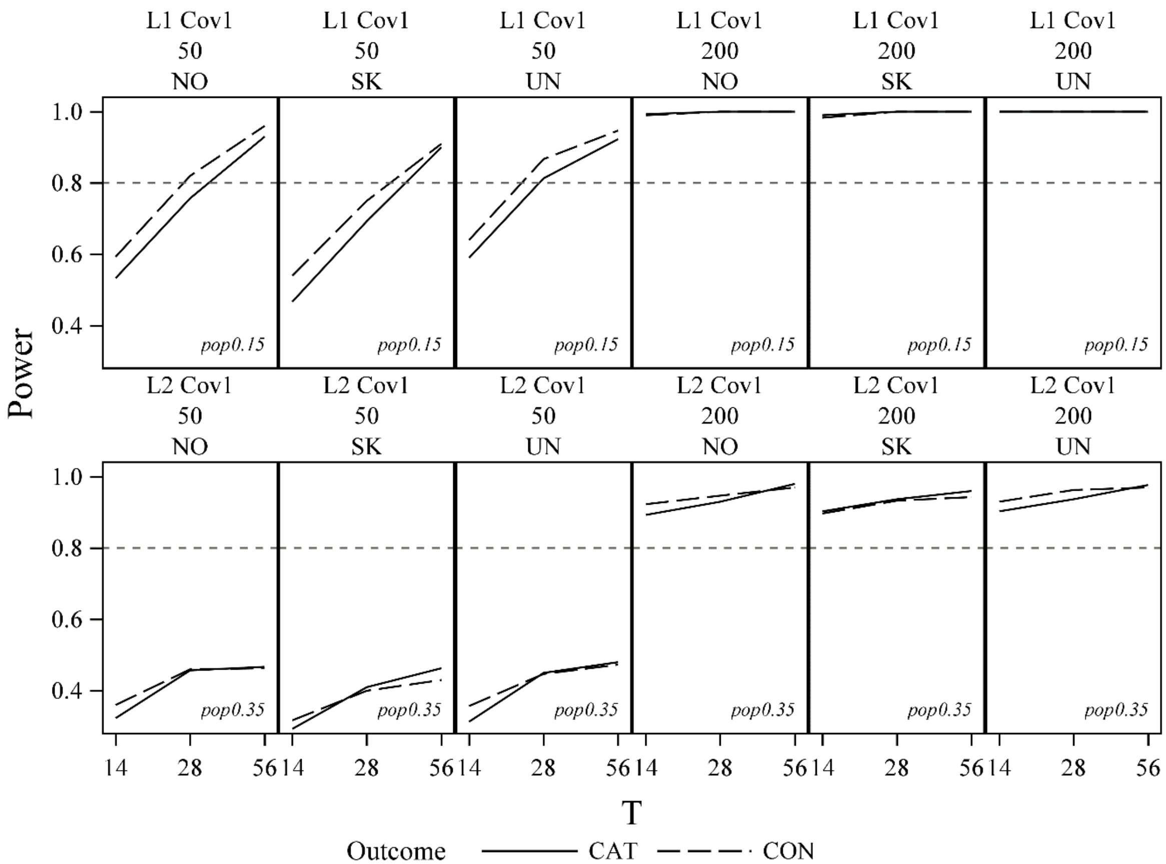

Figure 56. Type-I Error Rates for Actor and Partner Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

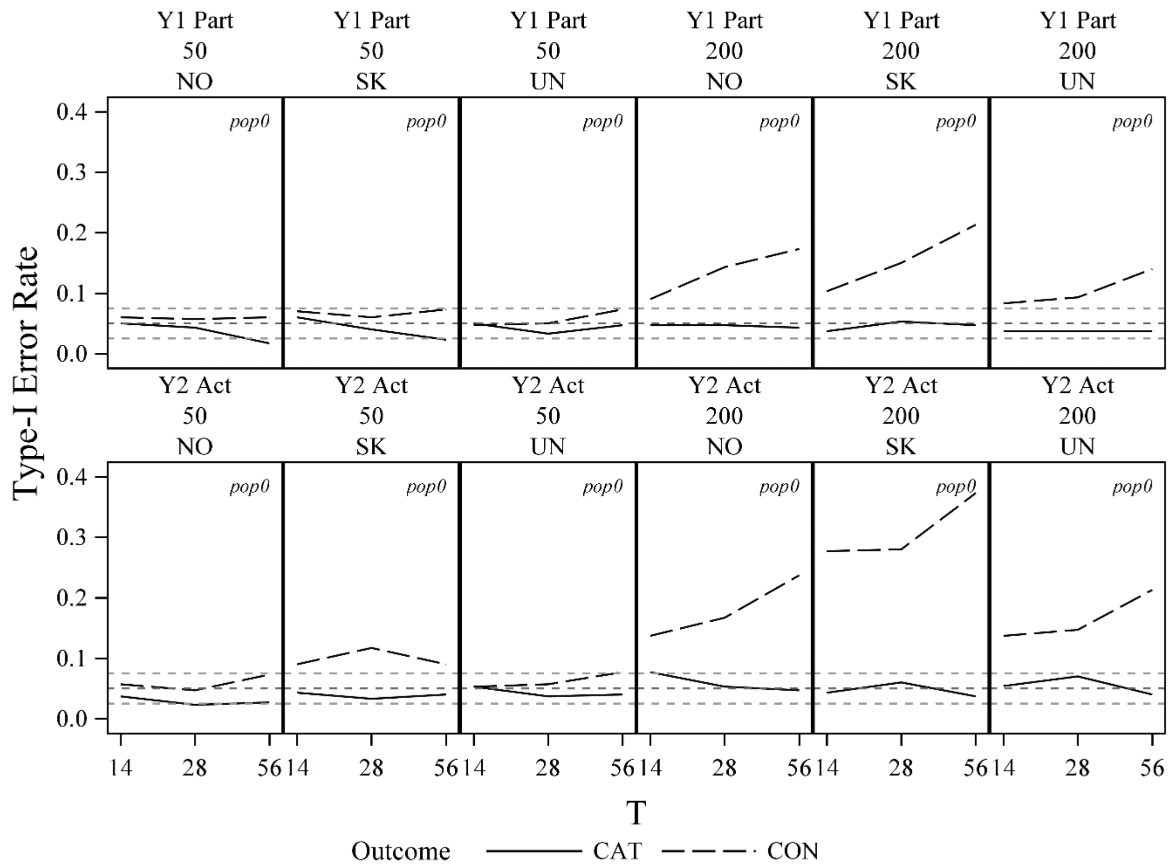

Figure 57. Estimated Type-I Error Rates for Autoregressive Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

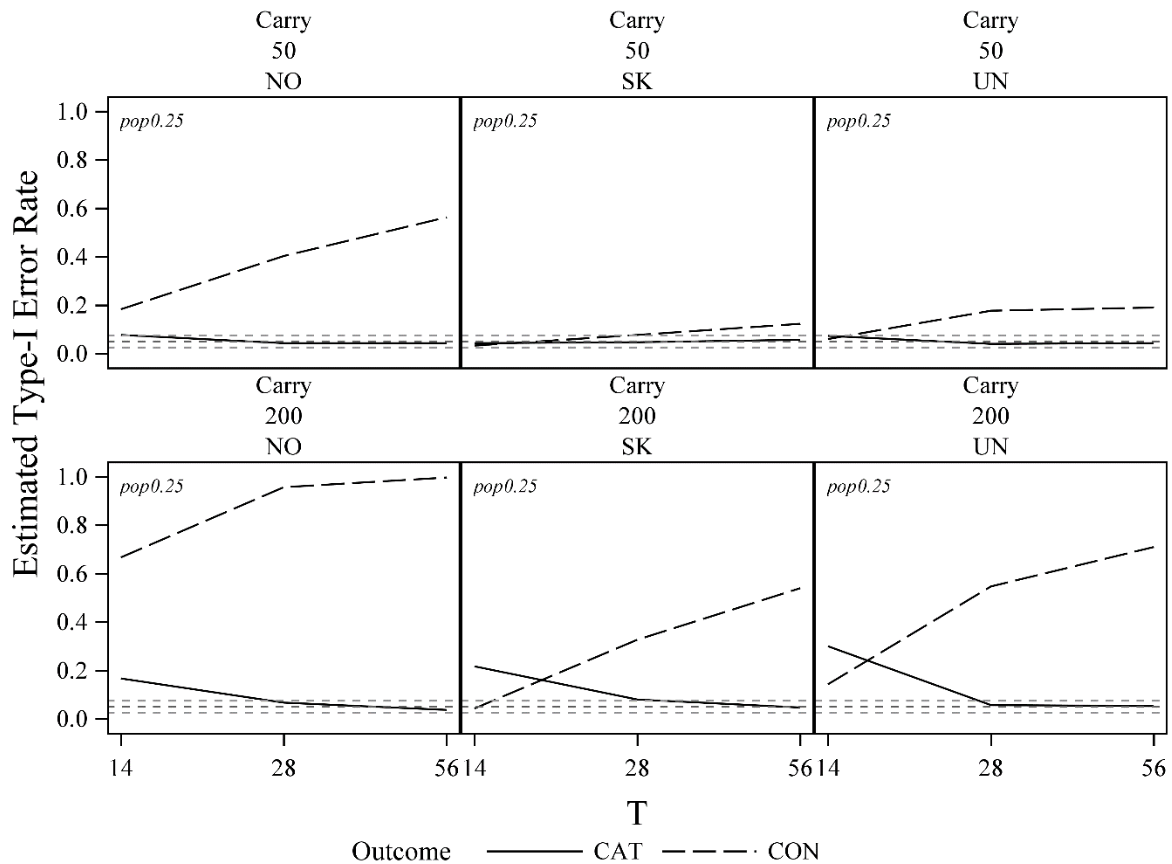

Figure 58. Type-I Error Rates for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

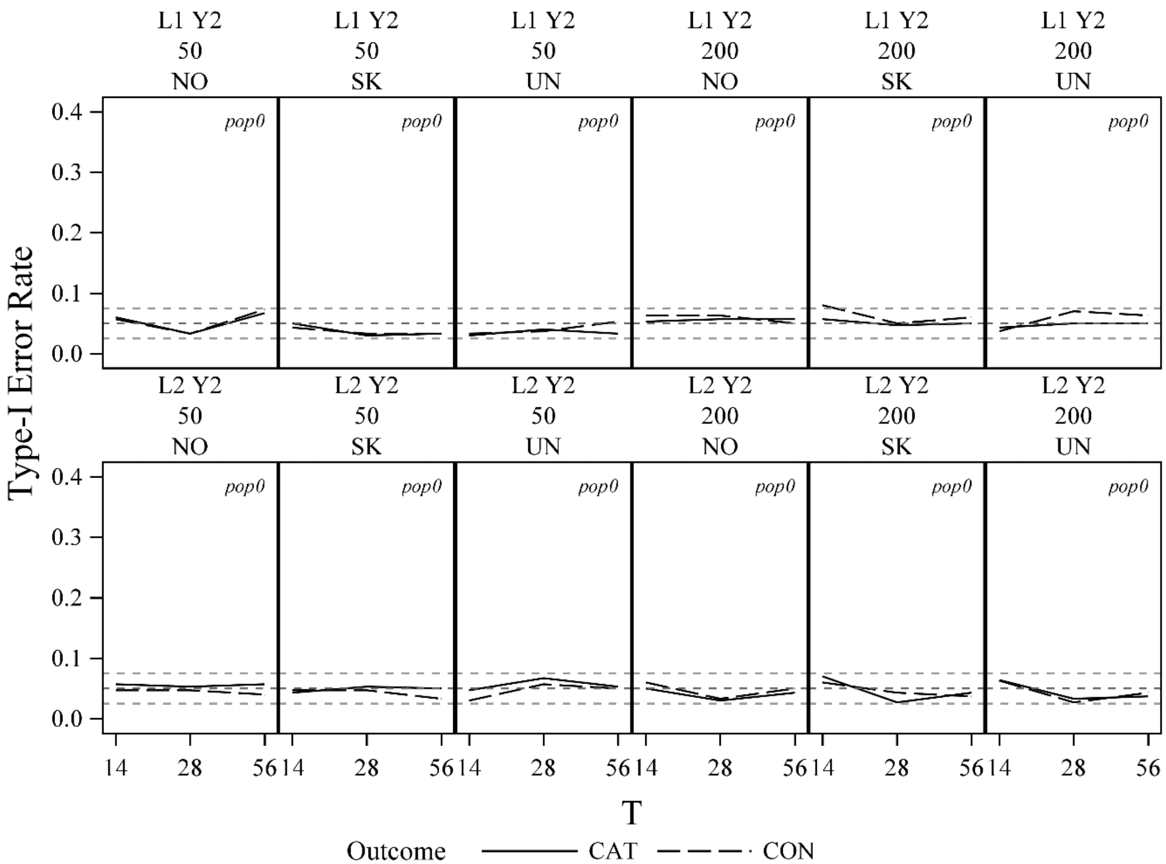

Figure 59. Type-I Error Rates for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 3-Category Outcomes

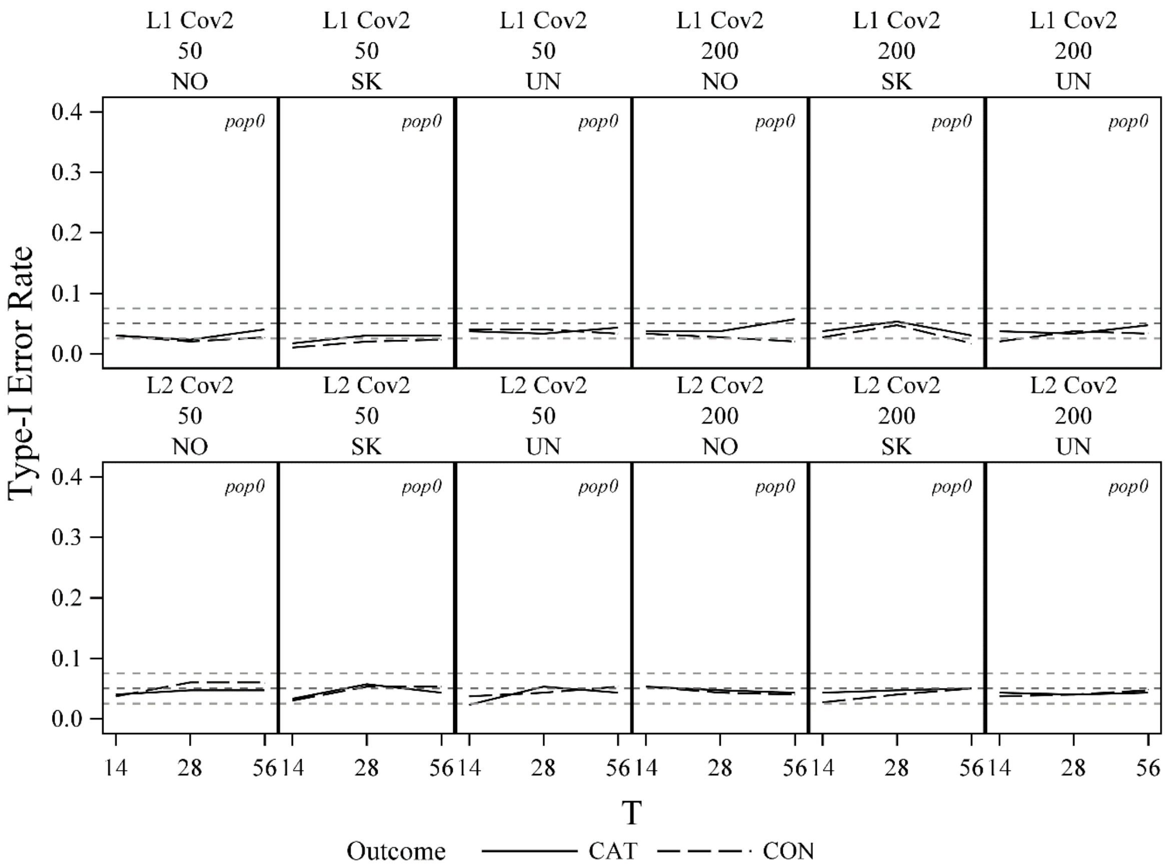

Figure 60. Average Bias for Actor and Partner Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

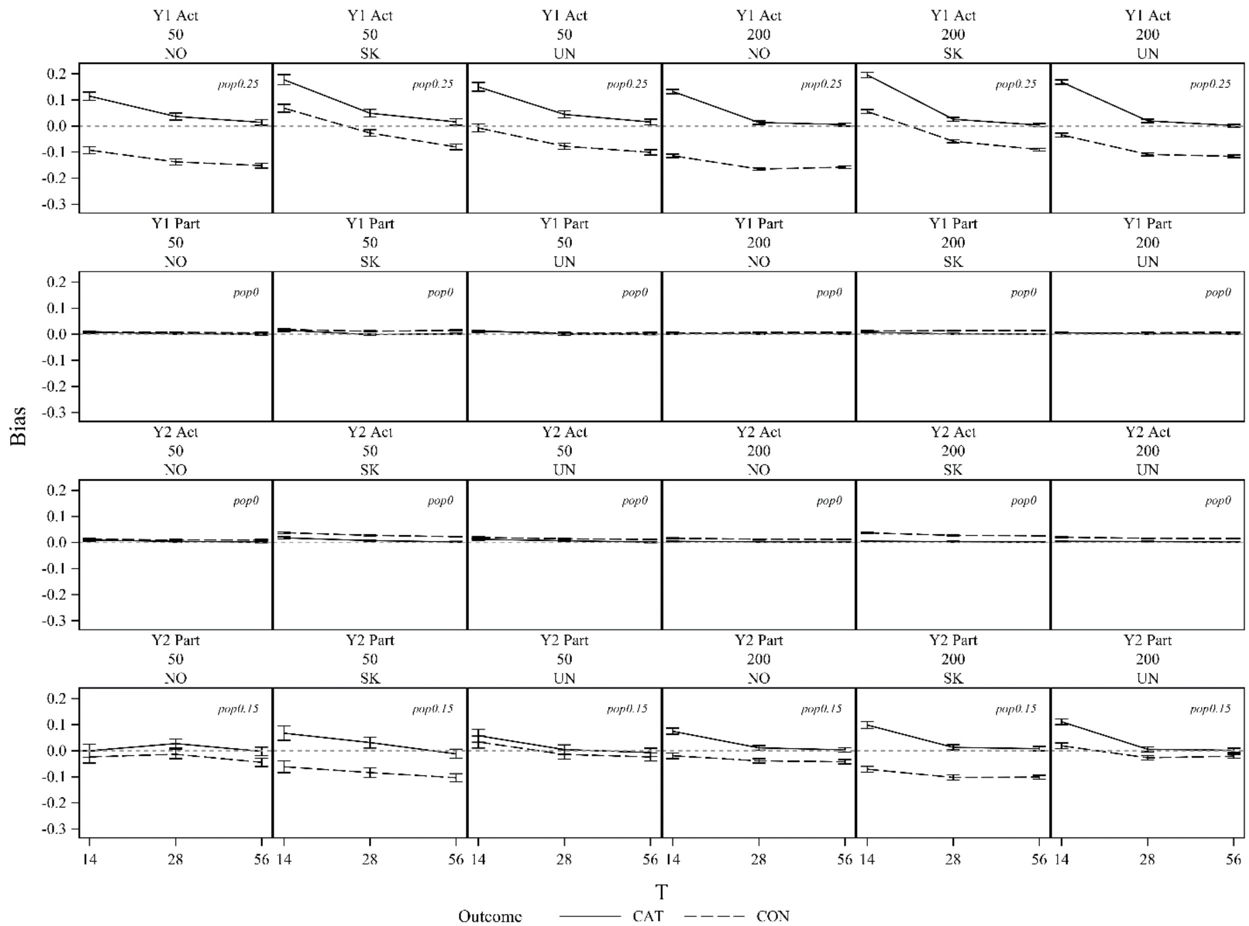

Figure 61. Average Bias for Autoregressive Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

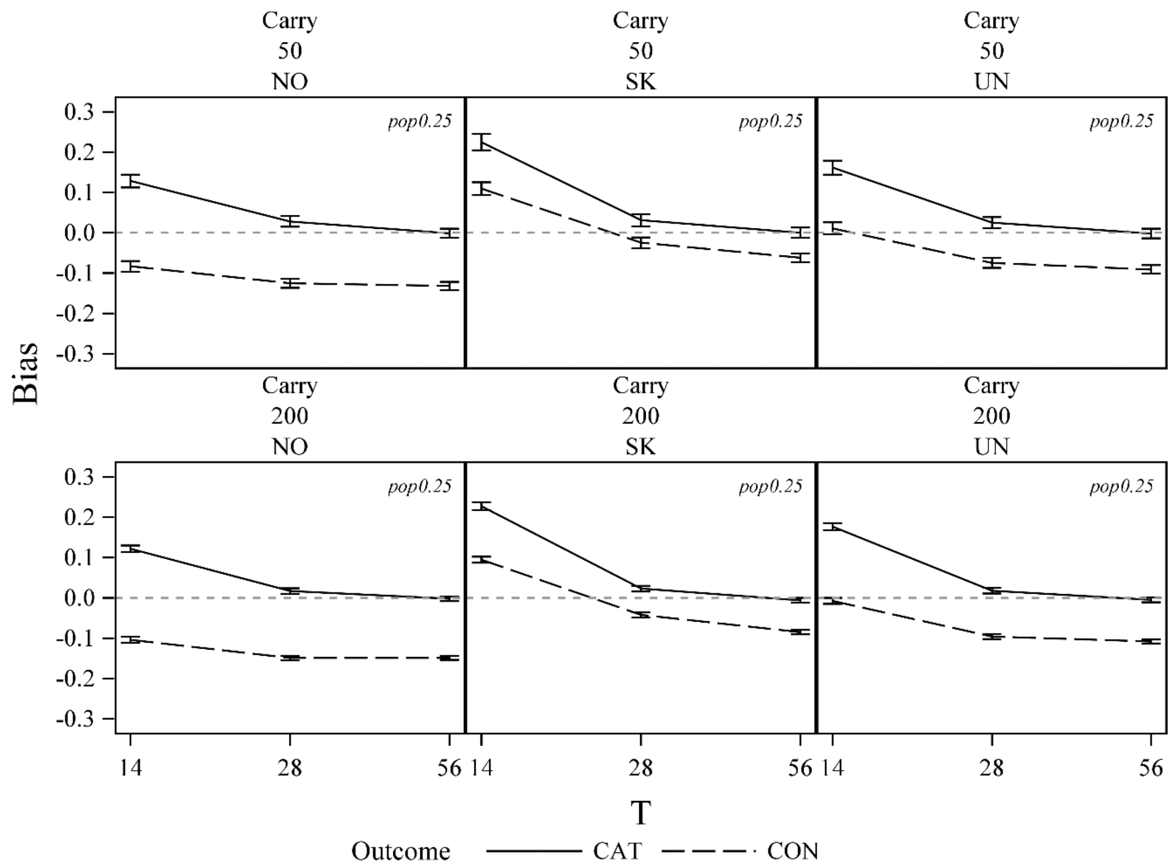

Figure 62. Average Bias for Standardized Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

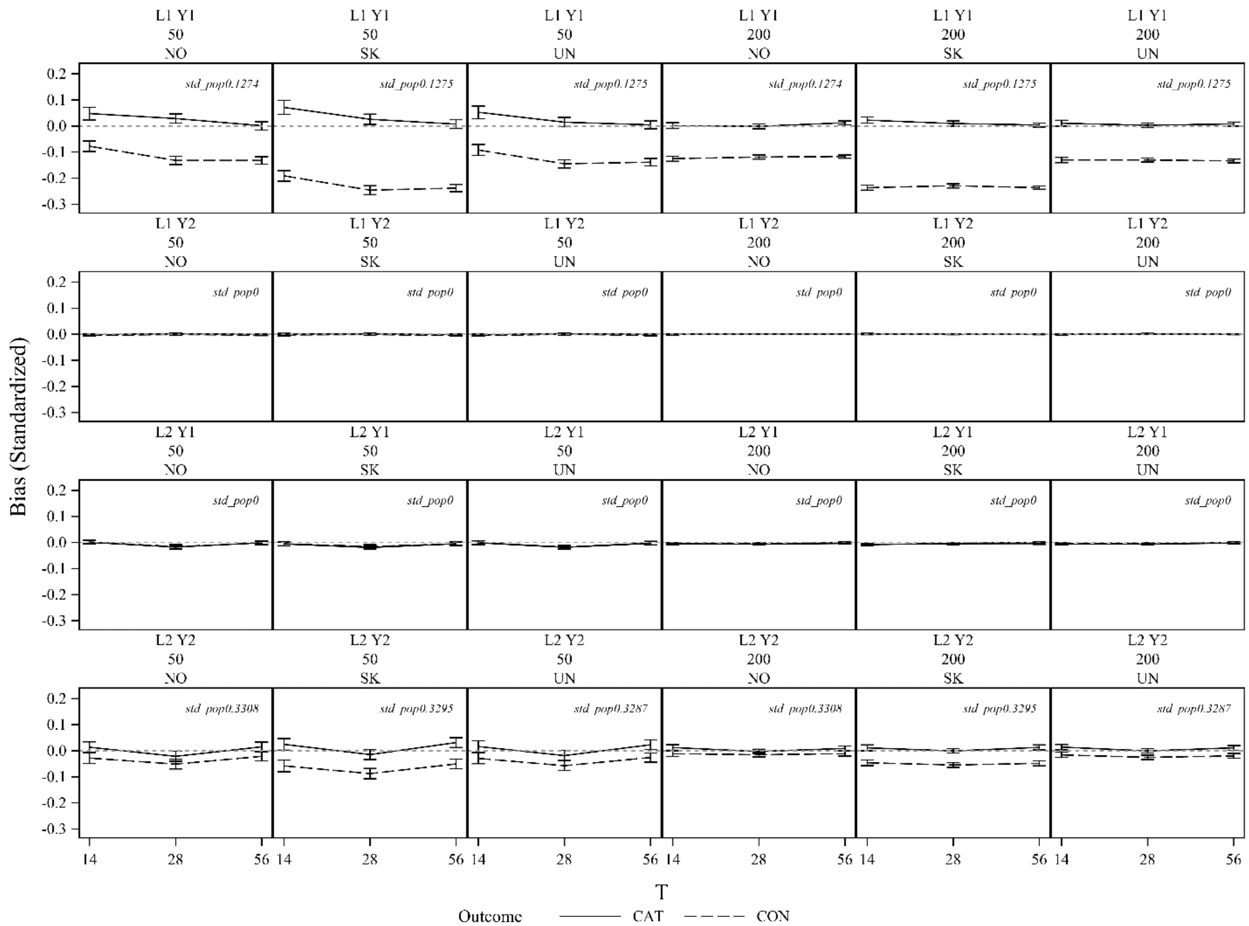

Figure 63. Average Bias for Standardized Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

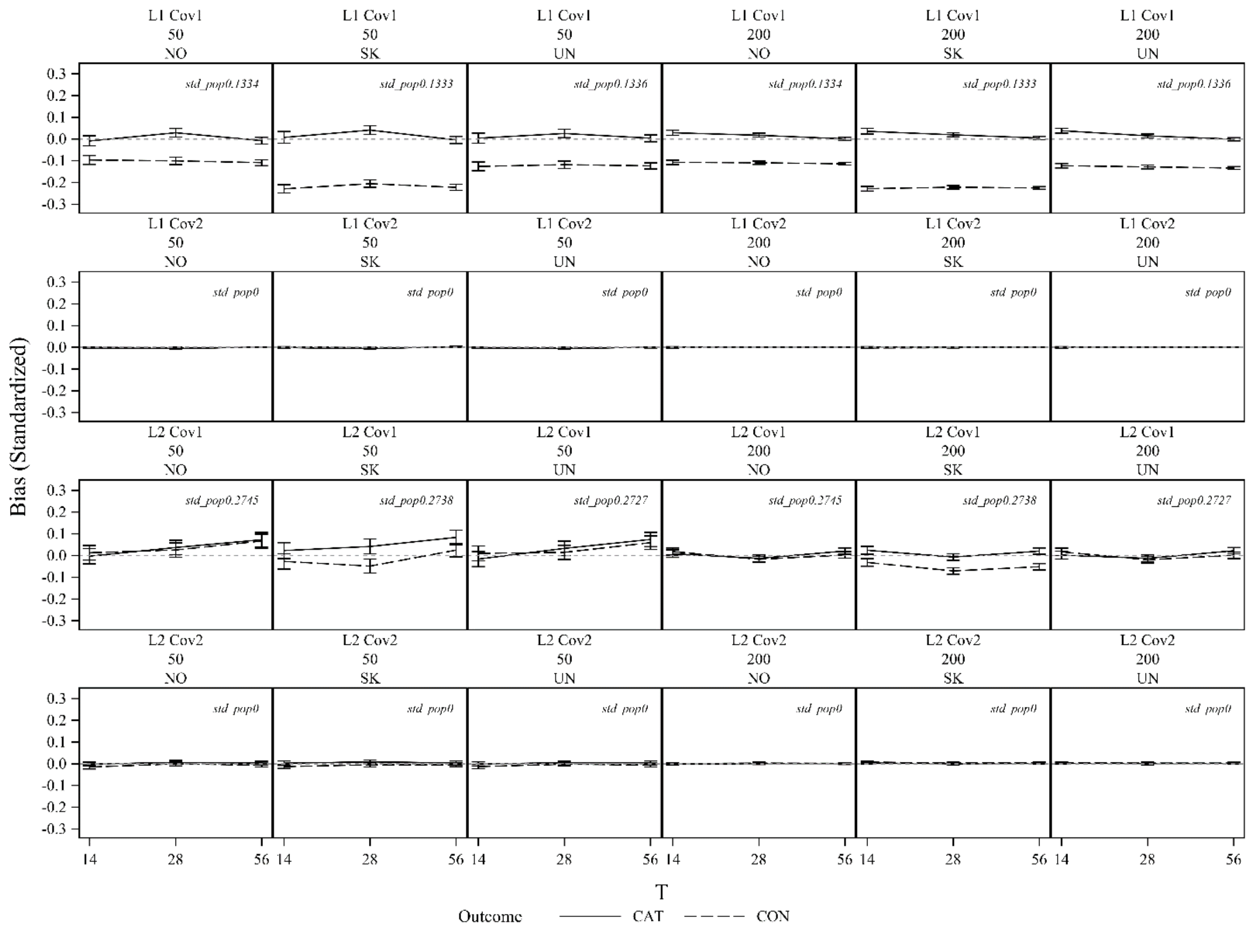

Figure 64. Average Median Value of the Posterior Distributions for Actor and Partner Random Effect Variances across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

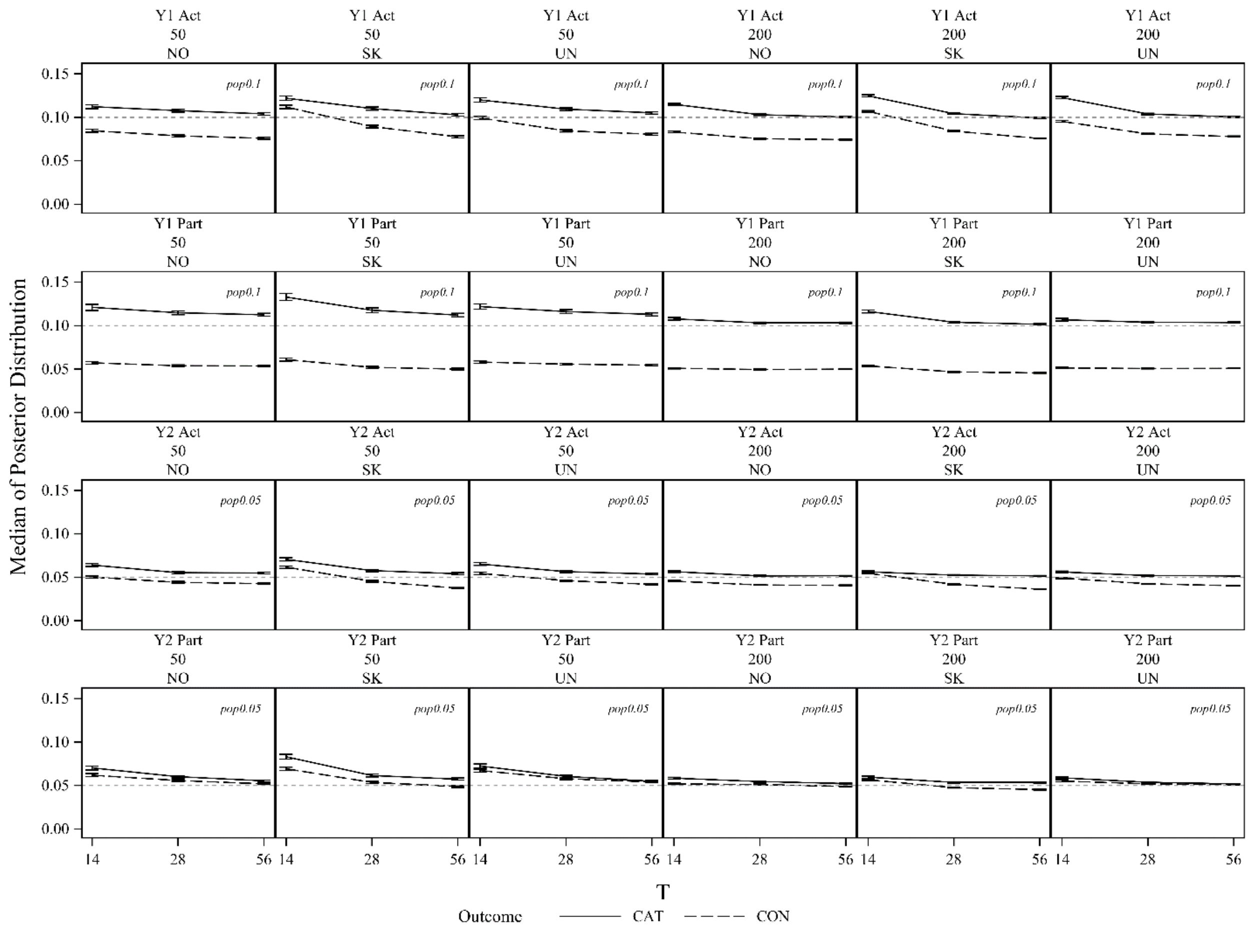

Figure 65. Average Median Value of the Posterior Distributions for Autoregressive Random Effect Variances across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

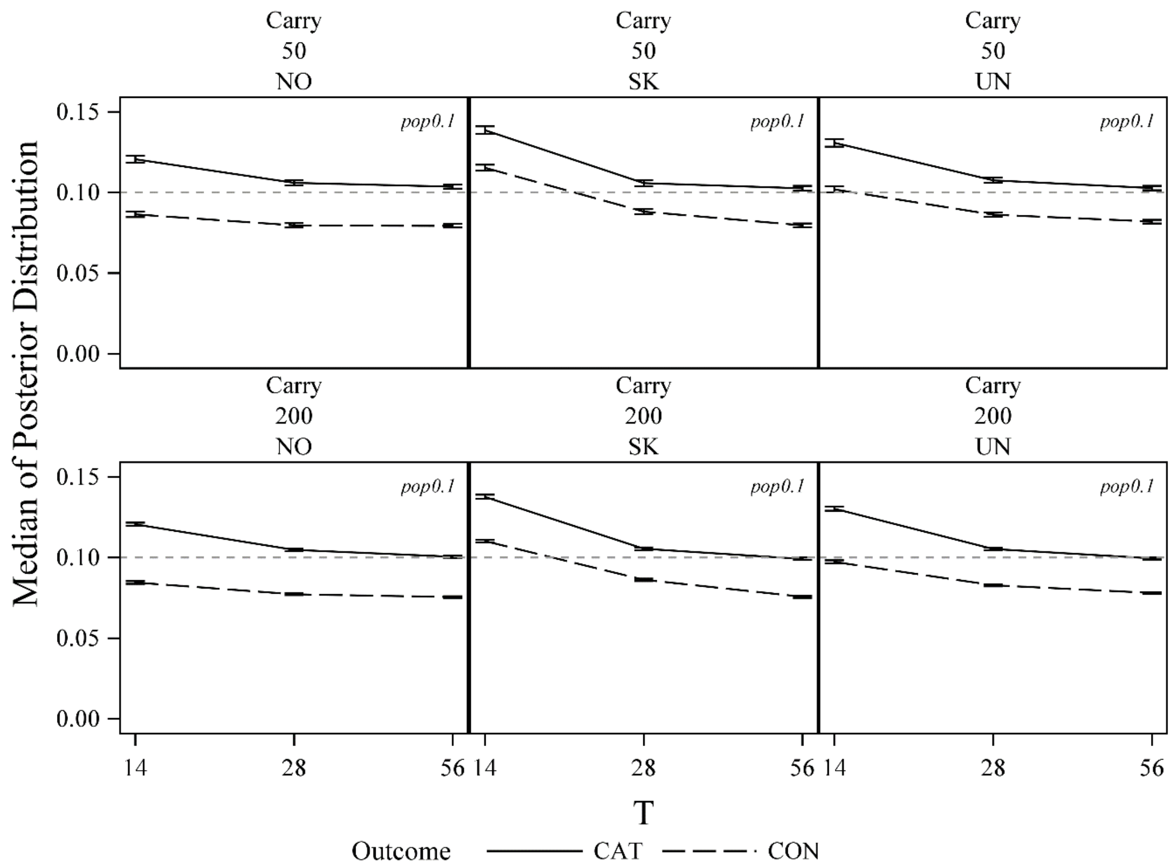

Figure 66. Average Median Value of the Posterior Distributions for Level-1 Covariate Random Effect Variances across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

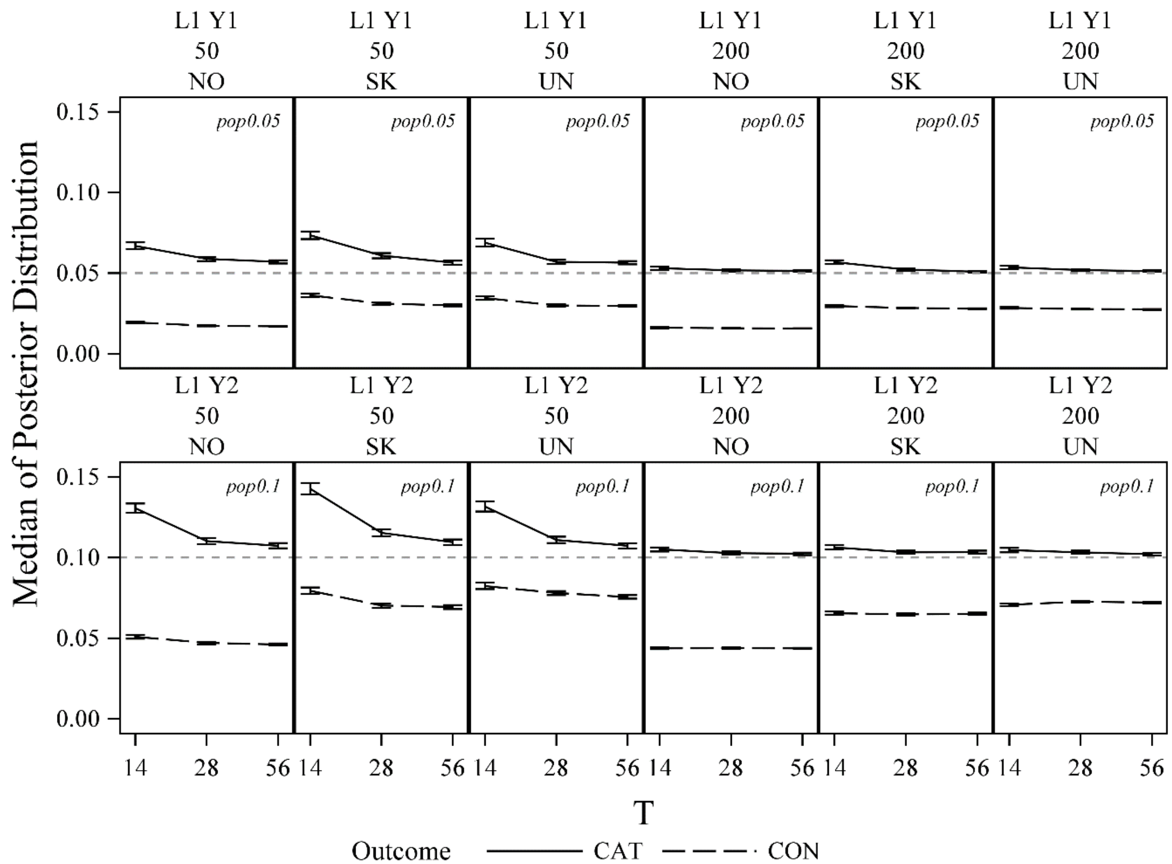

Figure 67. Average Median Value of the Posterior Distributions for Level-1 Covariate Random Effect Variances across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

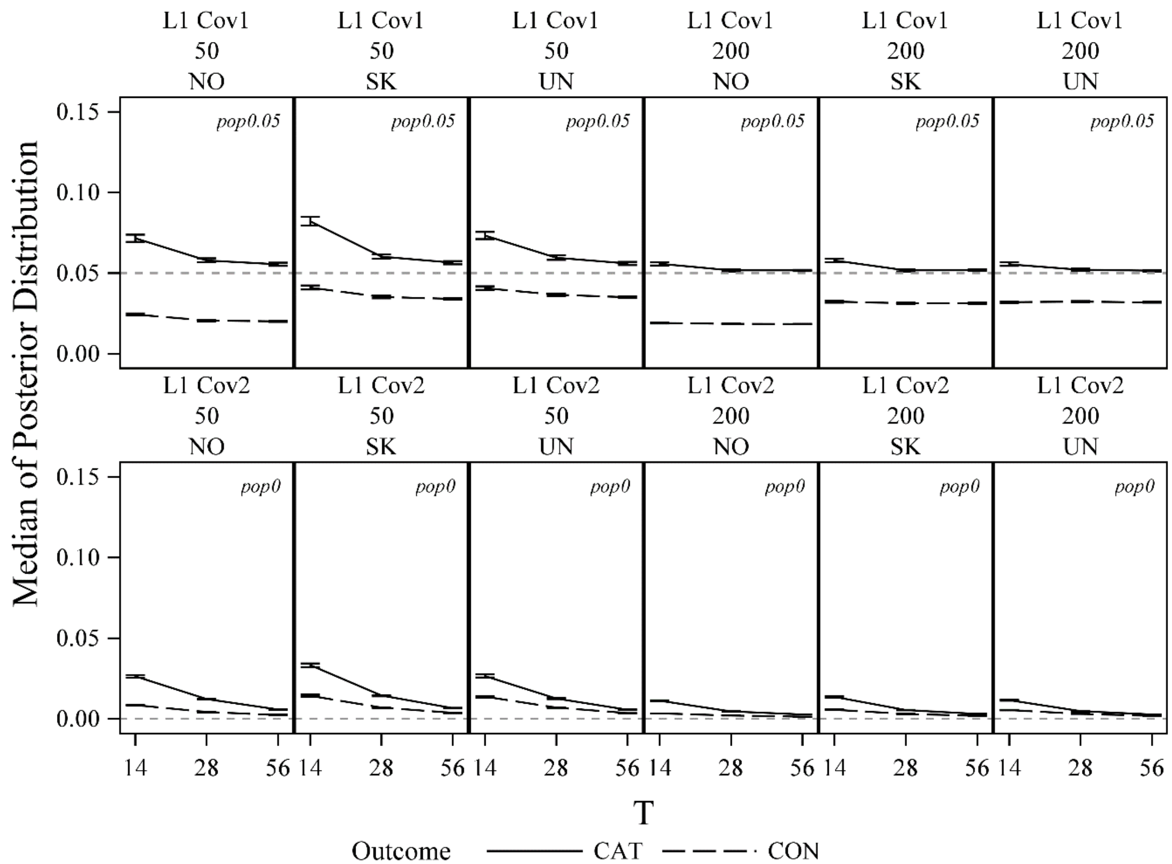

Figure 68. Efficacy Results for Actor and Partner Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

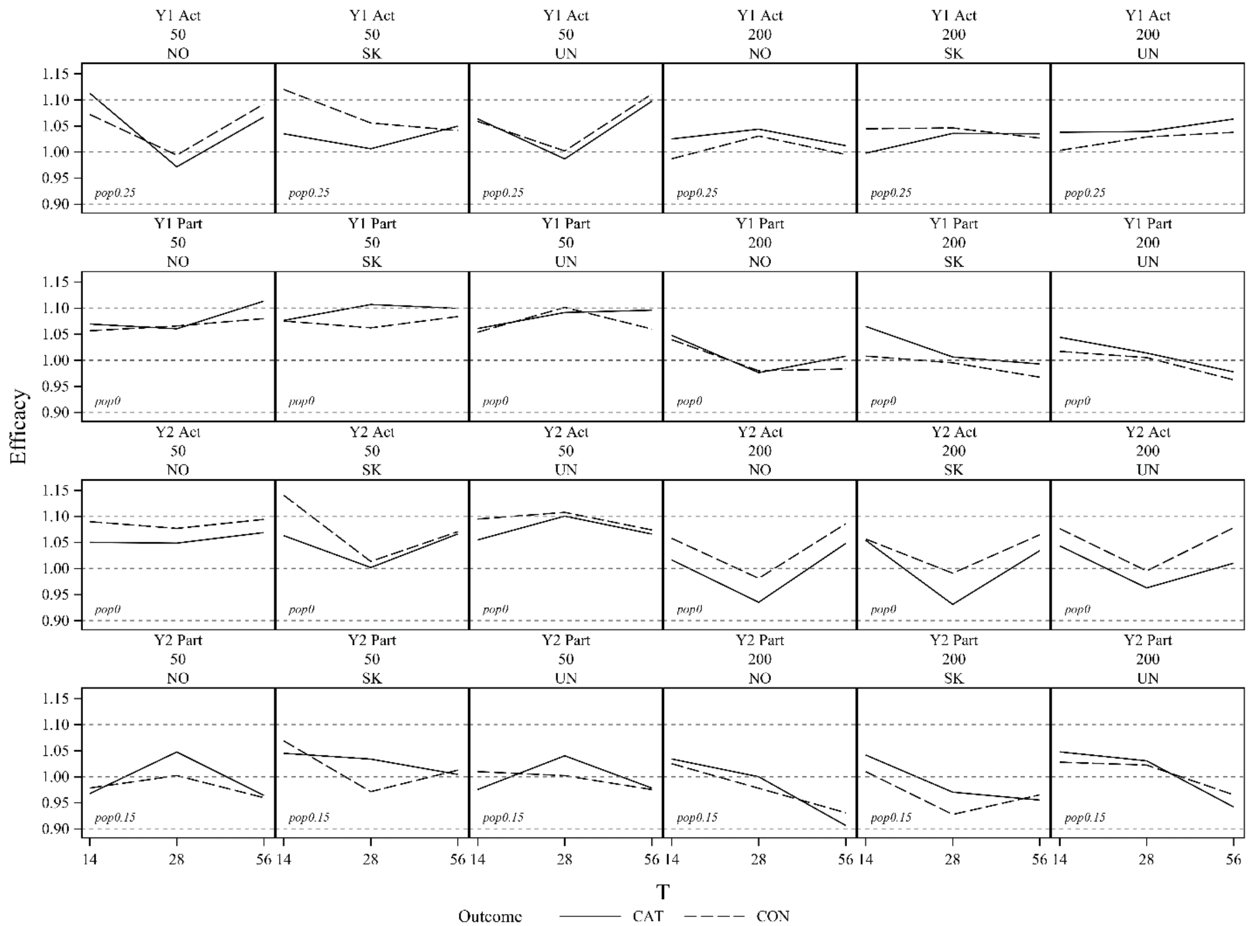

Figure 69. Efficacy Results for Autoregressive Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

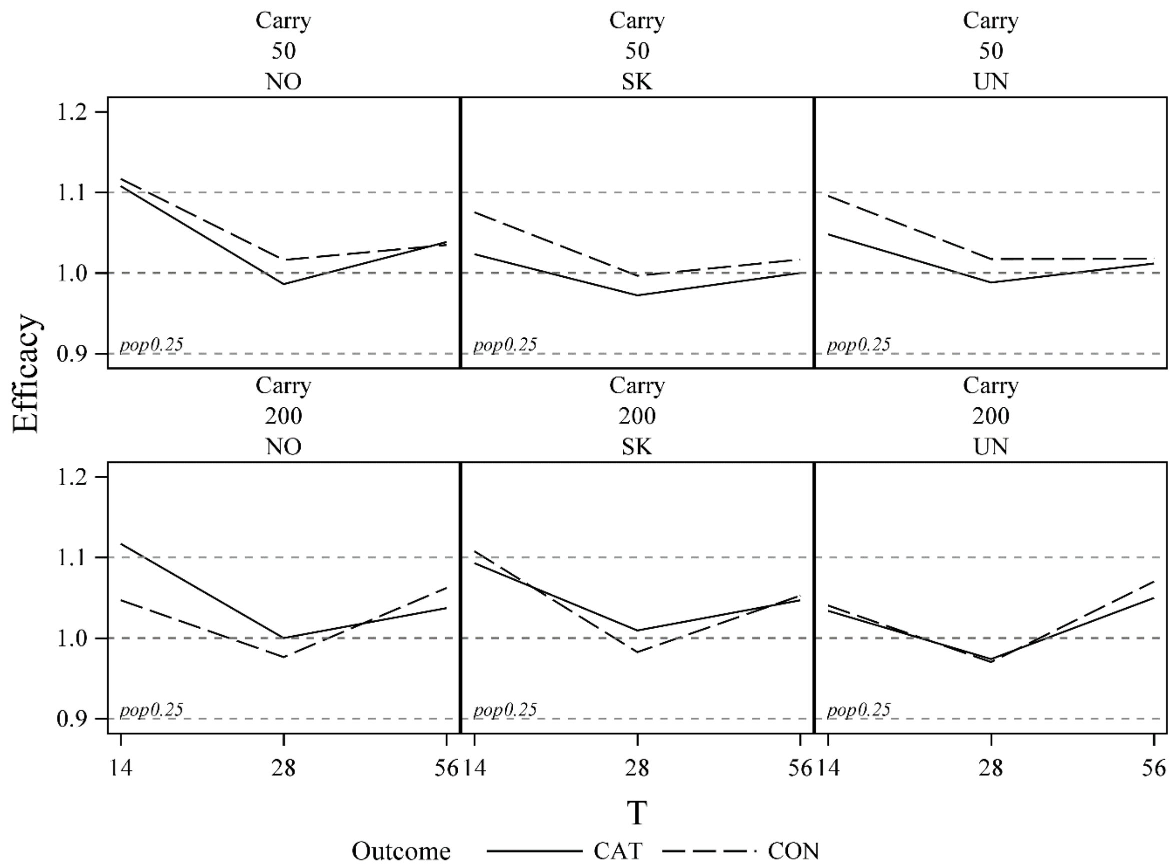

Figure 70. Efficacy Results for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

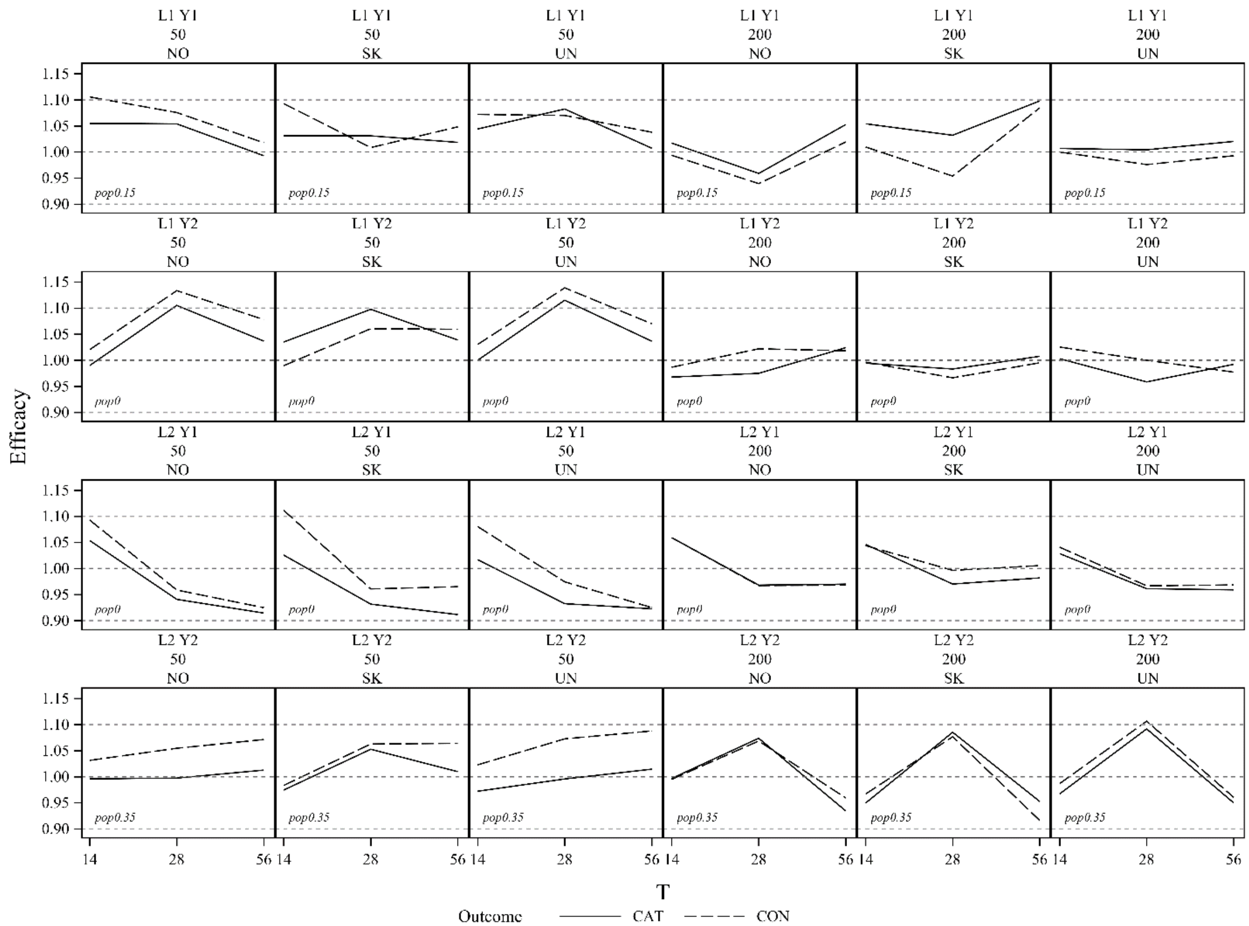

Figure 71. Efficacy Results for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

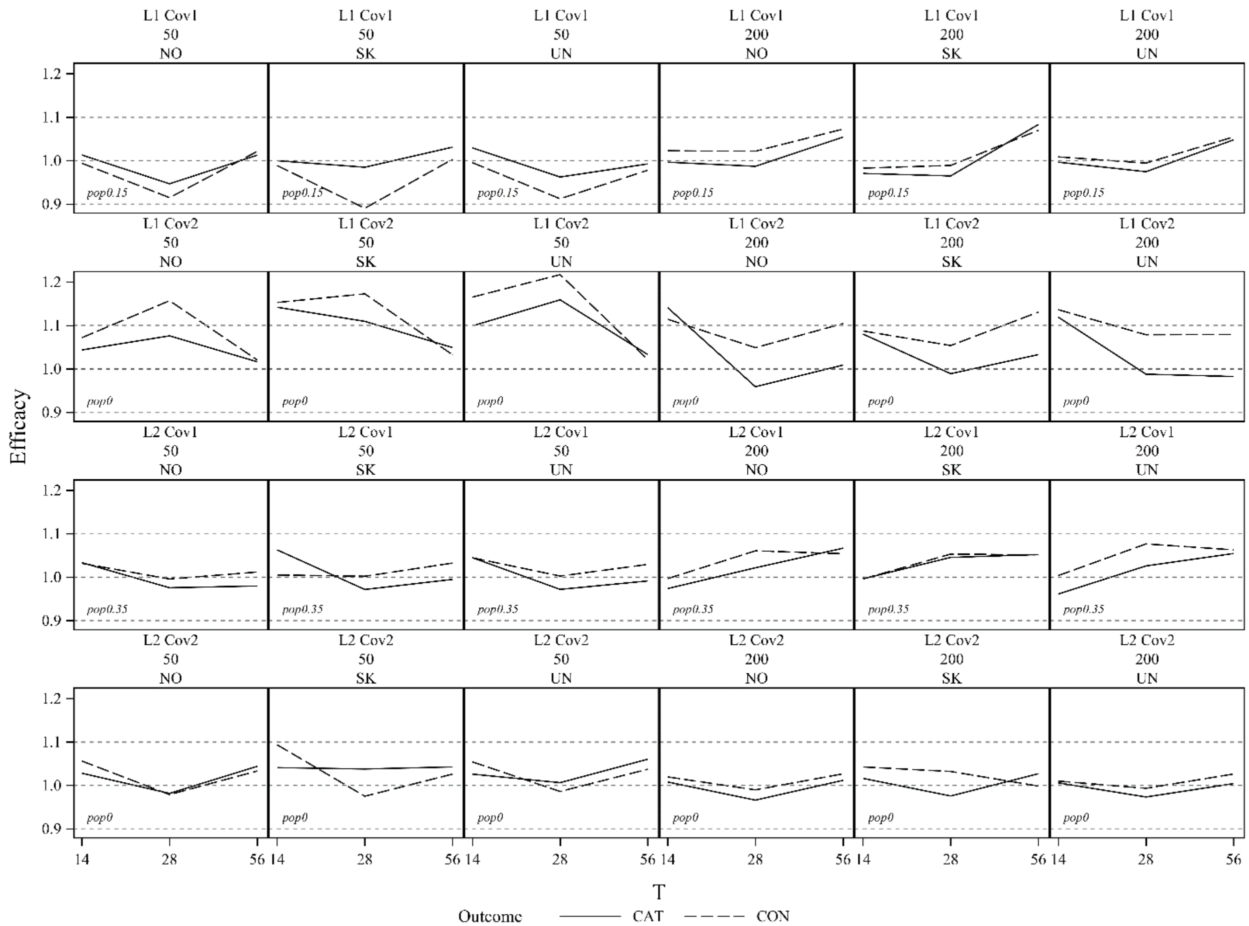

Figure 72. Efficacy Results for Actor and Partner Random Effect Variances across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

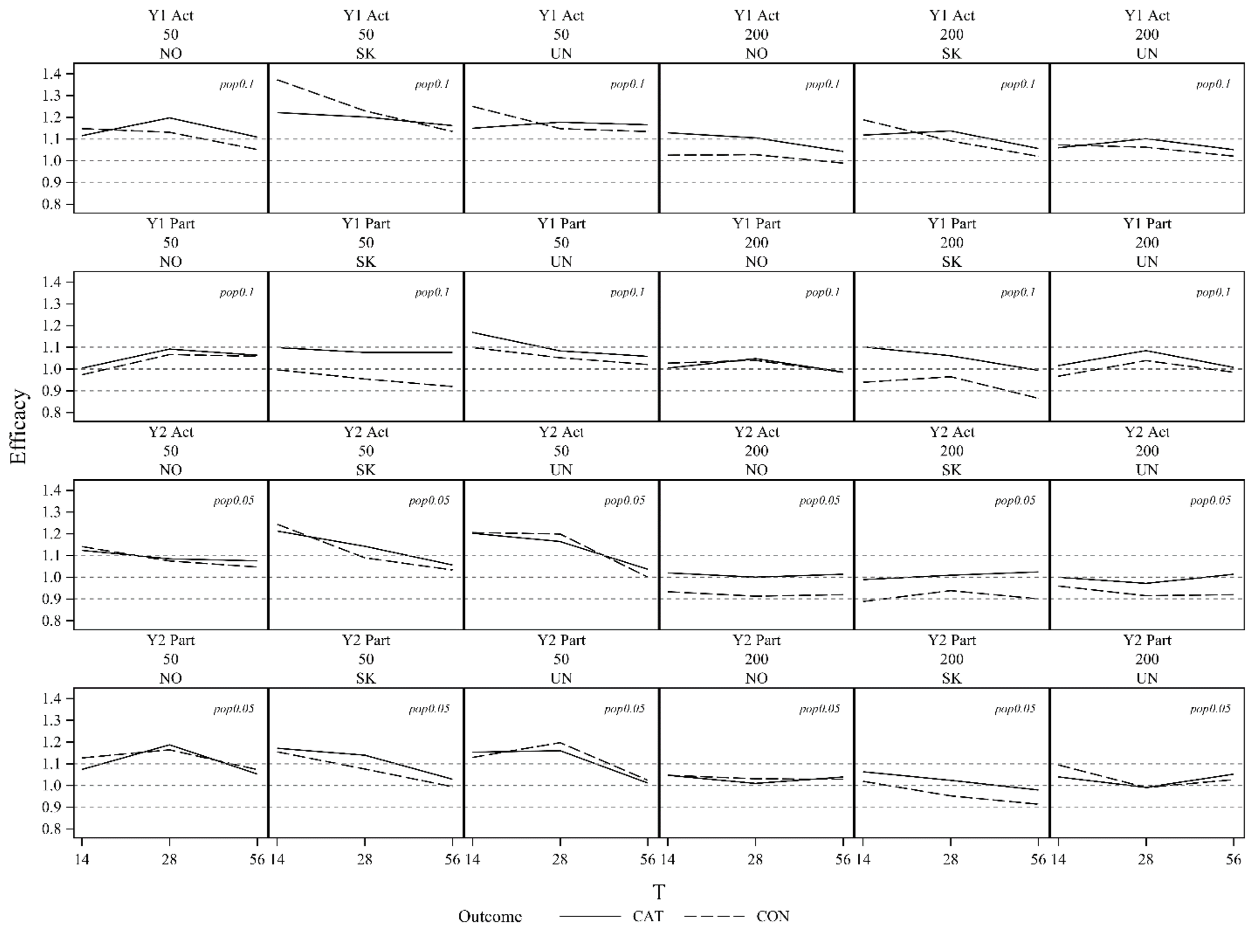

Figure 73. Efficacy Results for Autoregressive Random Effect Variances across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

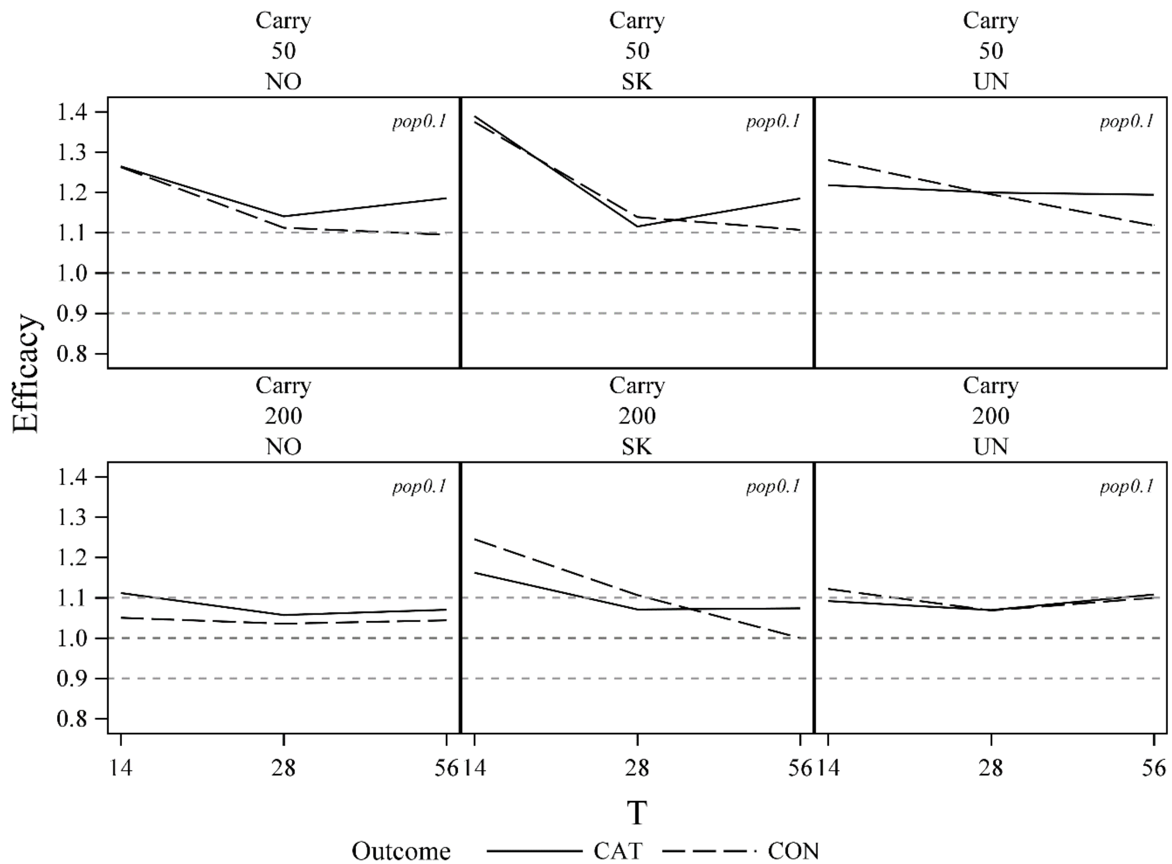

Figure 74. Efficacy Results for Level-1 Covariate Random Effect Variances across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

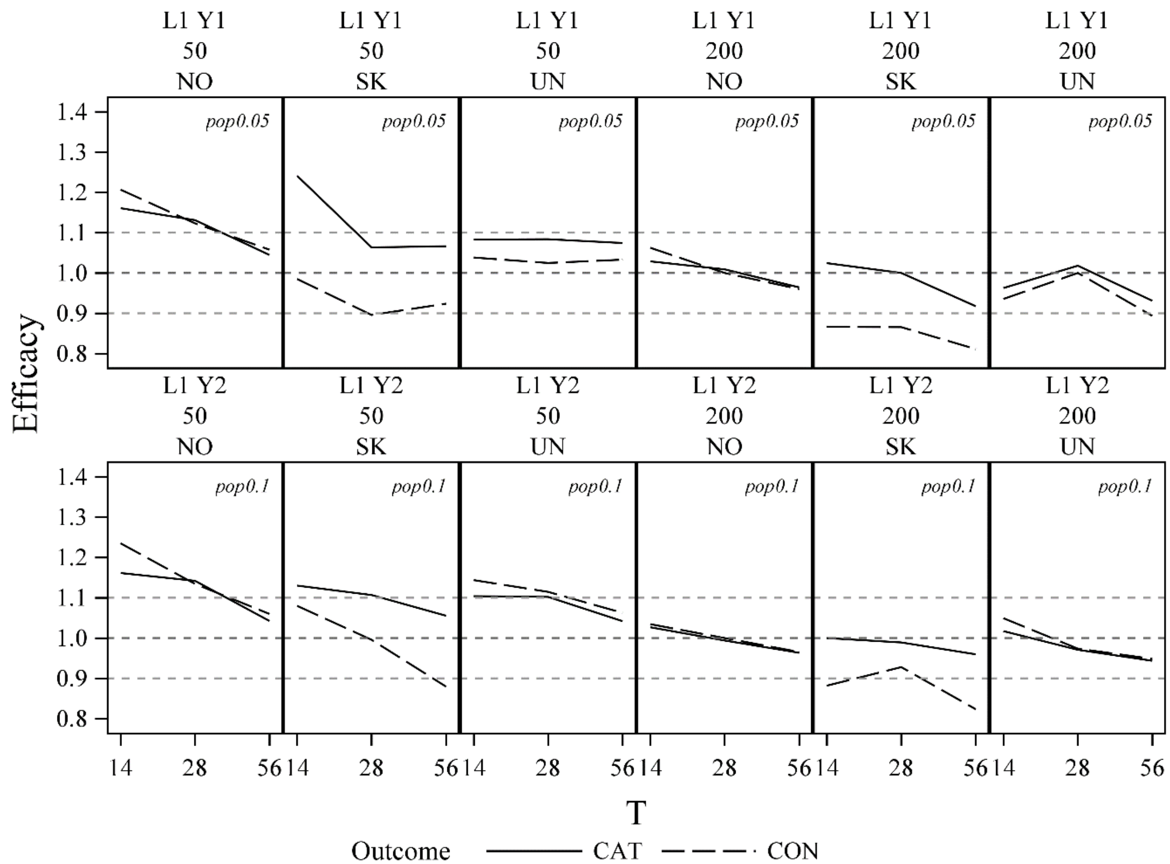

Figure 75. Efficacy Results for Level-1 Covariate Random Effect Variances across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

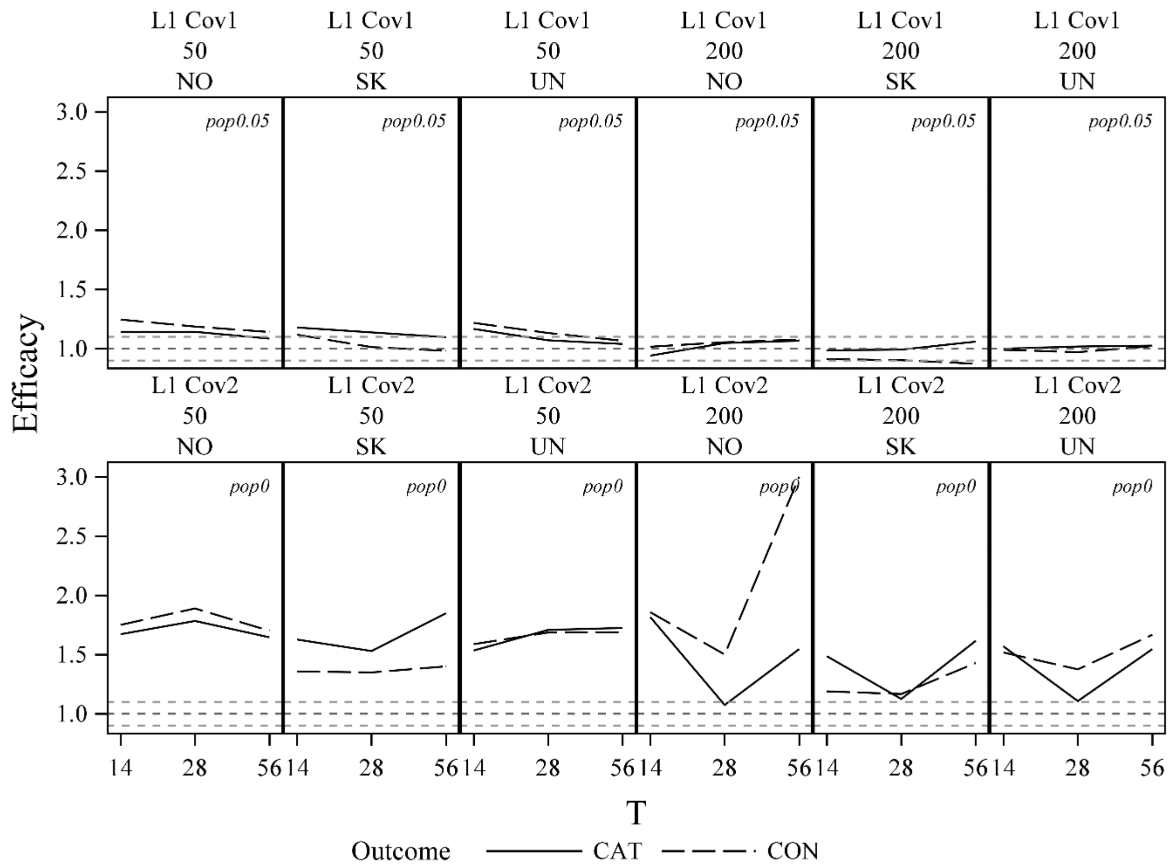

Figure 76. Power to Detect Significant Actor and Partner Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

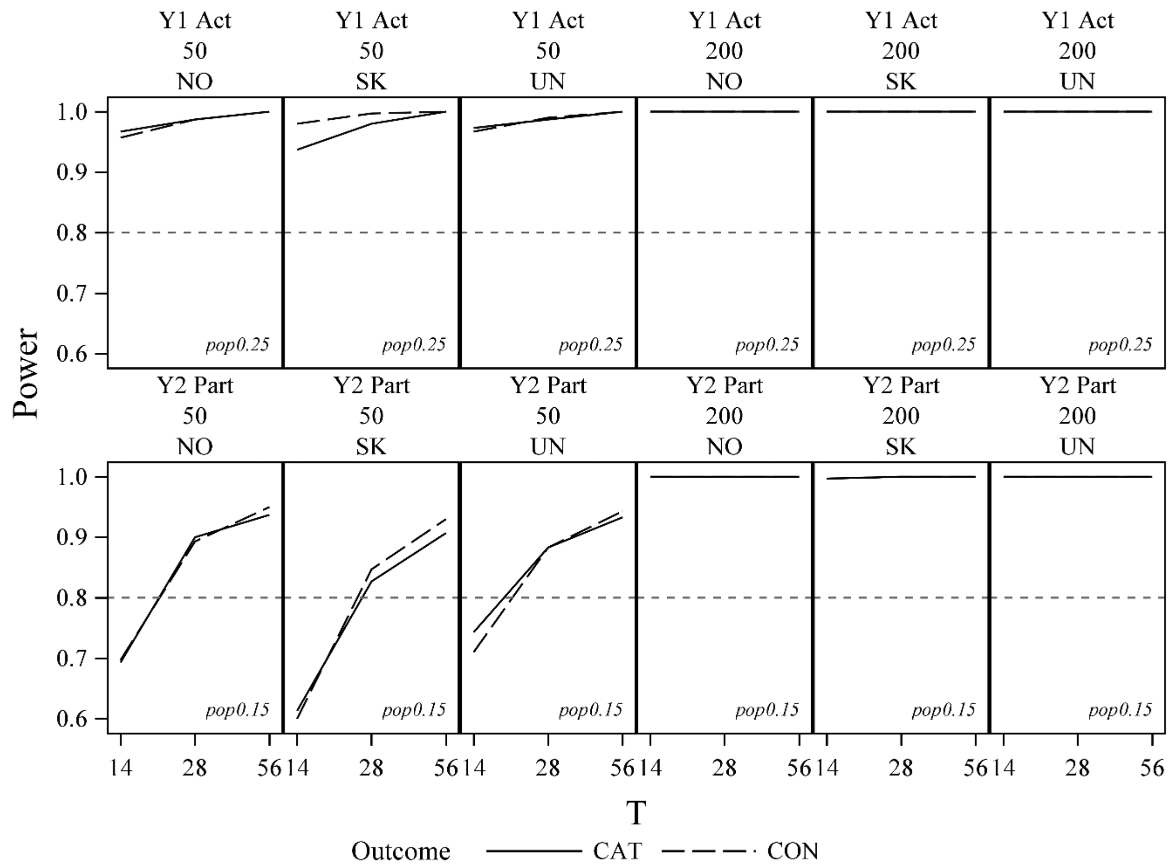

Figure 77. Power to Detect Autoregressive Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

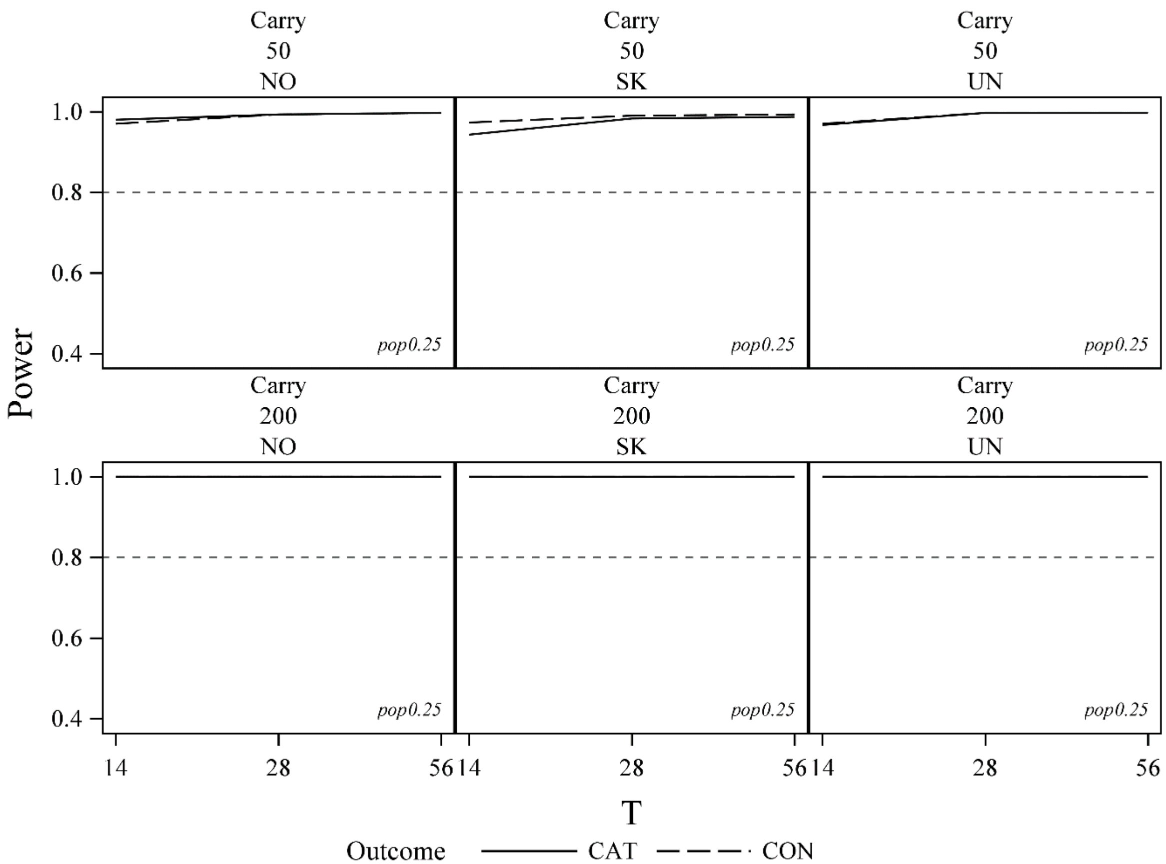

Figure 78. Power to Detect Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

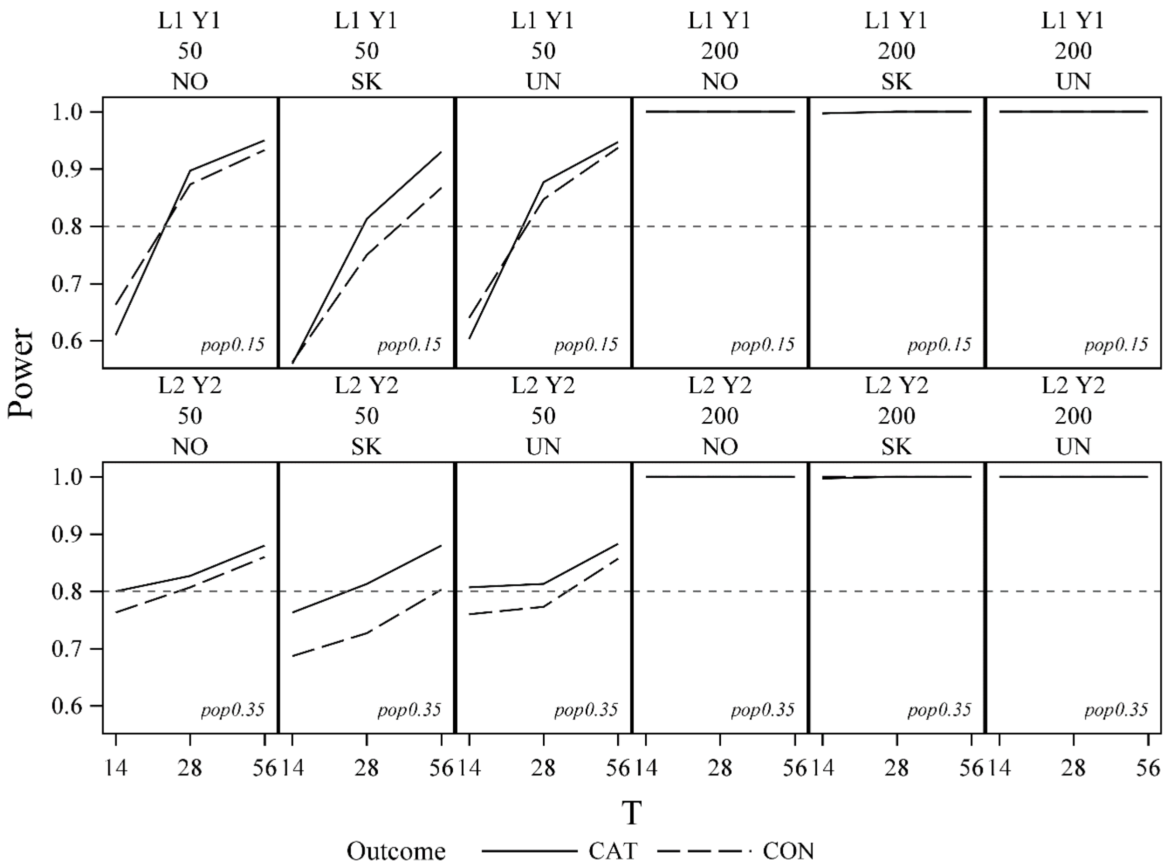

Figure 79. Power to Detect Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

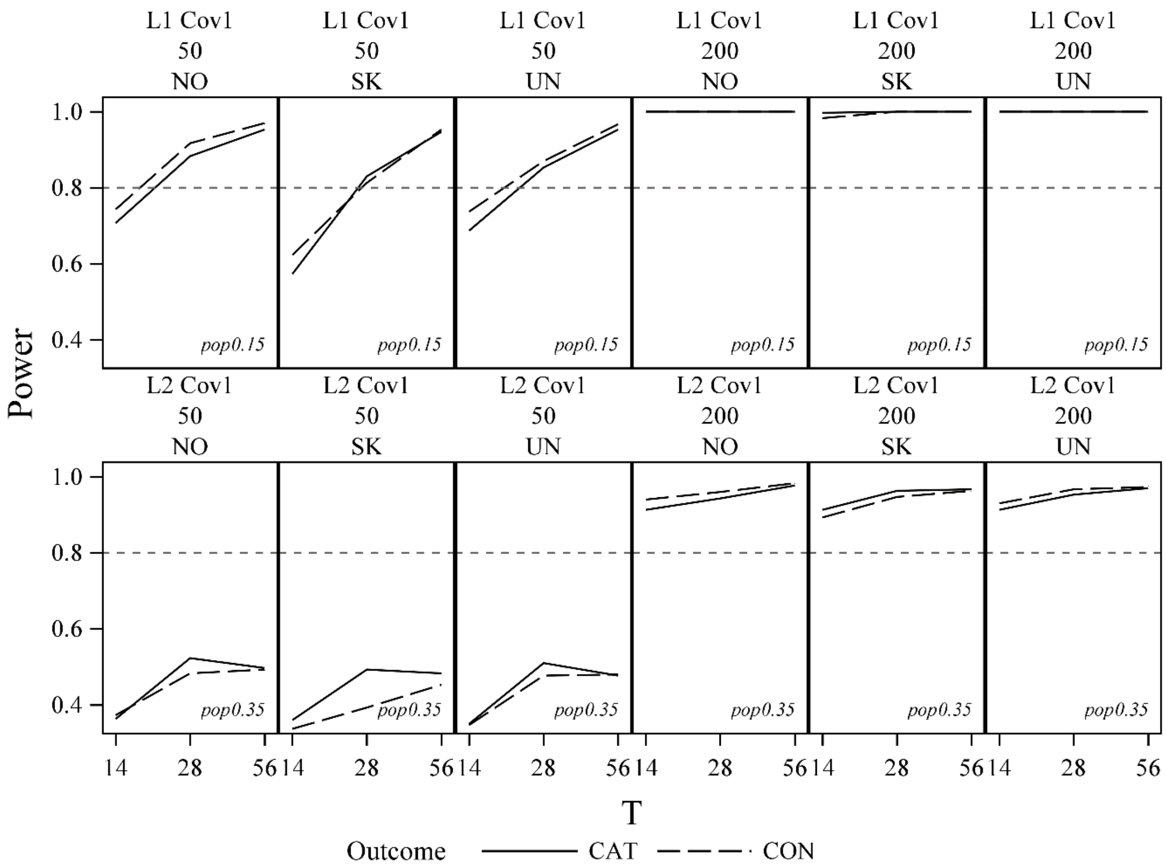

Figure 80. Type-I Error Rates for Actor and Partner Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

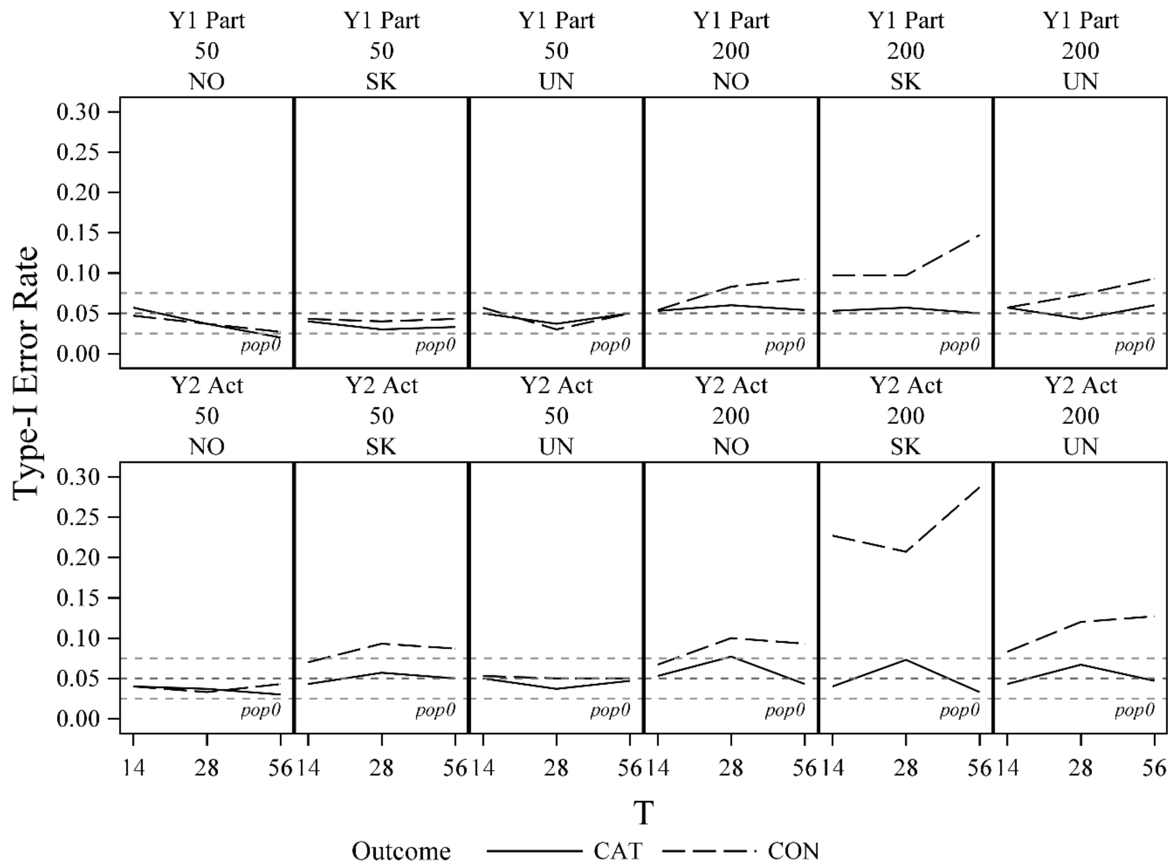

Figure 81. Estimated Type-I Error Rates for Autoregressive Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

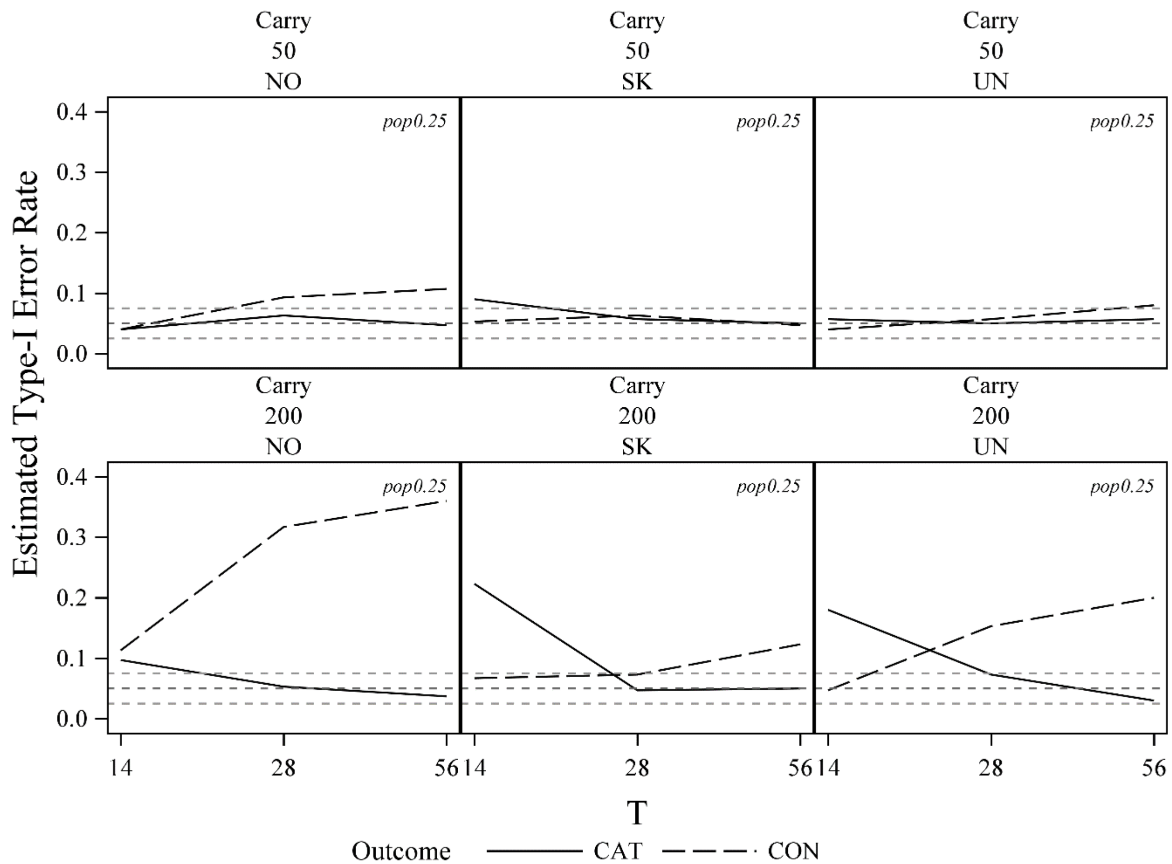

Figure 82. Type-I Error Rates for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Multivariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

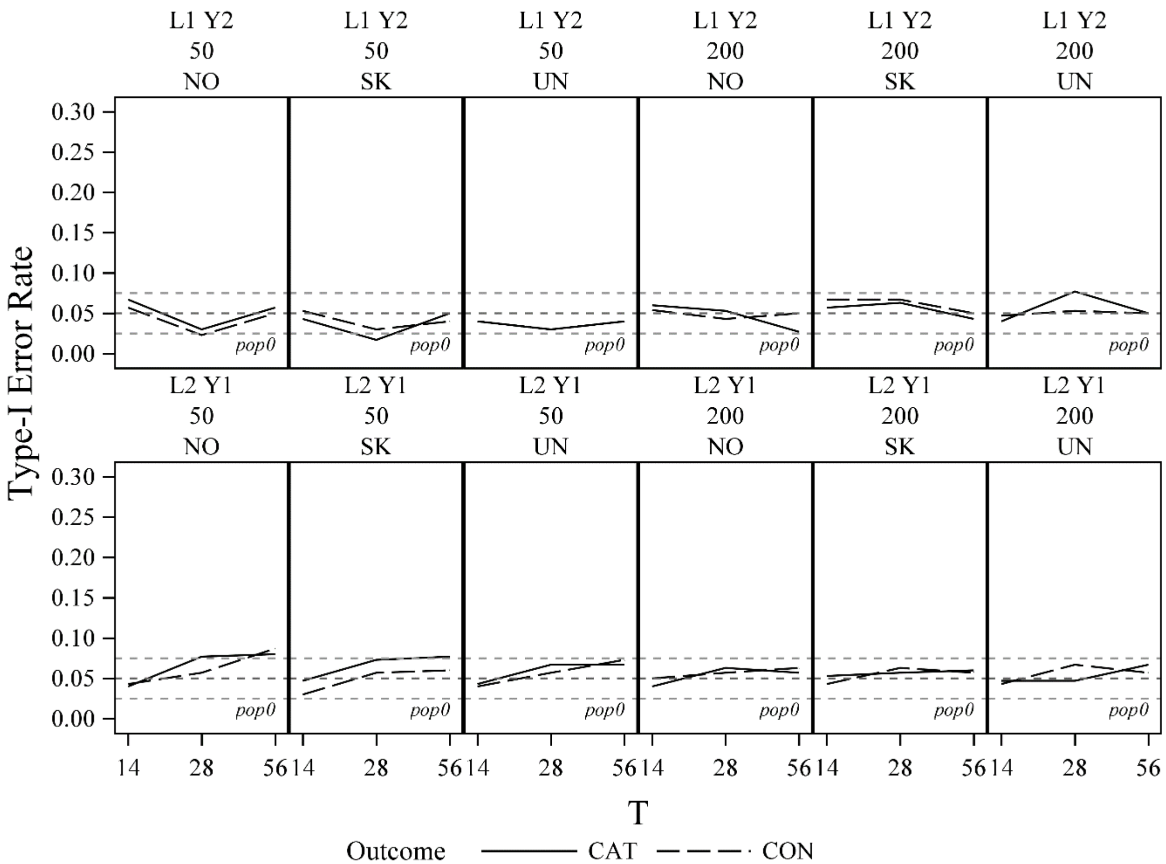

Figure 83. Type-I Error Rates for Level-1 and Level-2 Covariate Effects across Simulation Conditions for Univariate Categorical (CAT) and Continuous (CON) DSEM Models with 5-Category Outcomes

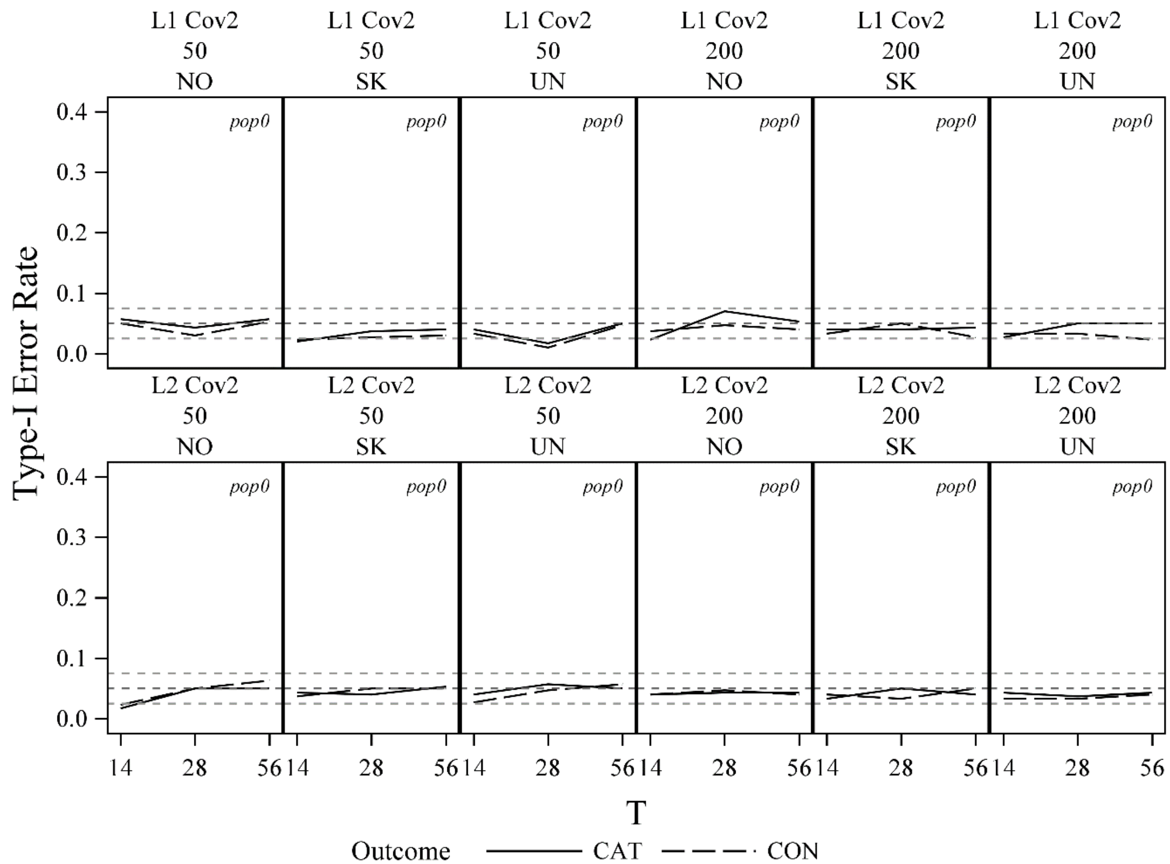

Figure 84. Average Bias for Actor, Partner, and Autoregressive Effects across Simulation Conditions for All Categorical DSEM Models

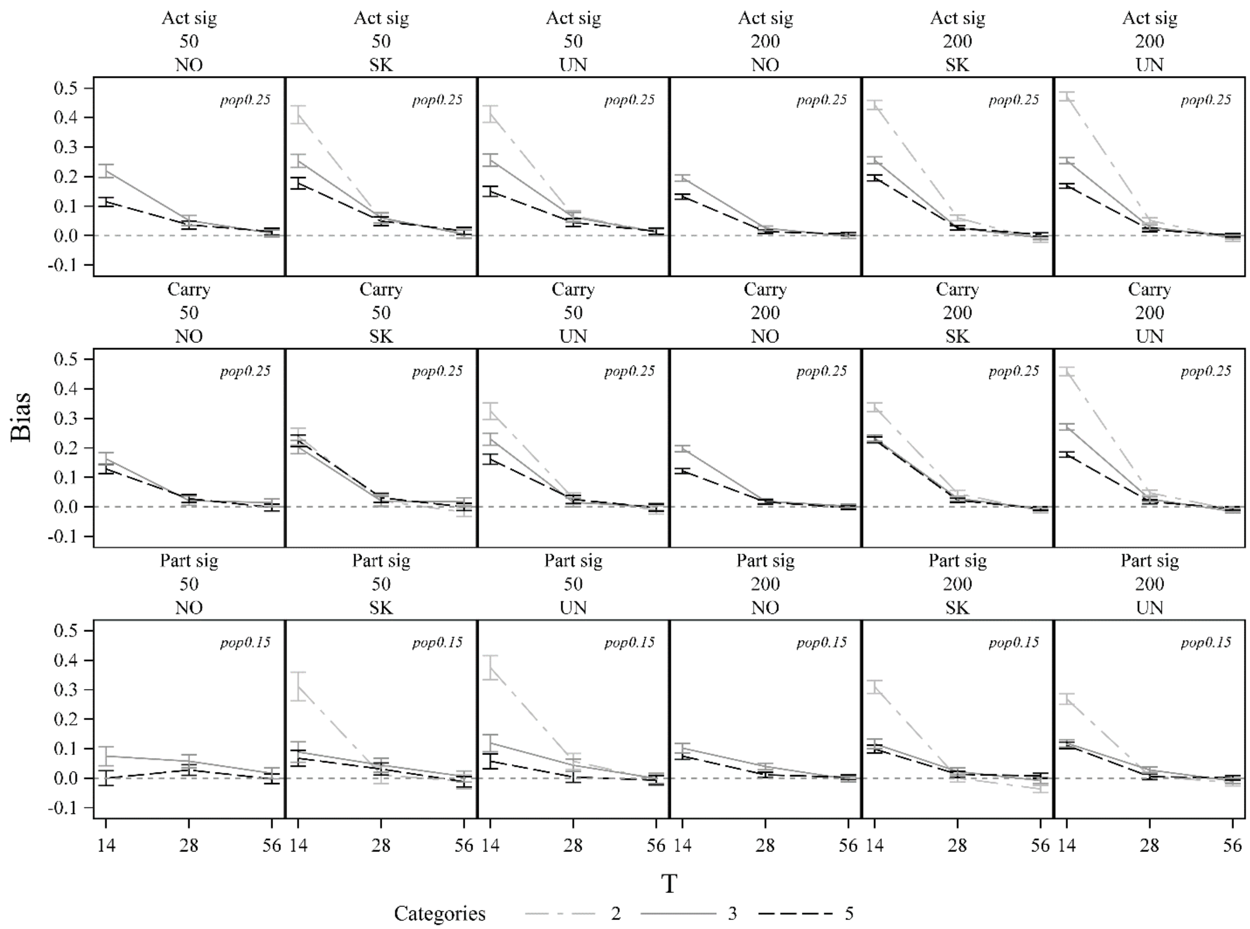

Figure 85. Average Bias for Standardized Level-1 and Level-2 Covariate Effects across Simulation Conditions for All Categorical DSEM Models

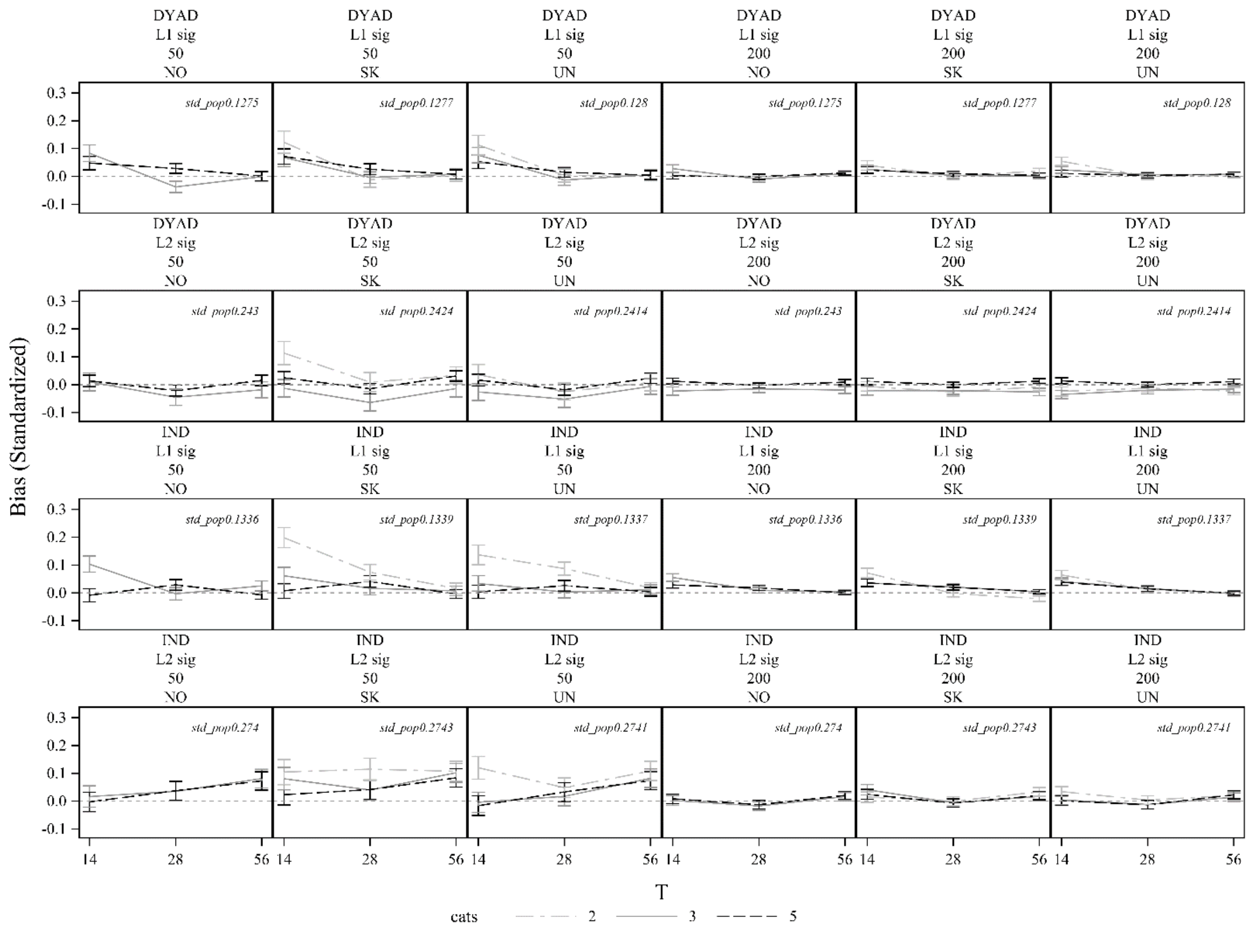

Figure 86. Average Median Value of the Posterior Distributions for Actor, Partner, and Autoregressive Random Effect Variances with a Population Value of 0.10 across Simulation Conditions for All Categorical DSEM Models

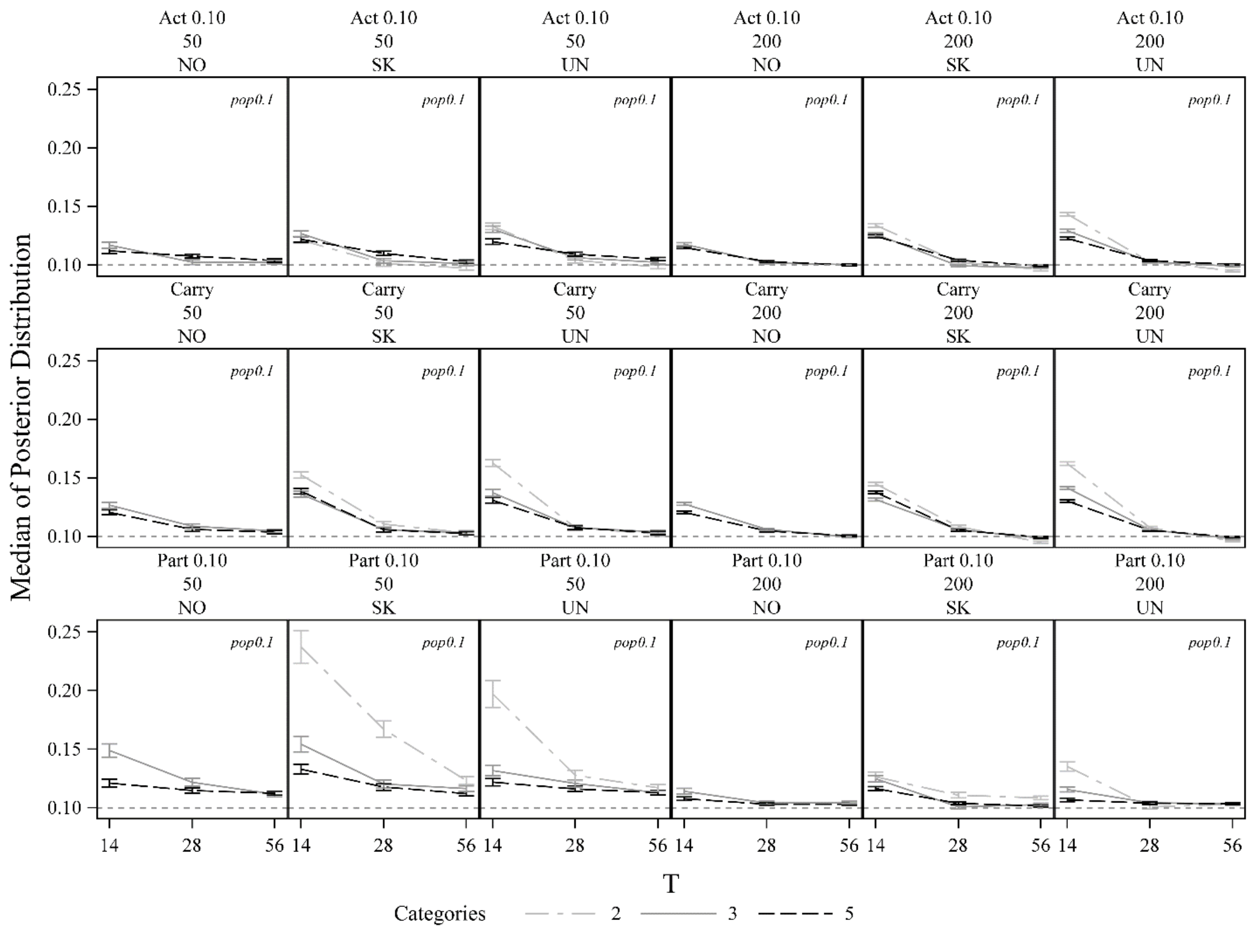

Figure 87. Average Median Value of the Posterior Distributions for Actor and Partner Random Effect Variances with a Population Value of 0.05 across Simulation Conditions for All Categorical DSEM Models

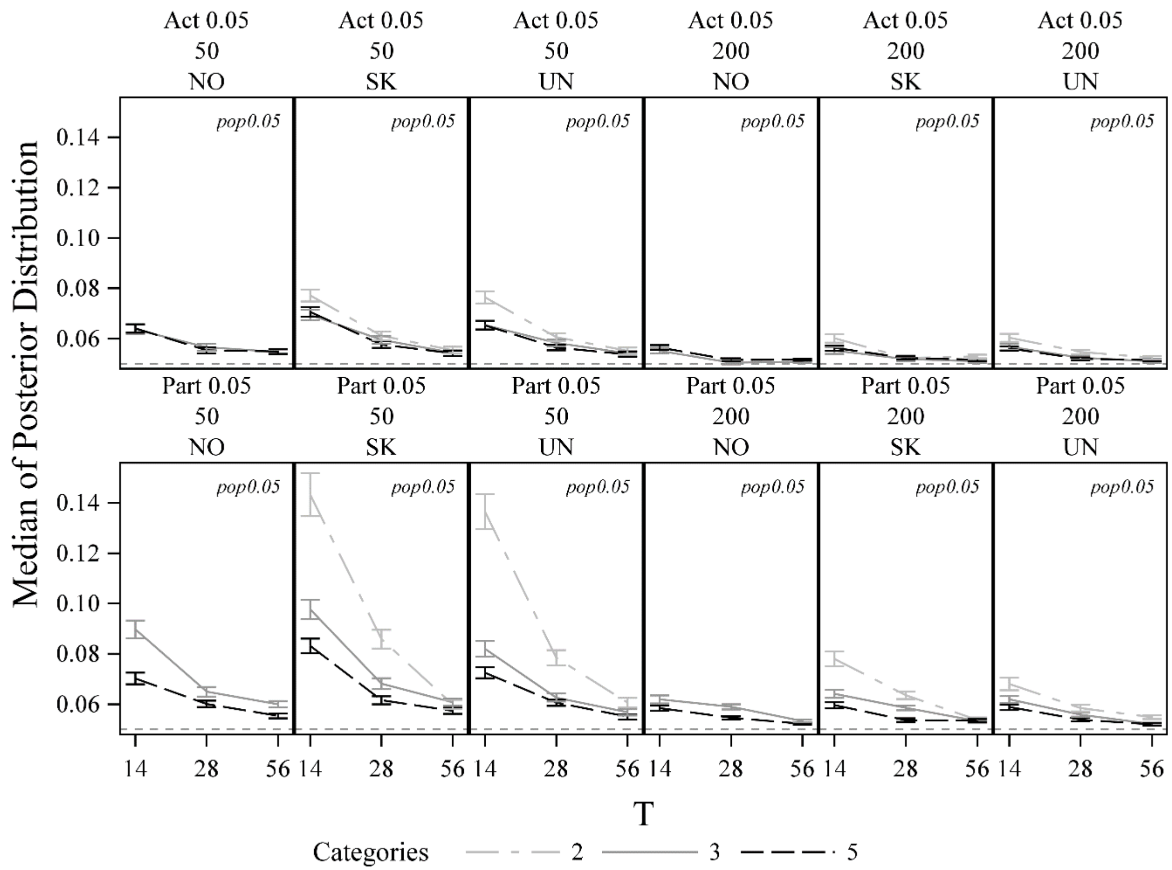

Figure 88. Average Median Value of the Posterior Distributions for Level-1 Random Effect Variances across Simulation Conditions for All Categorical DSEM Models

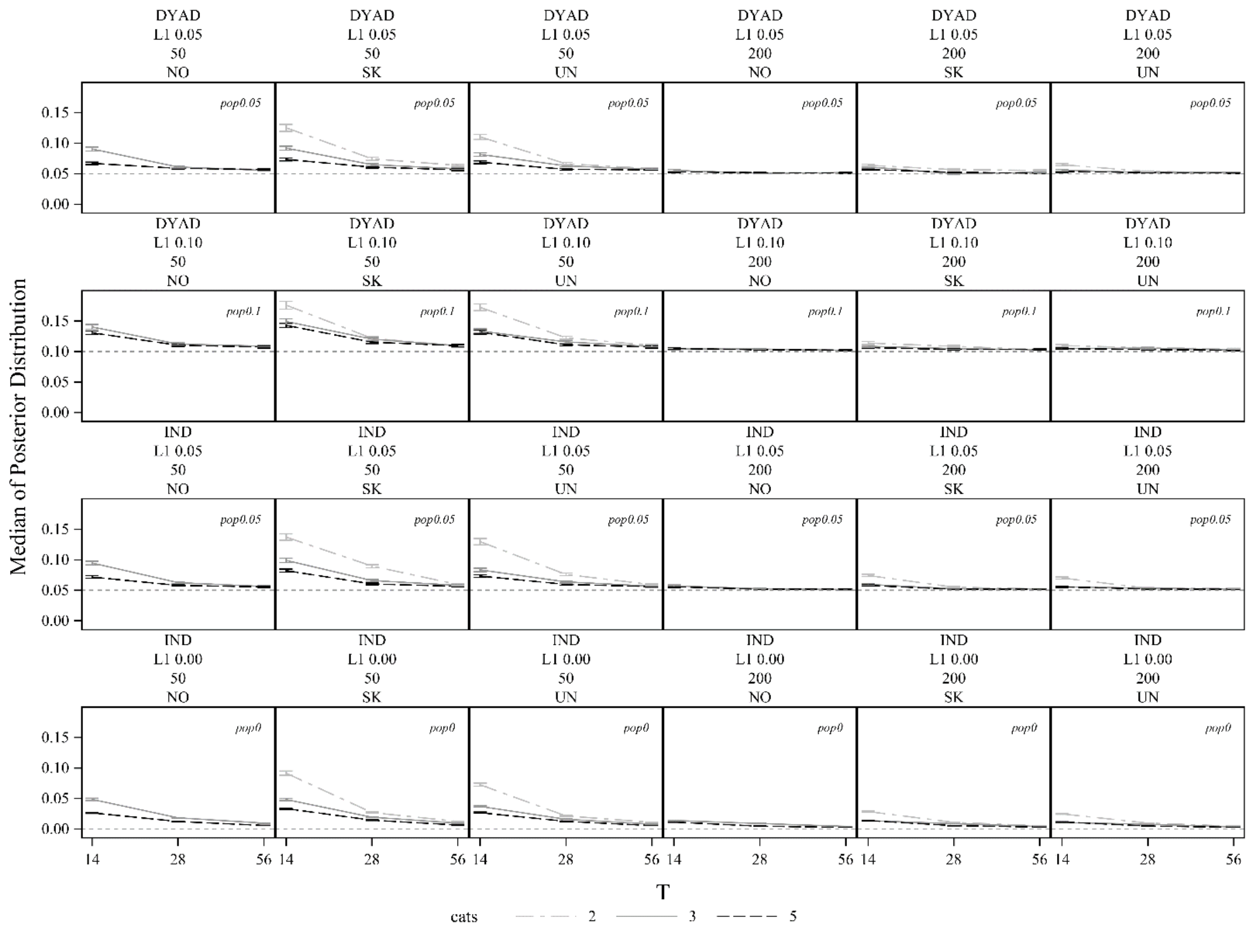

Figure 89. Power to Detect Significant Actor, Partner, and Autoregressive Effects across Simulation Conditions for All Categorical DSEM Models

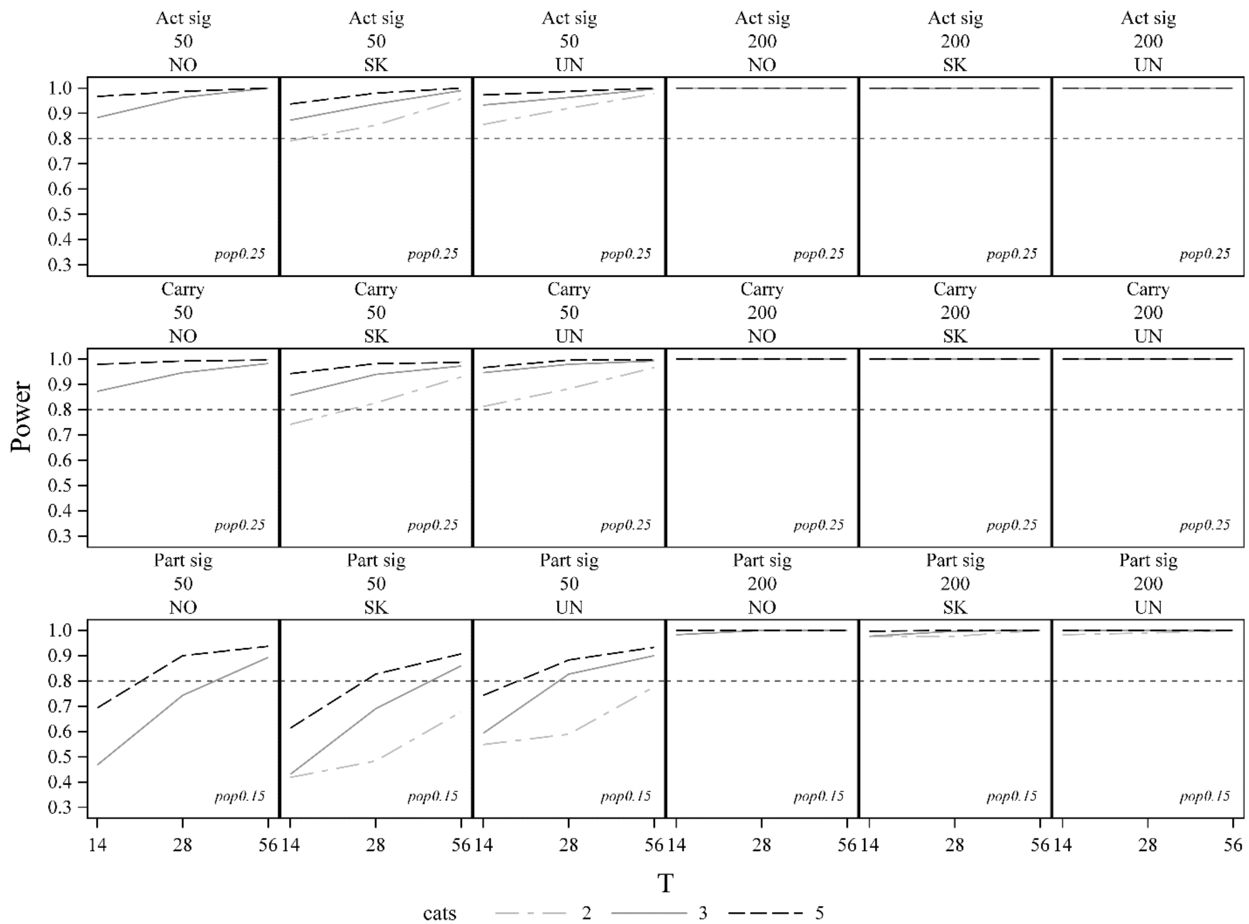

Figure 90. Power to Detect Significant Level-1 and Level-2 Covariate Effects across Simulation Conditions for All Categorical DSEM Models

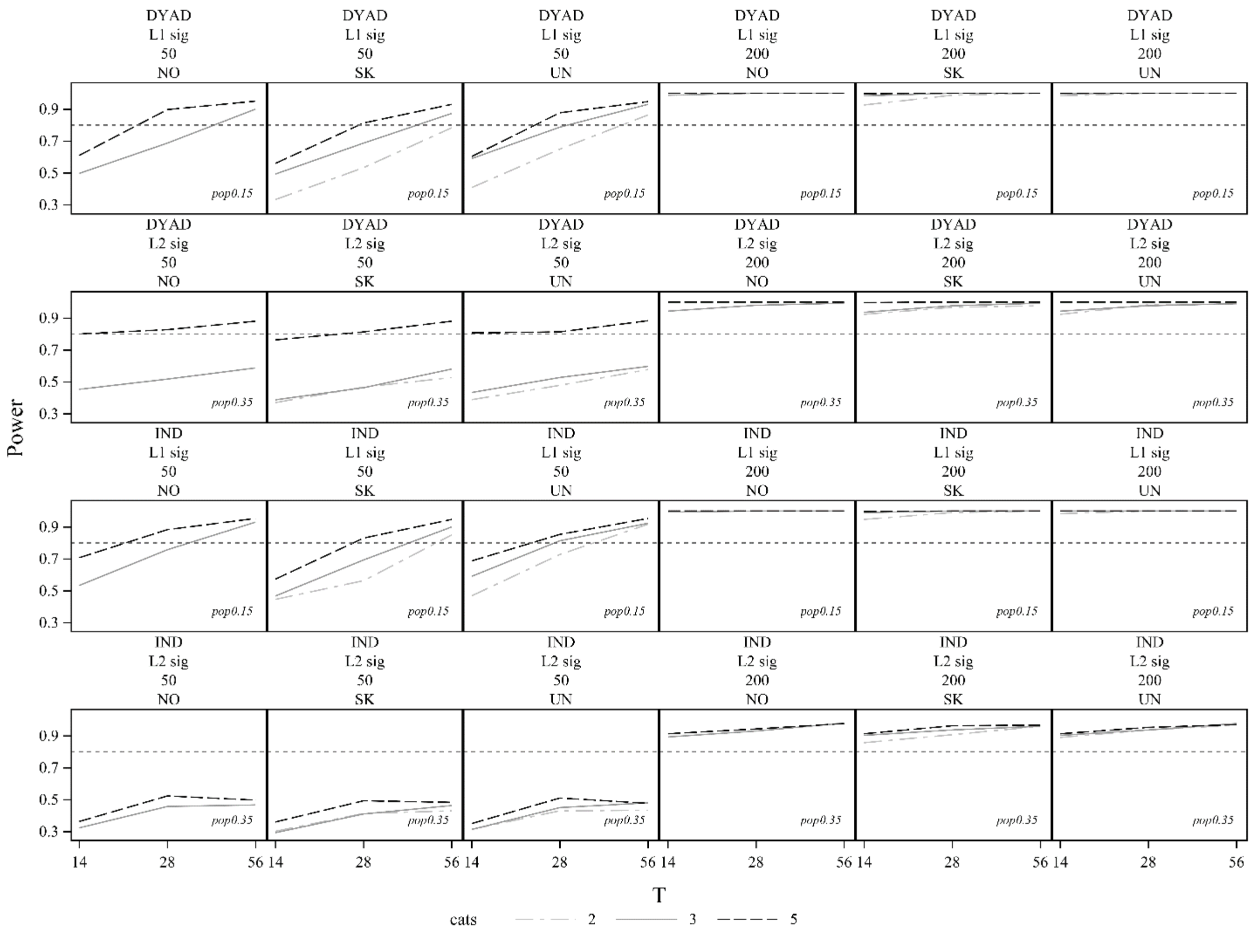

Figure 91. Type-I Error Rates for Actor and Partner Effects across Multivariate Simulation Conditions for All Categorical DSEM Models

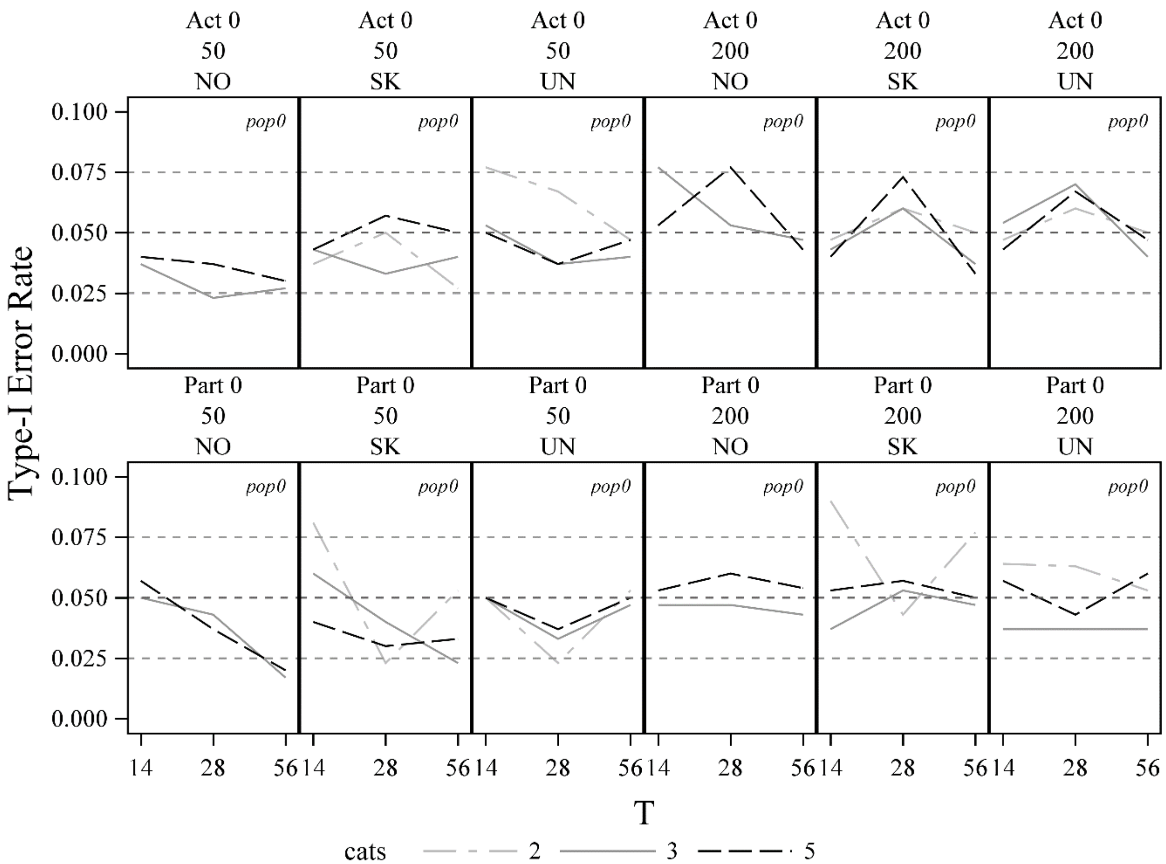

Figure 92. Type-I Error Rates for Level-1 and Level-2 Covariate Effects across Simulation Conditions for All Categorical DSEM Models

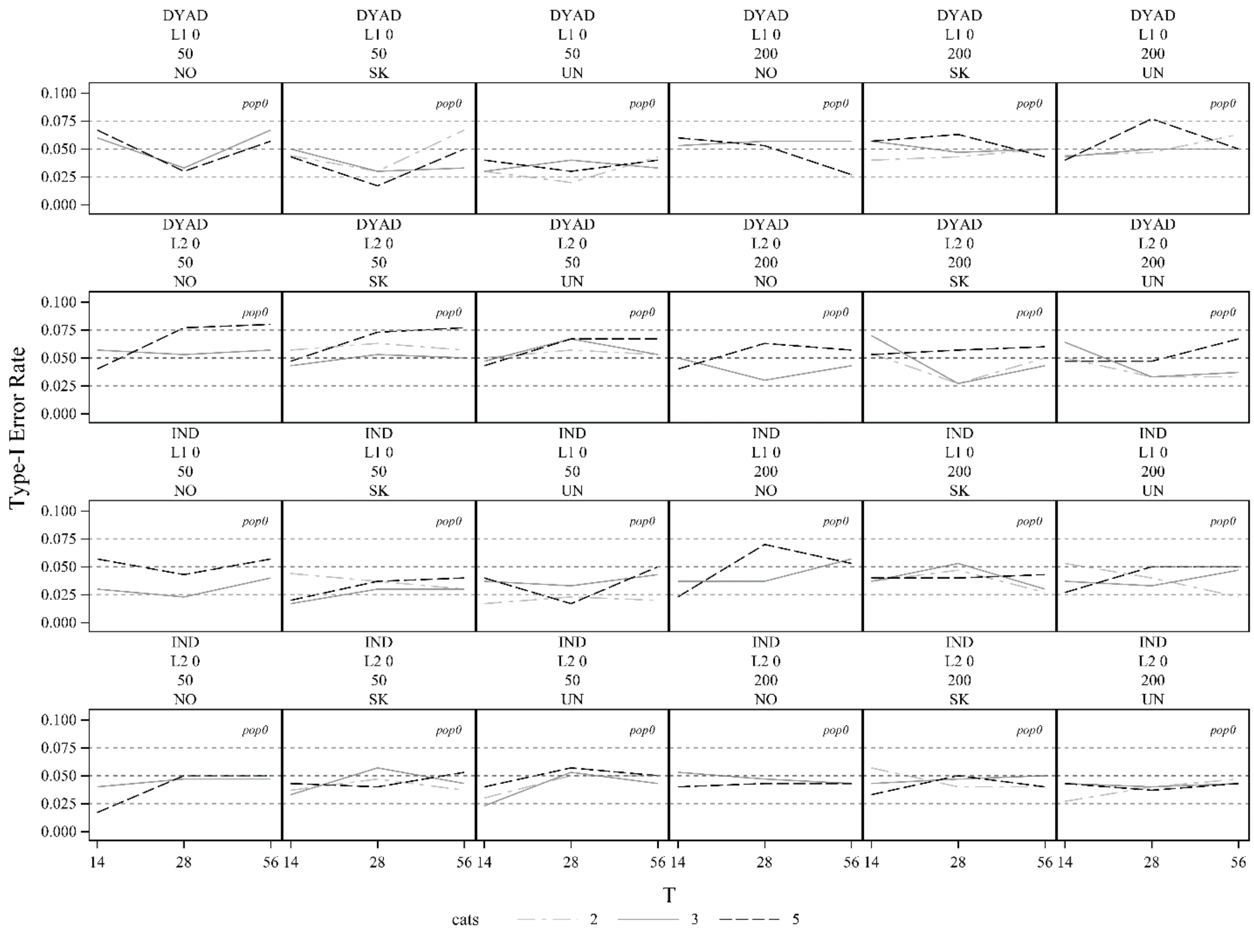

Figure 93. Average Bias for Actor, Partner, and Autoregressive Effects across Simulation Conditions for All Continuous DSEM Models

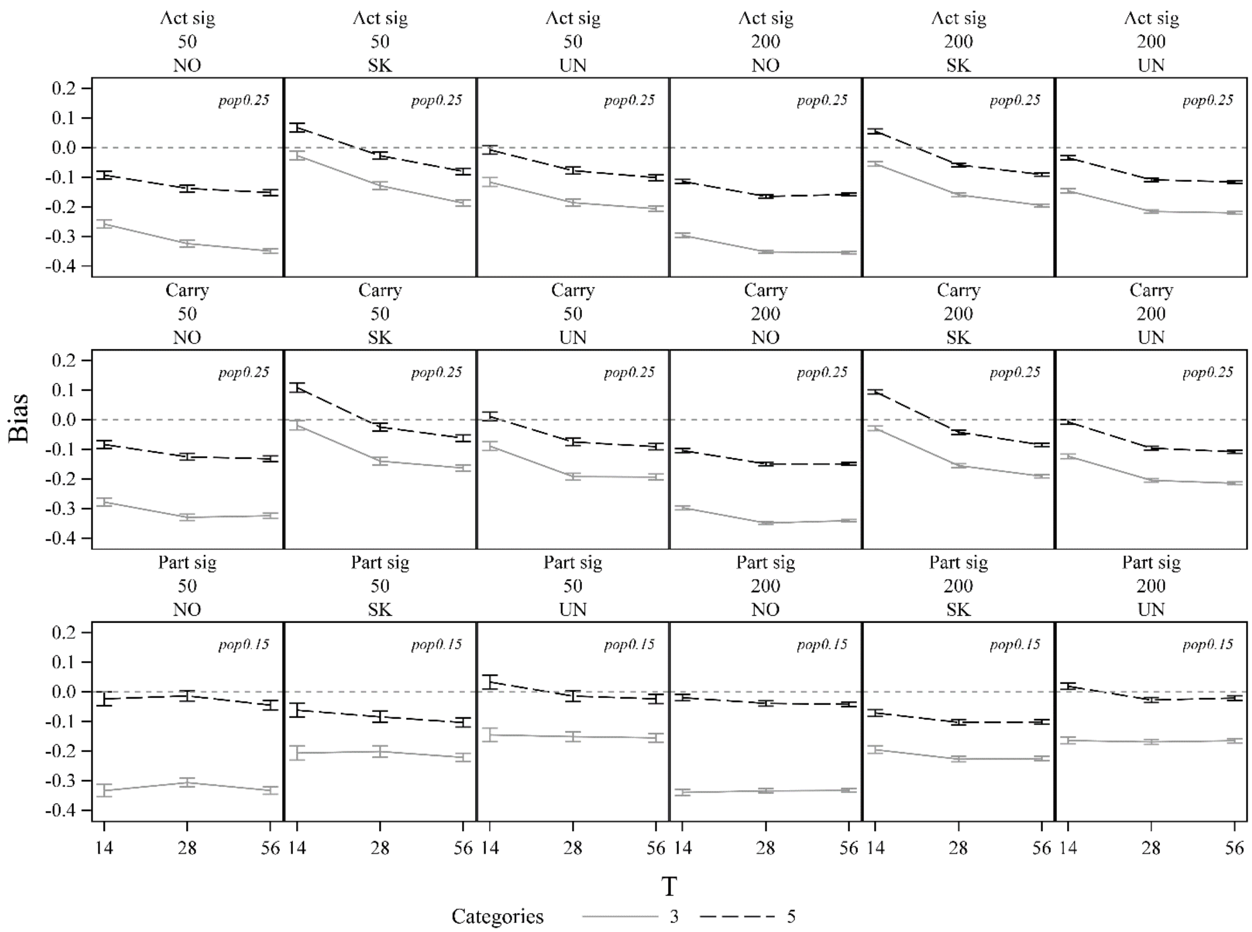

Figure 94. Average Bias for Standardized Level-1 and Level-2 Covariate Effects across Simulation Conditions for All Continuous DSEM Models

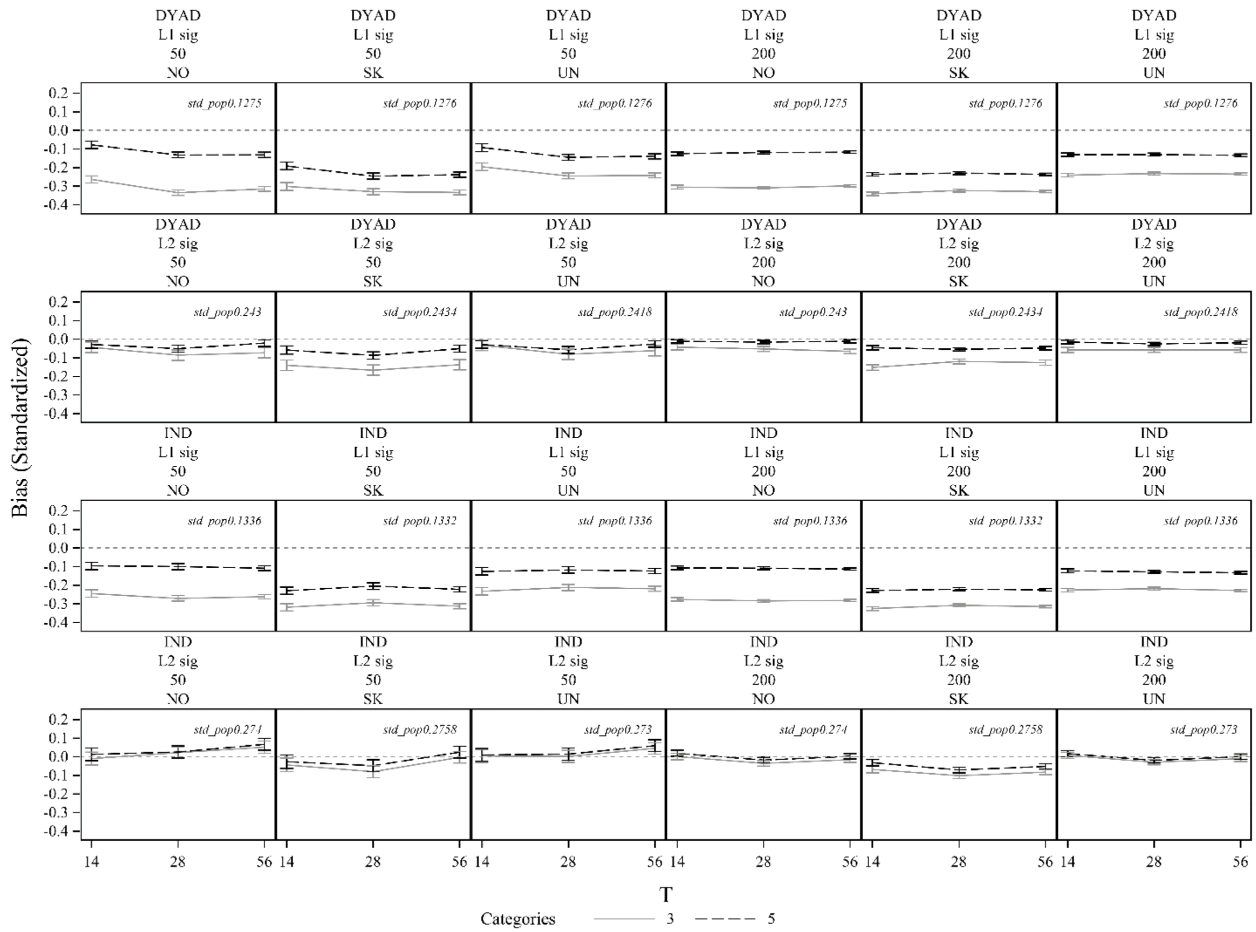

Figure 95. Average Median Value of the Posterior Distributions for Actor, Partner, and Autoregressive Random Effect Variances with a Population Value of 0.10 across Simulation Conditions for All Continuous DSEM Models

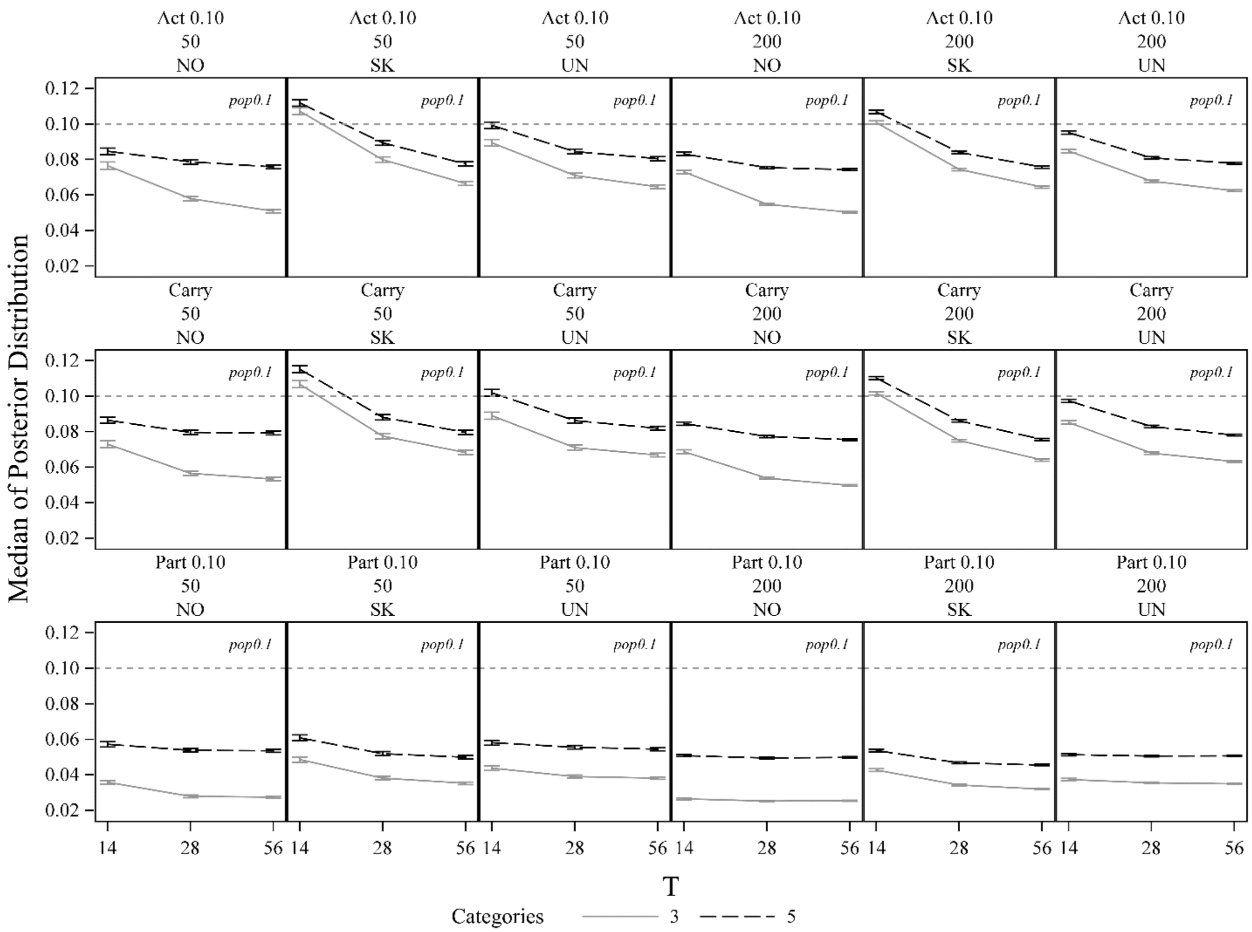

Figure 96. Average Median Value of the Posterior Distributions for Actor and Partner Random Effect Variances with a Population Value of 0.05 across Simulation Conditions for All Continuous DSEM Models

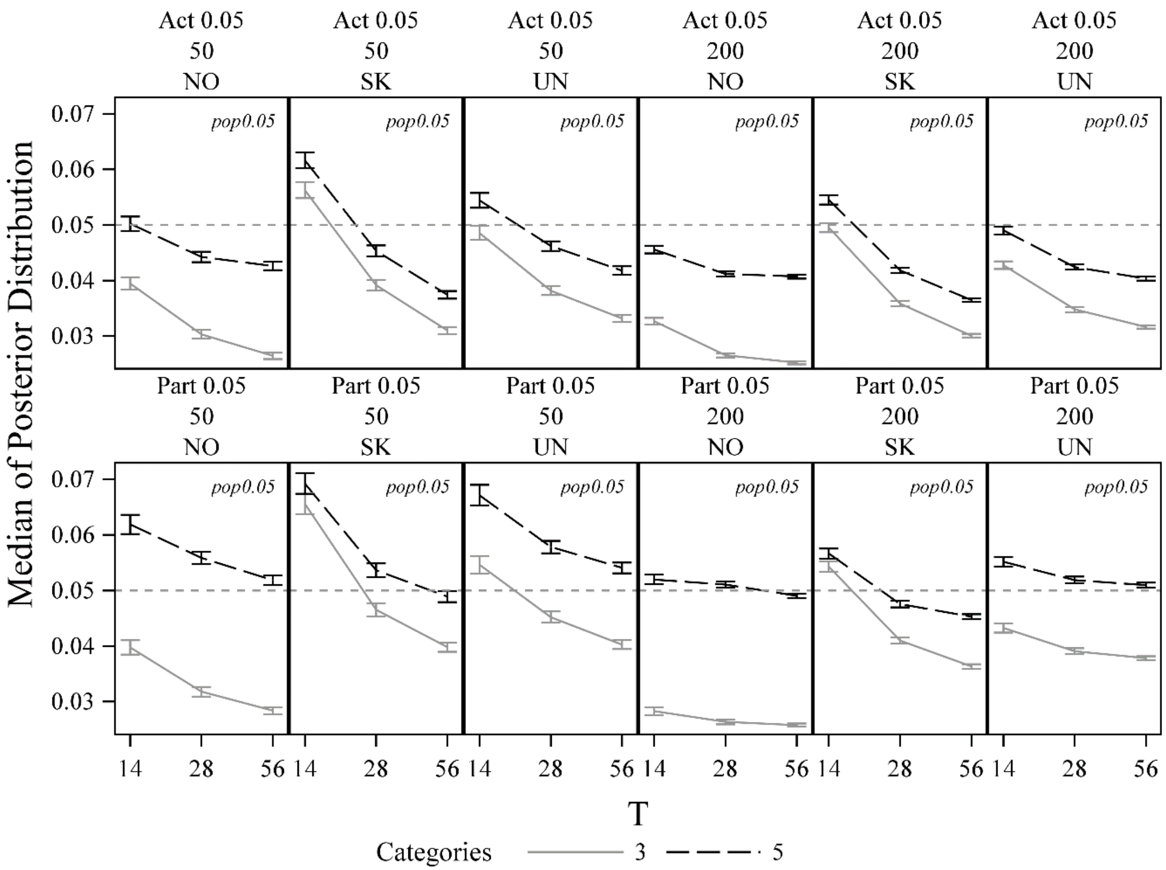

Figure 97. Average Median Value of the Posterior Distributions for Level-1 Random Effect Variances across Simulation Conditions for All Categorical DSEM Models

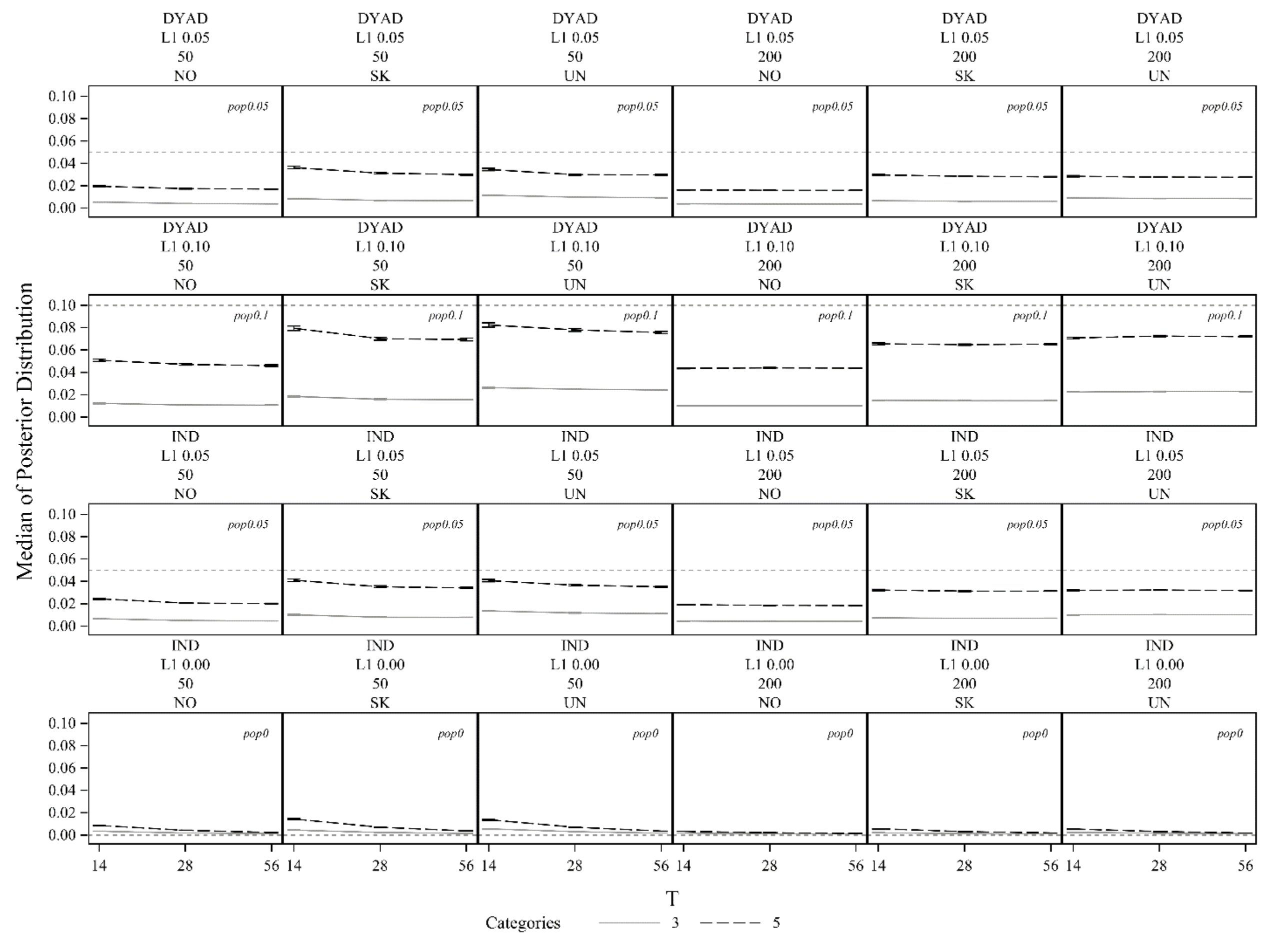

Figure 98. Power to Detect Significant Actor, Partner, and Autoregressive Effects across Simulation Conditions for All Categorical DSEM Models

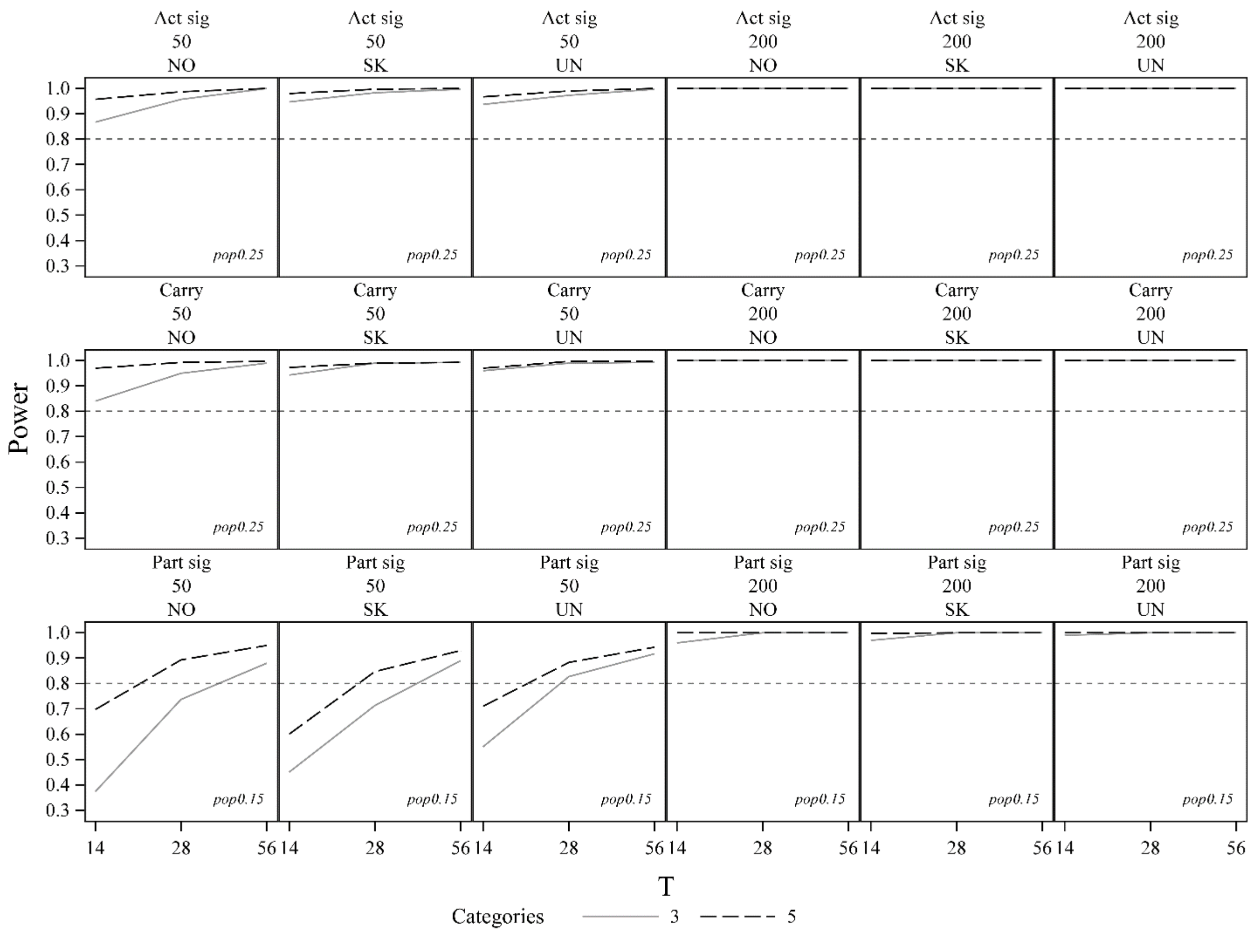

Figure 99. Power to Detect Significant Level-1 and Level-2 Covariate Effects across Simulation Conditions for All Categorical DSEM Models

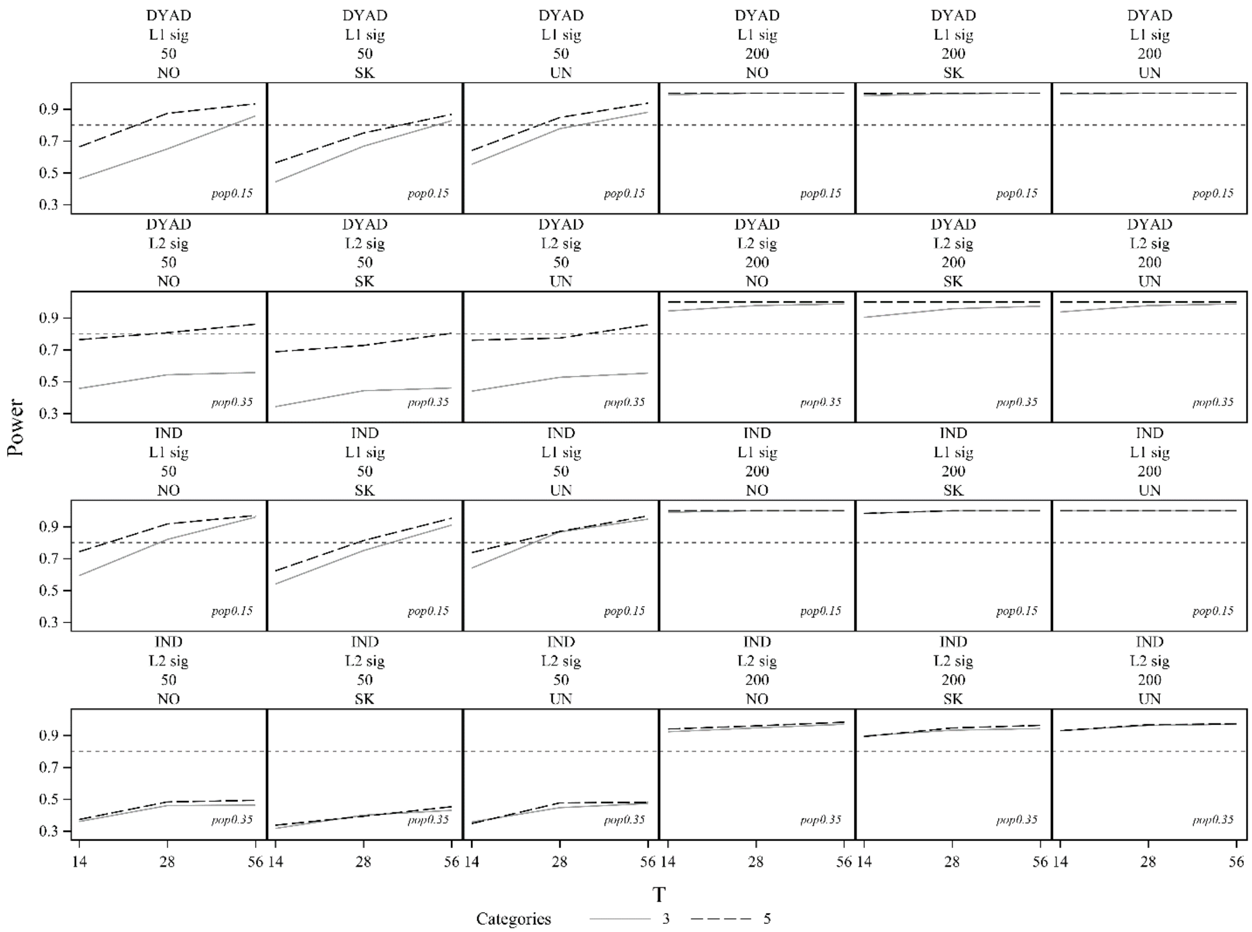

Figure 100. Type-I Error Rates for Actor and Partner Effects across Multivariate Simulation Conditions for All Categorical DSEM Models

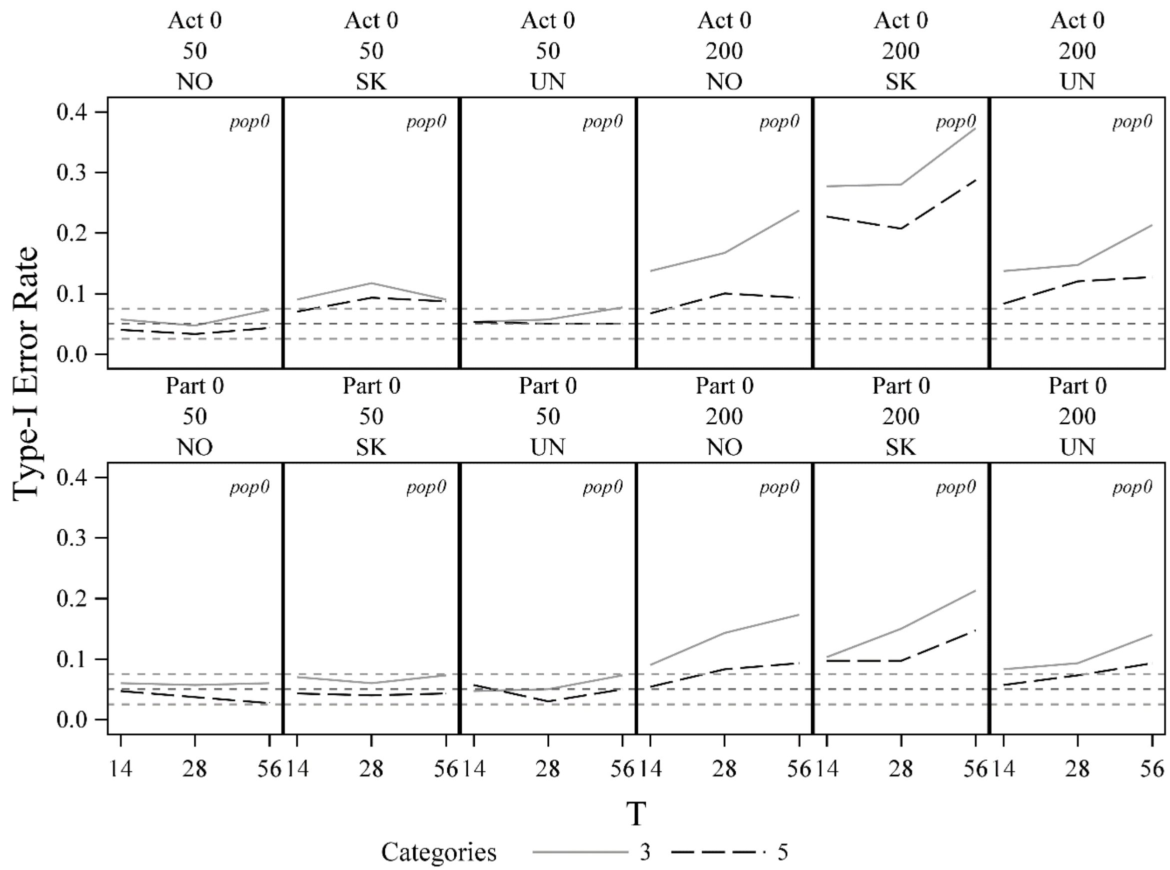

Figure 101. Type-I Error Rates for Level-1 and Level-2 Covariate Effects across Simulation Conditions for All Categorical DSEM Models