Dynamic Structural Equation Models: Promising Yet Concerning

Suryadyuti Baral & Patrick J. Curran*

Department of Psychology and Neuroscience, University of North Carolina - Chapel Hill, Chapel Hill, NC

https://doi.org/10.33697/ajur.2023.096

*Student: sonnet@email.unc.edu**
Mentor: curran@unc.edu

ABSTRACT
Dynamic Structural Equation Model (DSEM) is a powerful statistical modeling approach that has recently gained popularity among researchers studying intensive longitudinal data. Despite its exciting potential, the stability and replicability of DSEM is yet to be closely examined. This study empirically investigates DSEM using recently published data to explore its strengths and potential limitations. The results show that while some of its parameter estimates are stable, others are characterized by substantial variation as a function of seemingly innocuous initial model estimation conditions. Indeed, some parameters fluctuate between significance and non-significance for the same model estimated using the same data. The instability of DSEM estimates poses a serious threat to the internal and external validity of conclusions drawn from its analyses, challenging the reproducibility of findings from applied research. Given the recent focus on the replication crisis in psychology, it is critical to address these issues as the popularity of DSEM in psychological research continues to rise. Several potential solutions are investigated to address this problem and recommendations of best practice are offered to applied researchers who plan to use DSEM in intensive longitudinal data analysis.

KEYWORDS
Dynamic Structural Equation Model; Bayesian; Robust Estimation; Intensive Longitudinal Data

INTRODUCTION

Human psychology weaves together like an intricate tapestry, where the threads of cognition and behavior intertwine, creating a rich and complex individual. One of the most important goals in the behavioral sciences is to disentangle the threads and study the cause—effect pathways that shape the individual. In order to study such relationships, researchers need longitudinal data—that is, repeated assessments of individuals collected over the course of weeks, months or even years. With increasing recognition that life unfolds continuously over time, there has been a push towards intensive longitudinal data (ILD) analysis in which a large number of assessments are taken in shorter time intervals of days or even hours.[1] Although characterized by many strengths, ILD comes with its own set of unique challenges in measurement and modeling. A key challenge is to satisfactorily examine co-developmental processes where cause-and-effect influences are studied in two or more processes over multiple time points. Quantitative models like autoregressive cross-lagged (ARCL) have been developed for the analysis of longitudinal data, although it is increasingly appreciated that such traditional models fail to accurately capture co-developmental processes in ILD.[2] The tapestry of human psychology, quite unsurprisingly, is just too tightly knit and it has challenged the statistical might of generations of quantitative psychologists.

However, a novel method of modeling these dynamic processes has recently been introduced for the analysis of ILD: Dynamic Structural Equation Model, or DSEM.[3] DSEM has the potential to move well beyond the confines of conventional techniques. It provides a comprehensive framework for modeling and analyzing the reciprocal interplay of co-developing phenomena at both the individual and group level. Indeed, DSEM offers the prospect of testing research hypotheses in the behavioral sciences in ways not previously possible. DSEM's application in the myriad fields of social sciences has garnered significant attention. A recent study from McNeish and Hamaker[4] is an illustrative discussion on how DSEM can be employed in applied research. Yet, given DSEM's recent development and strong encouragement for use, almost nothing is known about the stability of its estimation and the replicability of its results. It is paramount that DSEM be subjected to rigorous scrutiny before its widespread adoption in applied research. The goal of our paper is to provide an initial examination of the stability and replicability of DSEM under conditions commonly encountered in applied research. We reanalyze previously published data across a range of initial conditions in order to demonstrate what aspects of the DSEM are and are not stable and replicable.

METHODS AND PROCEDURES

DSEM brings together elements of three well-established analytic methods: Structural Equation Modeling (SEM), multilevel modeling (MLM) and time series analysis.[3,4] Similar to MLM, DSEM accommodates nested structures in the data, where observations are clustered under higher level units (here, time nested within individuals). The model representing the higher-level unit is called the *between-person* model and the individual models nested under the higher-level unit is called the *within-person* model. DSEM uses time series analysis to model the autocorrelations in the within-person model for intensive longitudinal data. Moreover, incorporation of SEM allows DSEM to model the individual differences in the time-series parameters as latent variables. This integrative approach fills critical gaps left by individual methods. For instance, MLM handles interindividual variability but not latent variables. SEM deals with latent variables but lacks the flexibility for highly dense time measures and for random effects on model parameters. By bringing these methodologies together, DSEM emerges as a powerful tool capable of addressing the limitations of its constituent methods and offering a more comprehensive analytical framework for studying complex data structures.

DSEM provides both a powerful and flexible statistical framework, making it a valuable tool in the social sciences. A clinical psychologist interested in nicotine addiction and depression may hypothesize that depression and urge to smoke are entangled together in a codeveloping process.[5,6] Job stress and home stress may act as a catalyst to the situation by elevating an individual's level of depression and in turn entangling the urge to smoke and depression in a tighter yarn. McNeish and Hamaker[4] provide a guide to any researcher who wishes to test out such hypotheses using DSEM. The simulated dataset from the paper had measurements on depression, urge to smoke, home stress and job stress for 100 individuals across 50 discrete time points. This dataset is suitable to be analyzed using DSEM because it has a comparatively large sample of individuals, all measured over multiple time points. The presence of an observed variable with an autoregressive component (in case of urge to smoke) and without the autoregressive component (in case of depression) requires time series analysis. Proper treatment of the time invariant covariates of job and home stress necessitates the use of MLM. The latent variables used in the model draw heavily from the SEM literature. In short, the complexities of the dataset demand a flexible framework that attends to all these needs. Here we reanalyze the same data with our primary focus on the time-invariant covariate (TIC) DSEM as elucidated in McNeish & Hamaker.[4]

The path diagrams and representative equations[1-6] delineate the TIC DSEM. In this context, depression and urge to smoke are defined as time-varying covariates (or TVCs). For each individual, the urge to smoke at a time point (t) is dependent on their urge to smoke at the previous time point (t-1) and their depression at the same time point (t). This is represented in **Equation 1**. The random slopes at the *within-person* level resurface in the consequent **Equations 2-4**. At the *between-person* level, job stress and home stress affect the random slopes. Job and home stress are constant for each individual and therefore act as the TICs. The TICs are grand mean centered as opposed to Dep_i^b which is latent person-mean centered. In **Figure 1a**, the gamma (γ) variables represent the fixed effect of its predictor, and the u terms represent the random inter-subject effect. The tau (τ) variables stand for variances of the u parameters. For the physical description of each parameter, refer to **Table 1**.

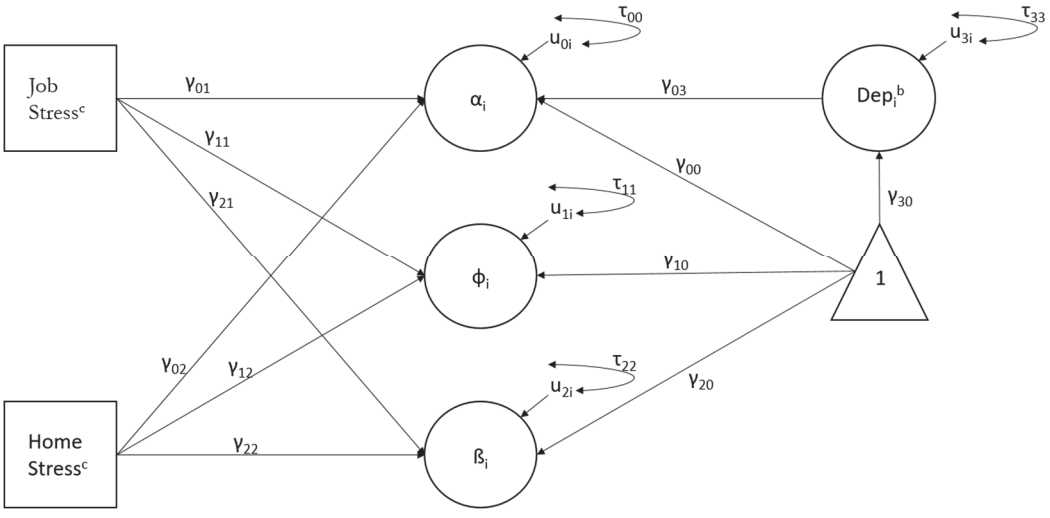

Figure 1a.

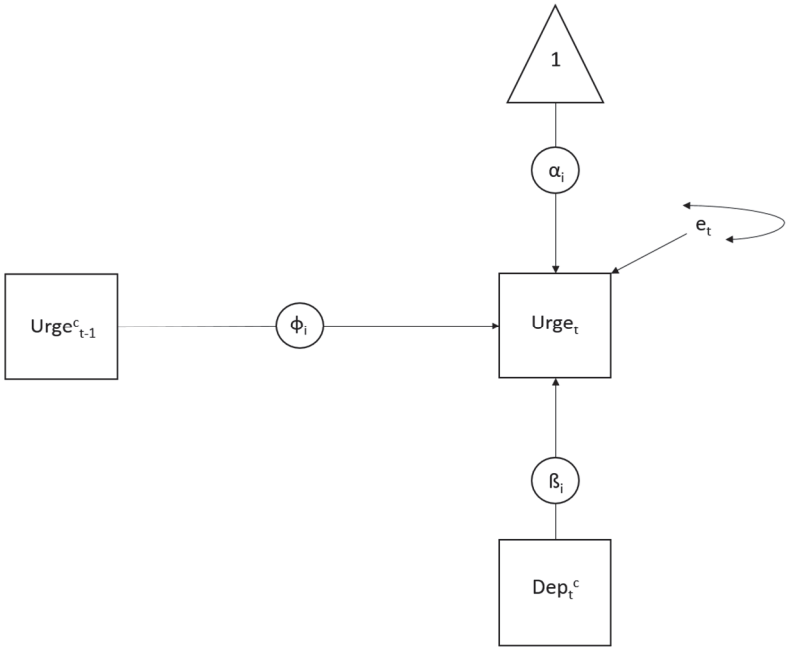

Figure 1b.

Figure 1. The figure displays the path diagrams of the between-person model (1a) and the within-person model (1b) for the TIC DSEM outlined in McNeish and Hamaker.[4] The rectangular boxes stand for variables that were directly measured and the circles represent latent variables estimated from the data. Note: path diagrams have been adapted from McNeish and Hamaker.[4]

$$Urge_{ti} = \alpha_i + \phi_i Urge^c_{(t-1)i} + \beta_i Dep^c_{ti} + e_{ti} \quad \textbf{Equation 1.}$$

$$\alpha_i = \gamma_{00} + \gamma_{01} JobStress^c_i + \gamma_{02} HomeStress^c_i + \gamma_{03} Dep^b_i + u_{0i} \quad \textbf{Equation 2.}$$

$$\phi_i = \gamma_{10} + \gamma_{11} JobStress^c_i + \gamma_{12} HomeStress^c_i + u_{1i} \quad \textbf{Equation 3.}$$

$$\beta_i = \gamma_{20} + \gamma_{21} JobStress^c_i + \gamma_{22} HomeStress^c_i + u_{2i} \quad \textbf{Equation 4.}$$

$$Dep^n_{ti} = Dep^c_{ti} + Dep^b_i \quad \textbf{Equation 5.}$$

$$Dep^b_i = \gamma_{30} + u_{3i} \quad \textbf{Equation 6.}$$

Including the latent person-mean centered Dep^c_{ti} in the within-person model and the latent person-mean Dep^b_i in the between-person model allows the complete disaggregation of the total effect of depression observed in the raw data (**Equation 5**). Latent person-mean centering approach treats Dep^b_i as an unknown quantity that has to be estimated and thus properly accounts for measurement error.[7] This inclusion allows researchers to discern the impact of unit change in a covariate on its outcome at a specific measurement occasion (the within-person effect). In this scenario, Dep^b_i influences α_i (**Equation 2**), the mean urge to smoke of an individual which also shows up at the within person level. Moreover, researchers can inspect how a one-unit change in the covariate mean across all measurement occasions affects the average of the outcome variable (the between-person effect). By including both effects in the model simultaneously, it is possible to investigate whether the within-person and between-person effects differ.

While DSEM is a versatile framework, it is quite complex and requires advanced methods for statistical estimation. The commercial software program Mplus,[8] the only package that currently implements DSEM, uses Bayesian estimation with a Gibbs sampler. The Gibbs sampling approach enables the estimation of parameters using conditional distributions, given that the conditional distributions are known and are easier to sample from compared to the unknown and often complex joint distribution.[9] It begins with an initial seed value for all the parameters. In the first iteration, it fixes the values of all parameters besides one and samples the value of the unfixed parameter from its conditional distribution. The estimator then selects a separate parameter, fixes the rest and samples from its conditional distribution. This process goes on until a new sample of values for all the parameters have been generated from their respective conditional distributions. The estimator continues sampling in multiple

iterations and the samples begin to approximate the joint distribution of the parameters. This technique is elegant because it uses the local dependencies and enables efficient exploration of high-dimensional spaces without having to directly sample from the often intricately complex joint distribution. In a model like DSEM with multiple parameters and a complex parameter space, Gibbs sampling is a logical option for estimation.

However, Gibbs sampling is not without its own set of challenges.[10,11] One key concern lies in its sensitivity to initial conditions, or the *seed values*. Since initial conditions guide its sampling of the posterior distribution, the first few samples can be biased, sometimes substantially so. Therefore, the initial samples are often discarded, a process referred to as the *burn-in*. In addition to burn-in, multiple sampling chains can be used to explore the space. The initial conditions assigned to the parameters might push the chains towards a local region of the distribution, making samples from the distribution seem biased. This underscores the importance of an extensive number of iterations to facilitate convergence.

To assess convergence, a commonly employed criterion is the Potential Scale Reduction Factor (PSRF)[12]. This criterion hinges on the idea of interchain variability, wherein a PSRF value equal to 1 means that the samples acquired from the chains are indistinguishable from each other. From a PSRF=1, it can be assumed that the chains are sampling from the same distribution—the true posterior distribution. McNeish and Hamaker[4] adhered to the default values within Mplus,[8] in which the first half of the samples were discarded as burn-in and two chains were used in exploration. The number of iterations was set to a maximum of 1000 and the convergence criteria called the Potential Scale Reduction Factor (PSRF) was set to the default value of 1.1.

While a PSRF value of 1 may take an unfeasibly long time to achieve, PSRF values close to 1 are achievable. However, this convergence criterion does not ensure that the samples obtained from the chains generate reproducible samples of the posterior distribution. It is possible that the stochastic algorithm becomes stranded in some region of the parameter space and requires a substantially longer period of time to converge to the true posterior distribution.[13] Moreover, PSRF is extremely dependent on the shape of the posterior. It has been observed that for heavily skewed posterior distributions PSRF does not converge to 1.0 even with increasing sample size.[14] Therefore, it is essential to use multiple random seeds or initial conditions to check if the chains are producing similar or the same parameter estimates. If, after a reasonably extensive list of initial conditions, each independent of the other, the parameter values rest stably at the same values, it can be said that the estimator is sampling from the true posterior distribution. Such assessment of parameter stability is critical while investigating model stability.

McNeish and Hamaker only used a single seed value and a single PSRF value for all of their analyses. It is unknown if their final solution was dependent on the initial conditions of chains and the default parameters of the convergence criteria. To assess the stability of the parameter estimates and their dependence on initial conditions, we first randomly sampled 1000 seed values (without replacement) from integers ranging from 1 to 100000 using R (version 4.2.2).[15] Next, we estimated 1000 separate DSEMs, each using one of the randomly sampled seeds. The models were run on Mplus using MplusAutomation[16] (version 1.1.0) in R. To further probe the stability of these estimates, we introduced variations in the estimator such that the maximum number of iterations allowed was systematically altered (values of 1000, 10000 and 30000) as well as the value of the PSRF (values of 1.1, 1.05, 1.01 and 1.005). These manipulations of the convergence criteria were aimed to provide a more nuanced understanding of the impact of such criteria and initial conditions in the estimation of DSEM. The parameter estimates, defined by default as the median of the parameter's posterior distribution, were recorded across the 1000 initial conditions encompassing the variations in the PSRF thresholds and iteration counts. The median values of each of the parameters for the 1000 different seed values were visualized through boxplots for each of the conditions. The analysis was done in R[15] and the visualizations were created using ggplot2 (version 3.4.4).[17] To reduce redundancy, for some of the conditions we report the mean, median, standard deviations and quartiles of the parameters instead of plotting the boxplots.

Besides the parameter estimates, a record was maintained concerning the inclusion of 0 in each parameter's credible interval in its posterior distribution. In Bayesian statistics, the posterior distribution of the parameters is used to discuss its estimate and precision. However, frequentists using Gibbs sampling to estimate DSEM might use the presence or absence of 0 in the credible intervals of the parameter distribution for significance testing. It should be noted that this is not a true null hypothesis significance test, as the concept of null hypothesis testing does not fit well in the pure Bayesian perspective. However, as seen in McNeish and Hamaker[4], researchers use Bayesian credible intervals for null hypothesis testing. We do not wish to deviate from the analysis perspective used in McNeish and Hamaker[4], therefore we use the inclusion of 0 in the Bayesian credible intervals as a test of significance. The following discussion will include both the stability of the parameter estimates and the significance of the parameters to foster a more holistic insight into the problem at hand.

RESULTS

We present the findings from the TIC DSEM analysis, as discussed in McNeish and Hamaker[4] in **Table 1**.

Effect	Notation	Posterior Median	95% Credible Interval
Intercept (Alpha): overall intercept capturing the baseline urge to smoke	γ_{00}	.06	[-.14,.30]
Intercept (Phi): overall autoregressive intercept capturing the baseline autoregressive effect	γ_{10}	.19	[.16,.22]
Intercept (Beta): overall effect of depression on the urge to smoke capturing the baseline effect	γ_{20}	.79	[.62,.95]
Intercept (Dep): overall effect capturing the baseline influence of depression on the urge to smoke	γ_{30}	.02	[-.01,.05]
Alpha on Job Stress: effect of job stress on the baseline urge to smoke	γ_{01}	.50	[.35,.65]
Alpha on Home Stress: effect of home stress on the baseline urge to smoke	γ_{02}	.33	[.19,.46]
Alpha on Dep: effect of the person mean of depression on the baseline urge to smoke	γ_{03}	-2.35	[-11.22,4.35]
Phi on Job Stress: effect of job stress on the autoregressive effect	γ_{11}	.12	[.09,.15]
Phi on Home Stress: effect of home stress on the autoregressive effect	γ_{12}	.06	[.03,.08]
Beta on Job Stress: effect of job stress on slope of depression on urge	γ_{21}	.29	[.11,.48]
Beta on Home Stress: effect of home stress on slope of depression on urge	γ_{22}	.35	[.17,.51]
Var. (Alpha): individual specific baseline urge to smoke	τ_{00}	.34	[.16,.49]
Var. (Phi): variance of the autoregression	τ_{11}	.01	[.00,.01]
Var. (Beta): variance of influence of depression on urge	τ_{22}	.64	[.47,.88]
Var. (Dep): variance of depression	τ_{33}	.01	[.00,.01]
Res. Var. (Urge): residual variance of urge to smoke	σ^2	1.14	[1.09,1.18]

Table 1. Estimates and 95% Credible Intervals for the TIC DSEM. Note: The data have been taken from McNeish and Hamaker[4] and these are estimates from one seed value with 1000 being the maximum number of iterations allowed for the estimator.

The analysis reveals significant time-invariant effects in the model. Covariates for the intercept predict that stresses in the work and home environment increase the baseline Urge to Smoke for individuals. Specifically, a one-unit increase in Job Stress leads to a .50 unit increase in the mean Urge to Smoke while one-unit increase in Home Stress leads to a .33 unit increase in the same. The estimates for γ_{11} and γ_{12} are non-null and indicate that Home Stress and Job Stress strengthen the carryover effects of Urge to Smoke by .06 and .12, respectively. γ_{21} and γ_{22} are significant as well. These parameters provide evidence that increases in Job Stress and Home Stress is similarly predictive of a stronger effect of Depression on Urge to Smoke. γ_{21} is estimated to be .29, which means a one—unit increase in Job Stress increases β_i, or the slope of depression on urge, by .29. Similarly, γ_{22} estimated to .35 means a one unit change in Home Stress increases the slope of depression on urge by .35 units. Notably, the person mean of Depression does not appear to have any significant effect on the person mean of Urge to Smoke as γ_{03} has 0 in its 95% credible interval. This lack of a significant effect can be interesting to clinical psychologists because it implies that treating only depression will not have any effect on smoking tendencies and vice versa. This necessitates targeted interventions for both depression and smoking behavior. However, before clinicians begin developing novel interventions, it is imperative to check the robustness of these estimates to save time, effort and financial resources. Thus, we tested if McNeish and Hamaker's reported estimates were a local solution using different seed values for the estimation.

Since the original paper had employed 1000 iterations as the convergence criterion, we began with the same. However, instead of using just one seed value, we initiated the sampler with 1000 *different* seeds. To our concern, a substantial subset of the seeds failed to converge. The fact that no valid results could be obtained from these instances was, in itself, a disconcerting finding. Therefore, we shifted our focus to the first 100 seeds that exhibited convergence. Figure 2 shows the estimates of the model parameters for the 100 initializations. Notably, the boxplots for most of the parameters show minimal variability indicating no sensitivity to the initial conditions of the sampler, affirming DSEM's reliability in most instances. However, a concerning degree of variability was observed in γ_{03}, the effect of the person mean of depression on the baseline urge to smoke, exhibiting marked sensitivity to the initial conditions of the sampler. While γ_{30}, or the mean value of depression, is not as variable as γ_{03}, three seeds among the 100 emerged as statistically significant. These findings indicate that key substantive conclusions from the TIC DSEM could be fully

dictated by the arbitrary seed value chosen for the analysis. The challenges of non-convergence and the parameter instability underscore the need for a higher number of iterations in the Gibbs sampler to achieve more stable estimates.

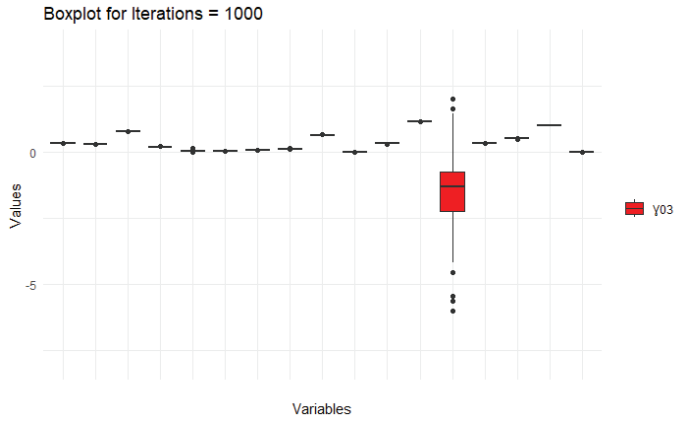

Figure 2. The figure displays boxplots for all TIC DSEM parameters based on 100 seeds, each subject to a 1000-iteration convergence criterion. γ_{03} clearly has the highest variability of all the parameters.

We extended the number of iterations substantially beyond the original recommendation and employed a convergence criterion of 10000 iterations. Under this criterion, all the seeds converged but the stability of parameter γ_{03} remained elusive. As seen in Figure 3, γ_{03} in this condition shows more variability than any of the other parameters. Moreover, 47 out of 1000 seeds yielded significant results for the mean value of depression (γ_{30}). This prompted an exploration of the convergence criterion of 30000 maximum iterations. The results persisted. γ_{03} showed no change in standard deviation, as can be seen in **Table 3**. γ_{30} had an equal number of significant seeds. These results concern us about potentially misleading significant results. If a random seed value can produce significance, researchers need to be cautious while interpreting results from DSEM.

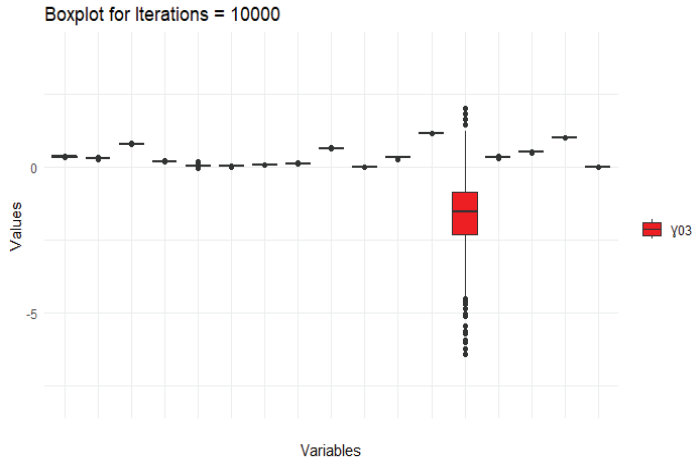

Figure 3. The figure displays boxplots for all TIC DSEM parameters based on 1000 seeds, each subject to a 10000-iteration convergence criterion. γ_{03} clearly has the highest variability of all the parameters. Note: This figure differs from Figure 2 on the crucial fact that for iterations set to 10000, all the models converged successfully, unlike iterations set to 1000. Therefore, this figure represents estimates from 1000 models as opposed to 100 from Figure 2.

Upon a thorough investigation into the convergence threshold used in the sampling algorithm, it became evident that the estimator relies on the similarity of samples from the posterior distribution. Using a higher number of iterations might result in the chains exploring the distribution better, but it does not ensure that the samples from the different chains will emerge similar to each other. Therefore, we employed a convergence threshold based on interchain variability called the Potential Scale Reduction Factor (PSRF). As described earlier, this criterion ensures that the estimator continues running until a predetermined value is attained.

The PSRF criterion is based on Gelman and Rubin's seminal paper of 1992, where they recommended PSRF values less than or equal to 1.1 as an indicator of convergence. Consequently, Asparouhov and Muthén[18] suggested that PSRF values between 1.1

and 1.05 can render samples from chains virtually indistinguishable for most models. Most studies have followed these recommendations and use PSRF values between 1.1 and 1.05. Thus, we set the PSRF convergence criterion to values of 1.1 and 1.05. Under both convergence criteria, γ_{03} exhibited variability. As shown in Figure 4, PSRF≤1.1 produced results similar to the iterations criteria. The critical difference lay in the fact that with stricter PSRF values one could significantly reduce the variability in the unstable parameter. Notably, the variability was significantly reduced when using PSRF≤1.05 compared to PSRF≤1.1 (see **Table 2**). The number of significant seeds for γ_{30} was 82 under PSRF≤1.1, while it reduced to 47 under PSRF≤1.05. It may be inferred that for models as complex as DSEM, lower convergence thresholds like 1.05 are better than the traditionally recommended value of 1.1. However, the result produced by PSRF≤1.05 was identical to those yielded by the convergence criteria based on the number of maximum iterations. In order to improve the stability of the parameter estimates, lower convergence thresholds need to be used.

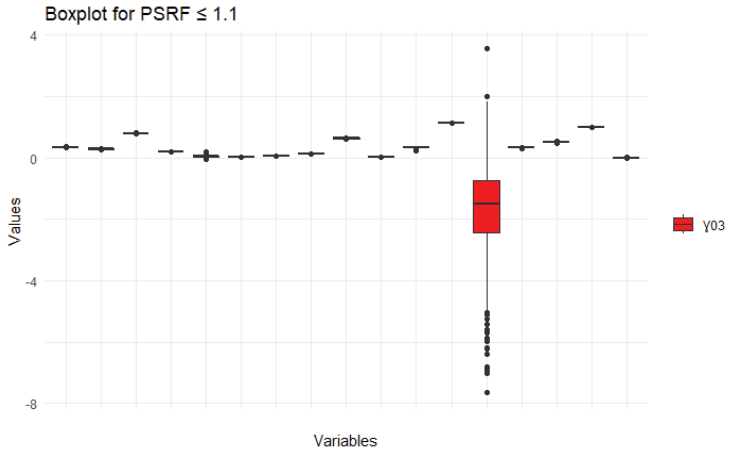

Figure 4. The figure displays boxplots for all TIC DSEM parameters based on 1000 seeds, each subject to a PSRF≤1.1 convergence criterion. γ_{03} clearly has the highest variability of all the parameters.

Lower convergence thresholds necessitate a departure from the standard values used with the PSRF criterion. Specifically, we tested thresholds as low as 1.01 and 1.005 to check if these bring stability to the parameter estimates. These values yielded a significant decrease in variability in the γ_{03} parameter as can be seen in **Table 2**. The number of significant seeds for γ_{30} dropped to 10 for PSRF≤1.01 and to only four seeds for PSRF≤1.005. The boxplots for the γ_{03} parameter under the seven different convergence criteria are shown in Figure 5 and it is apparent that the variability changes as a function of the convergence criterion. **Table 3** shows mean, median, standard deviation and the 25th and 75th percentiles of γ_{03} parameter for each condition of convergence. It reinforces the inference that the stricter PSRF criteria reduce the standard deviation of the unstable parameter. As noted earlier, PSRF≤1.05, iterations = 10000 and iterations = 30000 produce identical results. However, PSRF≤1.01 and PSRF≤1.005 produce significant reduction in variability. These results for the significance testing using pairwise Bartlett tests[19] with Benjamini—Hochberg correction[20] are tabulated in **Table 2**. Although the PSRF≤1.005 produces estimates with significantly the least amount of variability, it should be noted that computational time was a greater challenge in the case for PSRF≤1.005 as each seed took about 50 seconds to converge with the 1000 seeds taking almost 14 hours. However, each seed took approximately 18 seconds to converge under PSRF≤1.01 which added up to 5 hours for the 1000 models. The substantial investment in time prompts a careful consideration on the part of the researcher while weighing the trade-offs of each convergence threshold. For this model, considering the modest differences between 1.01 and 1.005 thresholds while looking at the mean, median and variability values, it would be wise to stick with 1.01 as the best convergence threshold.

The above analysis sheds important light on some critical characteristics of the DSEM framework. While a majority of its parameter estimates demonstrate robustness to variations in estimator initializations, certain parameters are highly sensitive to seed values. This sensitivity impacts the estimates of certain parameters and the credible intervals of some others. For any researcher following the standard practice of using singular seed values to initialize parameters, such sensitivity in estimates would remain unknown and drastically influence the interpretations of results. To mitigate this sensitivity, we systematically altered the convergence criteria and were able to increase robustness in the parameter estimates. To ascertain statistical significance of the reduction in variability of the parameter γ_{03}, we employed Bartlett's test for equality of variances.[19] To account for multiple comparisons, we used Benjamini—Hochberg false discovery rate correction.[20]

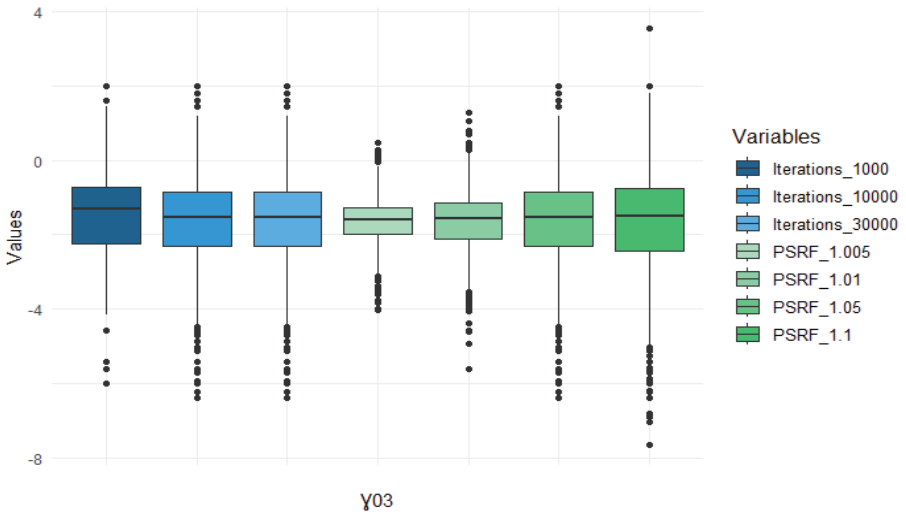

Figure 5. The figure displays boxplots for the γ_{03} parameter under all the 7 convergence criteria. The variability of the parameter changes as the convergence criteria changes with the least amount of variability being observed for PSRF≤1.005. Note: Iterations=1000 uses 100 models and their parameter estimates because many of the models failed to converge. All the other conditions allowed all the models to converge, therefore the other boxplots are made from estimates of 1000 models.

Group 1	Group 2	p value	p adjusted	sig
PSRF 1.1	PSRF 1.05	.00	.00	*
PSRF 1.1	PSRF 1.01	.00	.00	*
PSRF 1.1	PSRF 1.005	.00	.00	*
PSRF 1.1	Iterations 10000	.00	.00	*
PSRF 1.1	Iterations 30000	.00	.00	*
PSRF 1.05	PSRF 1.01	.00	.00	*
PSRF 1.05	PSRF 1.005	.00	.00	*
PSRF 1.05	Iterations 10000	1.00	1.00	
PSRF 1.05	Iterations 30000	1.00	1.00	
PSRF 1.01	PSRF 1.005	.00	.00	*
PSRF 1.01	Iterations 10000	.00	.00	*
PSRF 1.01	Iterations 30000	.00	.00	*
PSRF 1.005	Iterations 10000	.00	.00	*
PSRF 1.005	Iterations 30000	.00	.00	*
Iterations 10000	Iterations 30000	1.00	1.00	

Table 2. Pairwise Bartlett tests of the 6 groups with B—H FDR correction. Note: Iterations=1000 has not been included in this analysis due to the number of seeds that failed convergence.

PSRF 1.1	-1.66	-1.50	1.37	-2.45	-0.74
PSRF 1.05	-1.63	-1.53	1.15	-2.31	-0.86
PSRF 1.01	-1.65	-1.58	0.79	-2.10	-1.15
PSRF 1.005	-1.65	-1.61	0.60	-2.00	-1.26
Iterations 1000	-1.48	-1.32	1.36	-2.23	-0.73
Iterations 10000	-1.63	-1.53	1.15	-2.31	-0.86
Iterations 30000	-1.63	-1.53	1.15	-2.31	-0.86

Table 3. Mean, median, standard deviation, 25th and 75th percentiles for the estimate of γ_{03} parameter from 1000 models. For each condition, 1000 models with the same architecture but different initializations produced a posterior distribution for the γ_{03} parameter. The median of the posterior was treated as the estimate. Note: Iterations=1000 used 100 different seeds therefore the statistics come from the parameter estimates of those 100 models.

DISCUSSION

In this study, we explored the Dynamic Structural Equation Modeling (DSEM) framework with a focus towards the robustness of its parameter estimates. Using the TIC DSEM employed in McNeish and Hamaker,[4] we discussed the various capabilities of the framework. As was shown in the aforementioned paper, DSEM has the ability to study time-invariant covariates in intensive longitudinal data. The model allowed an extensive analysis of how stresses in the work and home environment can have severe impacts on the carryover effects of the urge to smoke and in strengthening the relationship between depression and smoking tendencies. Using DSEM researchers can study such behavior better and thereby improve intervention programs.

However, our analyses unveiled a critical insight – not all of the model's parameter estimates exhibited stability as a function of initial conditions. We used 1000 different seed values to initialize the Gibbs estimator. Across different initial conditions of the Gibbs sampler, a majority of the parameter estimates did not display variability. A few parameters proved to be exceedingly sensitive to seed values. The sensitivity affected the estimated values of certain parameters and the credible intervals of others. For the simulated dataset used in this study, the instabilities arose from γ_{03}, the effect of the person mean of depression on the baseline urge to smoke, and γ_{30}, or the mean value of depression. Unfortunately, it is difficult to make any inference as to why these parameters specifically show fluctuations. It is suspicious that both the variables are associated with the latent mean centered Dep, which may be the unstable component of the model. Here we can only ponder, but future research would do good to shed light on this issue. Although the underlying reasons are unclear, it should be acknowledged that this finding is disconcerting as model results are assumed to be stable across arbitrary initial conditions. Such lack of robustness can have cascading effects on the interpretation of results in academic papers. Yet all current recommendations for the widespread use of DSEM in practice do not address this issue. Therefore, it is imperative that applied researchers be aware of these issues before deciding to use DSEM in their research endeavors. Moreover, methods to reduce these instabilities need to be investigated.

To tackle the challenge of sensitivity to initial conditions of the Gibbs sampler, we systematically adjusted the convergence criterion. While McNeish and Hamaker[4] used an upper bound on the number of iterations as the convergence criterion, our investigation revealed that transitioning to PSRF thresholds results in diminished parameter variability. Going beyond the traditional recommendations of PSRF thresholds between 1.1 and 1.05, we employed stricter convergence thresholds of 1.01 and 1.005. To assess statistical significance of the reduced variability in the parameter we used Bartlett's test for the equality of variances. Benjamini—Hochberg false discovery rate correction was used to address multiple comparisons. PSRF≤1.01 and PSRF≤1.005 significantly reduced variability in the parameter estimate in comparison to PSRF≤1.05, as well as when the number of iterations was set to 10000 and 30000.

For a model as complex as TIC DSEM, which involves sampling from a complex posterior distribution, it is crucial to allow sufficient time for accurate exploration of the space. While more exploration of this complicated parameter space will allow more robust estimates, there is a trade-off between computational time and accuracy. Opting for a stringent threshold like 1.005 demands three times more computational time for convergence than a comparatively lenient threshold like 1.01. While the more stringent threshold will provide more robust estimates, the substantial increase in computational time makes it a less optimal choice. Consequently, we conclude that for the dataset at hand, a threshold value of 1.01 provides the best balance between robustness and computational efficiency. As DSEM gains traction in applied research, future investigators need to be mindful of these considerations and select the threshold that aligns best with their data.

CONCLUSION

This study explored Dynamic Structural Equation Models' robustness in parameter estimation, emphasizing its utility in studying time-invariant covariates within intensive longitudinal data. While DSEM provides a powerful tool for researchers to investigate complex behavioral dynamics, our paper reveals issues in the modeling framework that need to be accounted for. This was an initial investigation into the stability and replicability of DSEM on a simulated dataset. We discovered that a small subset of parameters is highly sensitive to initial conditions of the estimator. To address this lack in robustness and avoid local solutions, we propose transitioning to more stringent convergence thresholds. However, this comes at the cost of computational time. For the dataset used in this paper, a threshold of 1.01 is recommended to strike the best balance between robustness and efficiency. This underscores the necessity of researchers to be aware of the various challenges inherent to this novel framework.

While we have highlighted certain concerns within the DSEM framework, it needs to be emphasized that DSEM has immense potential in refining our understanding of dynamic human behavior as it unravels over time. Compared to other models in the literature like Autoregressive Cross Lagged (ARCL), DSEM does a better job in modeling co-developmental trajectories of psychological phenomena.[3,21] We encourage researchers to use DSEM in their endeavors, but with a recommendation to adhere to best practices that help navigate the associated challenges. In light of our analyses, we propose prioritizing the PSRF convergence criterion over criteria involving the number of iterations of the Gibbs sampler. Additionally, smaller PSRF values ($\leq$ 1.01) can yield more stable estimates, but the computational efficiency needs to be considered. Above all, we stress the importance of examining DSEM's estimates carefully before drawing conclusions. One may use different seed values for the estimation to test the robustness of the parameter estimates. However, there may be other and more effective ways of validating the results.

This study serves as an initial foray into the DSEM framework and is not meant to be comprehensive. There are a multitude of directions for future work some of which we mention here. We used a simulated dataset where the generating model was known. We acknowledge that our findings may be specific to this dataset, where the generating model was known. Future research would do well to explore simulations with different generating models to investigate if such issues consistently arise across datasets. Real world datasets, where the underlying model is unknown and complex, may present new challenges. Investigating if and how parameter instabilities show up under these real-world situations could yield valuable insights into the practical applications of DSEM. While this study unearthed the presence of instabilities in certain parameter estimates, we were not able to shed light on their underlying reasons. Simulations attempting to resolve these mysteries are highly recommended. Another promising direction of work lies in the exploration of priors for DSEM. We employed non-informative priors for our study. Future investigations could explore the sensitivity of parameters to mildly informative priors, as such adjustments might influence the posterior distribution, mitigate instabilities or improve computational efficiency. These are a few of the many avenues where more work needs to be done. Additionally, this study raised questions about the impact of Bayesian estimation in Mplus for the implementation of DSEM. Alternative software packages like JAGS/STAN should be considered to fit these models and test their robustness. At this point in time, DSEM can only be implemented in Mplus, opening up myriad opportunities for methodologists to build packages aimed at fitting these models. This is an ongoing quest for improvement that aims to empower researchers with a more robust and versatile tool to study human psychology. We hope that this study is just the first of many papers aimed at enhancing the DSEM framework.

ACKNOWLEDGEMENTS

The authors thank Dr. Daniel McNeish for providing access to the data from his paper which proved invaluable to this project. The guidance and support extended by the professors and graduate students at the L.L. Thurstone Psychometric Laboratory are sincerely appreciated. The authors express their gratitude towards UNC's Department of Psychology and Neuroscience for providing the financial support in acquiring the Mplus software license.

REFERENCES

1. Dunton, G. F., Rothman, A. J., Leventhal, A. M., & Intille, S. (2019) How intensive longitudinal data can stimulate advances in health behavior maintenance theories and interventions. *Translational Behavioral Medicine* 11(1), 281–286. *https://doi.org/10.1093/tbm/ibz165*
2. Curran, P. J., & Hancock, G. R. (2021) The Challenge of Modeling Co-Developmental Processes over Time. *Child Development Perspectives* 15(2), 67–75. *https://doi.org/10.1111/cdep.12401*
3. Asparouhov, T., Hamaker, E. L., & Muthén, B. (2017) Dynamic Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal* 25(3), 359–388. *https://doi.org/10.1080/10705511.2017.1406803*
4. McNeish, D., & Hamaker, E. L. (2020) A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods* 25(5), 610–635. *https://doi.org/10.1037/met0000250*
5. Roys, M., Weed, K., Carrigan, M., & MacKillop, J. (2016) Associations between nicotine dependence, anhedonia, urgency and smoking motives. *Addictive Behaviors* 62, 145–151. *https://doi.org/10.1016/j.addbeh.2016.06.002*

6. Dierker, L., Rose, J., Selya, A., Piasecki, T. M., Hedeker, D., & Mermelstein, R. (2015) Depression and nicotine dependence from adolescence to young adulthood. *Addictive Behaviors* 41, 124–128. *https://doi.org/10.1016/j.addbeh.2014.10.004*

7. Asparouhov, T., & Muthén, B. (2018) Latent Variable Centering of Predictors and Mediators in Multilevel and Time-Series Models. *Structural Equation Modeling: A Multidisciplinary Journal* 26(1), 119–142. *https://doi.org/10.1080/10705511.2018.1511375*

8. Muthén, B. O., Muthén, L. K., & Angeles. (1998-2017) *Mplus User's Guide. Eighth Edition* (Muthén & Muthén) 8th ed. Los Angeles, CA.

9. Dou, L., & Hodgson, R. J. W. (1995) Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation. I. *Inverse Problems* 11(5), 1069–1085. *https://doi.org/10.1088/0266-5611/11/5/011*

10. Christopher De Sa, Kunle Olukotun, & Ré, C. (2016) Ensuring Rapid Mixing and Low Bias for Asynchronous Gibbs Sampling. *PubMed* 48, 1567–1576.

11. Jensen, C. S., Kjærulff, U., & Kong, A. (1995) Blocking Gibbs sampling in very large probabilistic expert systems. *International Journal of Human-Computer Studies* 42(6), 647–666. *https://doi.org/10.1006/ijhc.1995.1029*

12. Gelman, A., & Rubin, D. B. (1992) Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 7(4), 457–472. *https://doi.org/10.1214/ss/1177011136*

13. Whidden, C., & Matsen, F. A. (2015) Quantifying MCMC Exploration of Phylogenetic Tree Space. *Systematic Biology* 64(3), 472–491. *https://doi.org/10.1093/sysbio/syv006*

14. Fabreti, L. G., & Höhna, S. (2021) Convergence assessment for Bayesian phylogenetic analysis using MCMC simulation. *Methods in Ecology and Evolution* 13, 77-90. *https://doi.org/10.1111/2041-210x.13727*

15. R Core Team. (2022) R: The R Project for Statistical Computing. R-Project.org. *https://www.r-project.org/* (accessed September 2023)

16. Hallquist, M. N., & Wiley, J. F. (2018) MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal* 25(4), 621–638. *https://doi.org/10.1080/10705511.2017.1402334*

17. Wickham, H. (2016) Create Elegant Data Visualisations Using the Grammar of Graphics. Tidyverse.org. *https://ggplot2.tidyverse.org* (accessed September 2023)

18. Asparouhov, T. and Muthén, B. (2010) Bayesian Analysis Using Mplus: Technical Implementation. Mplus Technical Report. *http://www.statmodel.com* (accessed September 2023)

19. Bartlett, M. S. (1937) Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series a - Mathematical and Physical Sciences* 160(901), 268–282. *https://doi.org/10.1098/rspa.1937.0109*

20. Benjamini, Y., & Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1), 289–300. *https://doi.org/10.1111/j.2517-6161.1995.tb02031.x*

21. Zhou, L., Wang, M., & Zhang, Z. (2021) Intensive Longitudinal Data Analyses With Dynamic Structural Equation Modeling. *Organizational Research Methods* 24(2), 219–250. *https://doi.org/10.1177/1094428119833164*

ABOUT STUDENT AUTHOR

Suryadyuti Baral is currently a senior at UNC - Chapel Hill. He is pursuing a double major in Psychology and Statistics. After graduation he wishes to pursue a PhD in quantitative or computational psychology.

PRESS SUMMARY

Dynamic Structural Equation Models (DSEM) have been widely advertised as a powerful and versatile modeling technique that can shed light on enduring inquiries in the field of psychology. However, our work has unearthed some disconcerting issues in the modeling framework. Turns out DSEM produces estimates that are very unstable and sensitive to arbitrary initializations of its estimation procedure. Its erroneous results can mislead applied researchers to form wrong conclusions from their data. Therefore, it is pivotal that researchers are not left uninformed of these concerns. The study exposes these shortcomings and offers directions for best practice to navigate the deficiencies for anyone who wishes to use DSEM in their work.