Structural Equation Modeling: A Multidisciplinary Journal, 25: 495–515, 2018

ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2017.1392862

Number of Subjects and Time Points Needed for Multilevel Time-Series Analysis: A Simulation Study of Dynamic Structural Equation Modeling

Mårten Schultzberg[1] and Bengt Muthén[2]
[1]*Uppsala University*
[2]*Muthén and Muthén, Los Angeles, CA*

Dynamic structural equation modeling (DSEM) is a novel, intensive longitudinal data (ILD) analysis framework. DSEM models intraindividual changes over time on Level 1 and allows the parameters of these processes to vary across individuals on Level 2 using random effects. DSEM merges time series, structural equation, multilevel, and time-varying effects models. Despite the well-known properties of these analysis areas by themselves, it is unclear how their sample size requirements and recommendations transfer to the DSEM framework. This article presents the results of a simulation study that examines the estimation quality of univariate 2-level autoregressive models of order 1, AR(1), using Bayesian analysis in M*plus* Version 8. Three features are varied in the simulations: complexity of the model, number of subjects, and number of time points per subject. Samples with many subjects and few time points are shown to perform substantially better than samples with few subjects and many time points.

Keywords: between-level mediation model, DSEM, power, sample size

Dynamic structural equation modeling (DSEM) provides new methods for analyzing intensive longitudinal data (ILD) such as ecological momentary assessments, experience sampling methods, and ambulatory assessments (Asparouhov, Hamaker, & Muthén, 2017a, 2017b). DSEM uses two-level modeling with time on Level 1 and individuals on Level 2. It models intraindividual changes over time and allows the parameters of these processes to vary across individuals using random effects. There are three key random effects of interest in psychological research of longitudinal data: random means, random autocorrelations, and random variances. As an extension of conventional multilevel modeling, DSEM allows random effects to be not only dependent variables regressed on Level 2 covariates but also predictors of various outcomes. The flexibility of DSEM is made possible using Bayesian estimation.

Correspondence should be addressed to Mårten Schultzberg, Department of Statistics, Uppsala University, Uppsala 751 05, Sweden. Email: marten.schultzberg@statistik.uu.se

Color versions of one or more of the figures in this article can be found online at www.tandfonline.com/HSEM

Hamaker, Asparouhov, Brose, Schmiedek, and Muthén (2017) illustrated different models and research questions that can be investigated within the DSEM framework using repeated measures of affect. As pointed out by the authors, many of these models are novel and need to be studied further. In this article, the sample size needed for good estimation properties is studied. In two-level time-series analysis, two aspects of sample size need to be considered when discussing finite sample properties: the number of individuals (N) and the number of time points (T) per individual. Different combinations of N and T are seen in different ILD application areas.

The following examples highlight the large variation in both N and T covered by ILD studies. In Hamaker et al. (2017), the data are of a daily diary type with 100 individuals measured for 100 days in a row (i.e., $N = 100$ and $T = 100$). In McAdams and Constantian (1983), intimacy and affiliation of 50 individuals are measured at seven random time points a day for 7 days (i.e., $N = 50$, $T = 49$). In Bolger and Schilling (1991), neuroticism and exposure to daily stressors are measured for 339 individuals once a day for 6 weeks (i.e., $N = 339$, $T = 42$). In the Shiffman and Waters (2004) study of smoking behavior and relapse, data of an ecological momentary assessment

(EMA) type are collected where $N = 215$ and $T = 100$. L. H. Cohen et al. (2008) studied effects of psychotherapy with a sample of $N = 62$ and $T = 7$. In Trull et al. (2008), 60 individuals ($N = 60$) are measured six times per day for 4 weeks ($T = 6 \times 28 = 168$) to investigate affect instability in borderline personality disorder. In Jongerling, Laurenceau, and Hamaker (2015), the multilevel autoregressive model is discussed and exemplified with a daily diary data set of women's positive affect, where 89 women ($N = 89$) completed daily diaries for 42 consecutive days ($T = 42$). For more examples of longitudinal studies with different kinds of data see, for example, Bolger and Laurenceau (2013).

This article aims to give insights and guidelines regarding the data requirements of some of the two-level time-series models available in M*plus* Version 8 (Muthén & Muthén, 2017) through the DSEM framework of Asparouhov et al. (2017b). To investigate the small sample properties of the estimation of these models, a Monte Carlo simulation study is conducted. The main focus of the simulation study is the N and T requirements for different combinations of the two. Which features of the models require large N and T, respectively, and to what extent can a large N compensate for a small T and vice versa?

There have been several previous Monte Carlo simulation studies on the topic of two-level models. Yuan and MacKinnon (2009) made a case for Bayesian estimation as compared to classical maximum likelihood (ML) estimation in a simulation study of one- and two-level mediation analysis. However, that work considers individuals on Level 1 as opposed to Level 2, which is considered in this article. Krone, Albers, and Timmerman (2016) found similar performance of Bayesian Markov Chain Monte Carlo (MCMC) and ML estimation of two-level AR(1) models in a small simulation study. The article considers the four combinations of $N = 10, 25$ and $T = 10, 25$. Jongerling et al. (2015) carried out a simulation study comparing different estimation methods for a multilevel autoregressive model. The methods considered are different kinds of ML and Bayesian estimation. The sample sizes considered are $N = 20, 50, 100$ and $T = 10, 20, 50$. The Jongerling et al. (2015) article is highly relevant for this study. First, most features pointed out by the authors as missing in common multilevel software (e.g., random innovation variance and multivariate models on the within level) are implemented in M*plus* Version 8. Second, their simulation study covers one of the nine models covered in this article. The overlap between the studies is limited, however, because the focus of this article is the estimation of DSEM models in M*plus* version 8, which is not included in Jongerling et al. (2015). To the best of our knowledge, the estimation performance of the usage of random coefficients for autoregressive models of order 1 (AR(1)) as mediators (i.e., both independent and dependent variables on the between level) has not been investigated.

The rest of this article is structured as follows. The next section gives an outline and setup of the Monte Carlo simulation study and the models considered therein. The section following that contains the results from the Monte Carlo study. After this, caveats and important considerations when simulating DSEM are identified. Finally, the last section contains a discussion with guidelines and concluding remarks.

MONTE CARLO SETUP

In this section the Monte Carlo simulation study is motivated and described in detail. First, the sequence of models considered is defined. The effect sizes and R^2 of these models are then discussed. Finally, the evaluation measures of the simulation study are defined.

Model Sequence

The models considered in the simulation study are based on Hamaker et al. (2017). The intent is to investigate key parts of these models. Figure 1 displays the model diagrams of the nine model variations considered in the Monte Carlo simulation. The figures follow path analysis convention and are consistent with the notation in the *M*plus *User's Guide* (Muthén & Muthén, 2017). A rectangle is an observed variable and a circle is a latent variable. A single-headed arrow from A to B means that A affects B and corresponds to a regression slope. A single-headed arrow that does not start at an observed or latent variable is a residual with its corresponding variance. A filled small circle is a random coefficient. A filled circle on the middle of a line is a random slope. A filled circle at the end of a single-headed arrow is a random intercept. A single-headed arrow starting in a filled circle is a random residual variance.

In Models 1 through 3 (Figure 1a–**c**), the three parameters of interest are allowed to be random. Starting with only a random mean in Model 1, Model 2 adds a random autoregressive coefficient, and Model 3 adds a random residual variance. In Models 4 through 6 (Figure 1d–**f**), the same pattern is repeated but here the random coefficients are regressed on the added between-level covariate **W**. Finally, in Models 7 through 9 (Figure 1g–**i**), a dependent variable Z is added and regressed on the random coefficients and **W**. This sequence gives guidelines for the rather different settings in which the random coefficients are random, random and used as dependent variables, or random and used both as dependent and independent variables.

All models considered in the study are nested in the model displayed in Equations 1 and 2, corresponding to Level 1 and Level 2, respectively,

$$Y_{it} = \varphi_i Y_{w,it-1} + \varepsilon_{it}, \tag{1}$$

where the predictor $Y_{w,it-1}$ is latent variable-centered as $Y_{w,it-1} = Y_{it-1} - \alpha_i$, $\varepsilon_{it} \sim N(0, \sigma_i^2)$ for all $t = 1, \ldots, T$, and

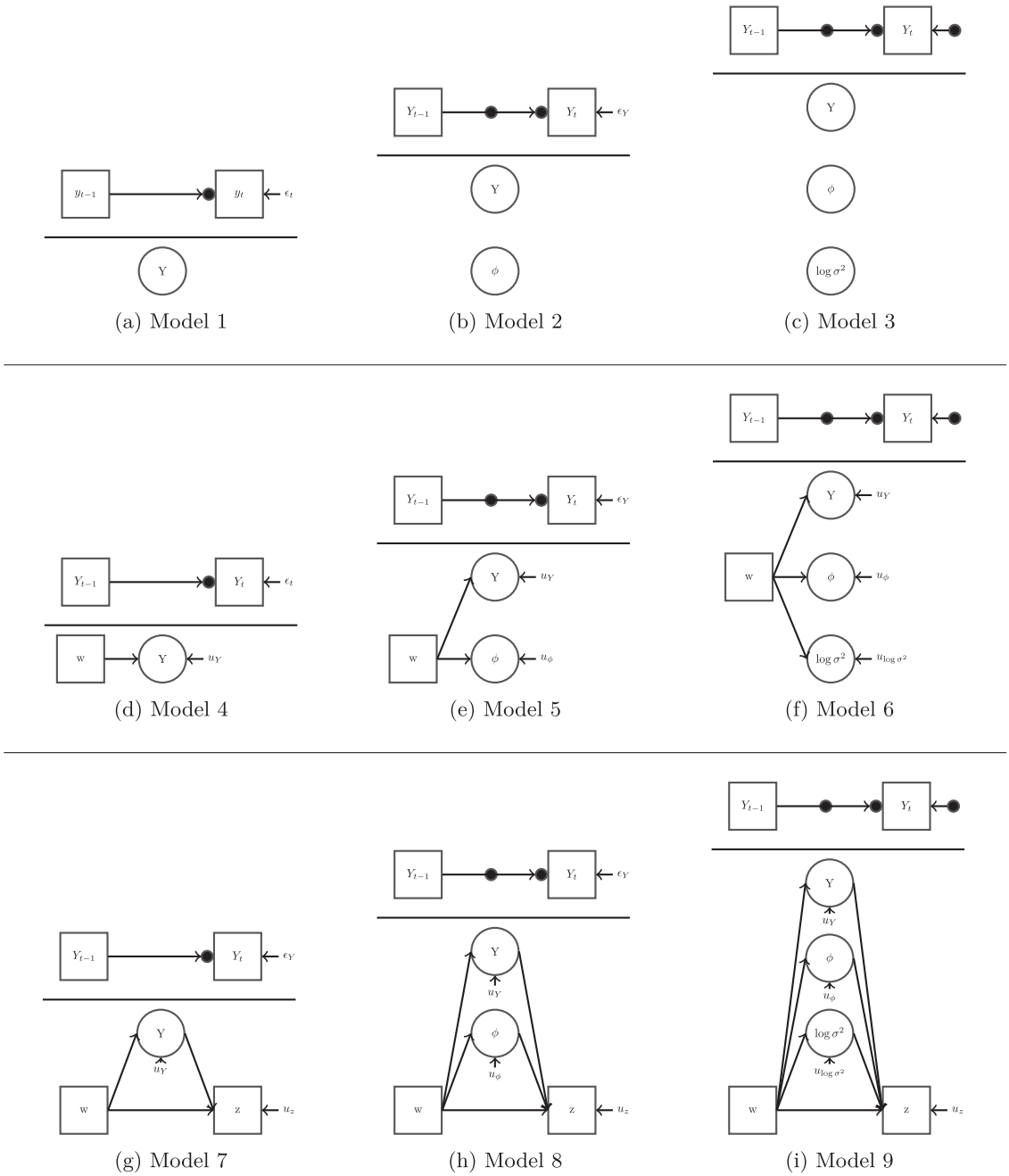


FIGURE 1 Model diagrams of the nine models considered in the Monte Carlo study.

$$
\begin{aligned}
\alpha_i &= \gamma_{00} + \gamma_{01} W_i + u_{0i}, \\
\varphi_i &= \gamma_{10} + \gamma_{11} W_i + u_{1i}, \\
log\sigma_i^2 &= \gamma_{20} + \gamma_{21} W_i + u_{2i}, \\
Z_i &= \beta_{30} + \beta_{31} W_i + \beta_{32}\alpha_i + \beta_{33}, \varphi_i + \beta_{34} log\sigma_i^2 + u_{3i},
\end{aligned} \tag{2}
$$

where $\mathbf{u}_i \sim N_4(\mathbf{0}, \boldsymbol{\Sigma})$ for all i = 1, …, N. This is a two-level time-series model with random mean, random autoregressive parameter, and random residual variance. Equation 2 displays the between-level mediation model with the random coefficients as mediators. Note that, to better fit the normal assumption, the logarithm of the variance is modeled on the between level. The logarithm of the within-level residual variance is referred to as "logv" for short. The M*plus* notation convention uses the name of the within-level dependent variable as the name of the random mean on the between level. Accordingly, the random coefficients are referred to as Y or random mean, φ or random autoregressive coefficient, and logv or logarithm of the random residual variance. On the between level, the random coefficients are all regressed

on the covariate **W**. In addition, a between-level dependent variable Z is regressed on all the random coefficients and **W**. All technical details of the implementation of DSEM models in M*plus* Version 8 can be found in Asparouhov et al. (2017b). The other eight models in the sequence can be obtained from this model by constraining parameters to zero. The models in this article are estimated under the correct model assumption, with the exception of the model misspecification sensitivity analysis performed later.

R^2 and Effect Sizes

To mimic analysis of real data as far as possible, the population parameter values are chosen to achieve certain R^2 and effect sizes. The effect sizes and R^2 used in data generation are specified for each model in the online supplemental materials (www.statmodel.com). Generally, R^2 is set to values between 0.2 and 0.5. To formalize the effect sizes, the framework suggested by J. Cohen (1992) is used. A weak effect is around a 0.3 *SD* change in the dependent variable for a 1 *SD* change in the independent variable. A moderately strong effect is around a 0.5 *SD* change in Y for a 1 *SD* change in X. The effect sizes in the simulation studies are set to 0.2 to 0.3. Given the type of between-level model in the set of Models 7 through 9, the effect size and R^2 in the regressions of the random coefficients on **W** are constrained by the effect sizes and R^2 of the regression of Z. For this reason the effect sizes and R^2 differ to some extent between models.

Evaluation Measures

For each cell of the Monte Carlo study, 500 replications were used. This means that 500 data sets were generated and analyzed. The Monte Carlo output given in M*plus* summarizes the results of the 500 replications. For each parameter considered in the output, five measures are given to evaluate estimation performance. The definition of each measure used to evaluate the estimation quality is given in Equation 3.

$$\begin{aligned}
\text{Relative bias} &= \frac{\text{Average estimate}}{\text{True value}} \\
\text{SE/SD} &= \frac{\text{Average SE estimate}}{\text{Standard deviation of the estimates from the replications}} \\
\text{MSE} &= \frac{\sum_{i=q}^{\#reps} (\text{Estimate}_i - \text{True value})^2}{\text{Number of replications}} \\
95\%\ \text{Coverage} &= \frac{\text{Number of credible intervals covering true value}}{\text{Number of replications}} \\
\text{Power} &= \frac{\text{Number of credible intervals not covering zero}}{\text{Number of replications}}
\end{aligned} \tag{3}$$

Here the parameter estimate is taken as the median of the posterior distribution of the parameters. The choice of the median is discussed later. The Bayesian 95% credible intervals, used in the 95% coverage and power measures, are based on percentiles of the posterior distribution of the corresponding parameter. However, because the intervals are evaluated from a frequentist's point of view, they are referred to as confidence intervals (CI) throughout the article.

The relative bias takes the value one when the point estimate is unbiased. A value of, for example, 1.5 implies that the average estimate is one and a half times larger than the true parameter.

In the *SE/SD* measure the empirical standard deviation (*SD*) is used; that is, the standard deviation of the point estimates from all replications. The *SD* is compared to the average of the *SE* estimates over the replications. If the *SD* and *SE* are similar, the *SE* estimate captures the true variability of the estimates and the *SE/SD* measure will be close to one.

Mean squared error (MSE) can be defined as a function of bias and standard error. If the bias and *SE* are small the MSE is close to zero. This is used as an overall precision measure in the result section, especially for comparison between different N and T allocations.

The 95% coverage measure gives the proportion of replications for which the CI covered the population value.

Power is the proportion of 95% CIs that did not cover zero, that is the proportion of parameters that would be interpreted as significantly different from zero. Power is presented only for between-level regression slopes.

When inspecting the 95% coverage and the power results, it should be noted that the accuracy of the percentile estimates on which the interval is based is a function of number of iterations of the Bayesian algorithm, in this case the number of iterations of the Gibbs sampler. Because high accuracy in estimation of small and large percentiles requires a very large number of iterations such as 10,000 or 50,000, it is not feasible to attain perfect accuracy in this simulation study due to computational time. Ninety-five percent coverage between 0.92 and 0.98 is considered good coverage, which is accurate enough for all the purposes of this article. All technical details of the M*plus* setup and example syntax can be found in the Appendix.

RESULTS

This section highlights patterns of estimation behavior for key parameters in the models. Note that all results, including results not presented here, for all estimated parameters and all the models are available in the online supplementary materials (www.statmodel.com), together with the specific effect sizes and R^2 for all models. Three different cases are considered in this section. Case 1 has a fixed number of subjects (N = 200) where the number of time points per subject varies as T = 10, 15, 20, 25, 50, 100, 200. This serves to illustrate how many time points per subject are needed in a design with many subjects (200). Case 2 has a fixed number of time points per subject (T = 100) where the number of subjects varies as N = 10, 15, 20, 25, 50, 100,

200. This serves to illustrate how many subjects are needed in a design with many time points per subject (100). Case 3 has N and T equal and each varies as 10, 25, 50, 75, 100, 150, 200, 300. This serves to give insights into how small N and T can be in combination.

As a first step, the results of these three cases are compared. Looking at the results together makes the comparison of N and T allocations easier, which might be helpful for designing studies. In addition to the comparisons of the cases, detailed results for each or some of the cases are given for chosen models. We then give the result of a model misspecification sensitivity analysis. Using different sample sizes, this section illustrates what happens if a parameter that is random in the population is fixed in analysis.

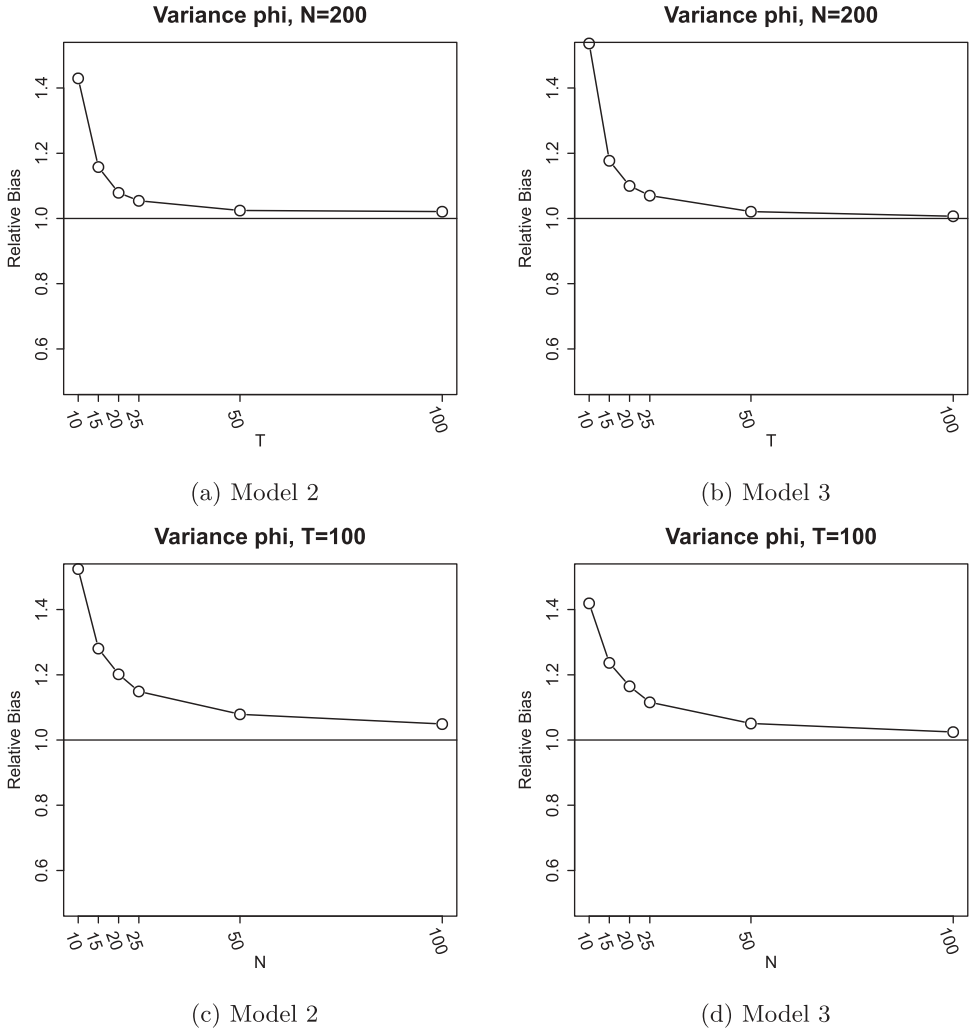
N and T Comparisons

This section gives the results for Models 1 through 3, 4 through 6, and 7 through 9 separately. Note that the x-axis of the plots in this section uses N*T, N or T depending on which cases are considered.

Models 1 through 3

For Models 1 through 3, the means of the random coefficients have relative bias and *SE/SD* close to 1 for $T \geq 10$ and $N > 15$. The variances of the random coefficients require N and T of between 50 and 75 for similar good performance. Figure 2 displays the bias of the variance of the random φ for Models 2 and 3 as functions of N and T, respectively. A very small T gives a somewhat larger relative bias than a very small N. For N and T larger than 20, however, increasing N reduces the relative bias faster than increasing T.

Coverage is between 0.92 and 0.98 for all parameters. Power is presented only for between-level regression slopes. Because of this, power is not relevant for Models 1 through 3. Because the performance is good and similar for all three ($N = 200$, $T = 100$, $N = T$) cases, Models 1 through 3 are not discussed further here.



(a) Model 2

(b) Model 3

(c) Model 2

(d) Model 3

FIGURE 2 Relative bias of the variance of the random coefficient φ. $N = 200$ and $T = 100$ for Model 2 and 3.

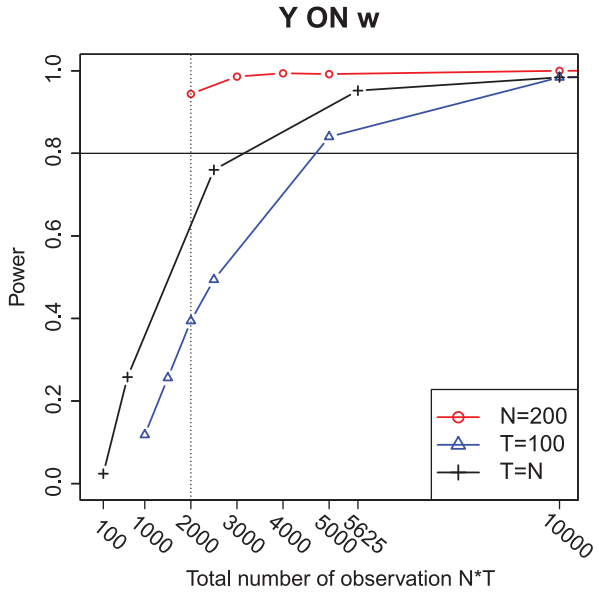


FIGURE 3 Power for between-level regression slopes of Model 4 for all three cases $N = 200$, $T = 100$ and $N = T$.

Models 4 through 6

The relative bias and *SE/SD* measures of Model 4 indicate that the precision in the slope of Y regressed on **W** is good for all N and T. The only measure that substantially differs between the cases is power. Figure 3 displays the power of the slope of the random mean Y regressed on **W** for the $N = 200$, $T = 100$ and $N = T$ cases. The $N = 200$ case consistently has the highest power in the range of total number of observations 2,000 to 10,000; that is, $N = 200$ and T between 20 and 50. $N = T$ is the second best option in this interval. This indicates that there is no benefit of having a large T without a substantial N, whereas a large N always seems beneficial.

Figure 4 displays the power and MSE for the two between-level slopes of Model 5 for all three cases. The power patterns are very similar to those of Model 4. MSE follows the same pattern as the power, where $N = 200$ has the lowest (i.e., best) MSE, followed by $N = T$ and finally $T = 100$. All cases have MSE close to zero and power close to one for a total number of observations larger than 10,000.

Figure 5 displays the $T = 100$ case for Model 5. When $T = 100$, Model 5 requires $N \geq 50$ for power above 0.8 together with low MSE. For similar quality of estimation in the $N = 200$ case, $T \geq 10$ is sufficient.

MSE of Model 6 is close to zero for all sample sizes except for $N = T = 10$. MSE of Model 6 is slightly lower than that of Model 5. A probable explanation for this is that the R^2 is a few percent higher for Model 6 than Model 5. The power and MSE of the very small sample sizes seem sensitive to effect size and R^2. This is confirmed by results from running the same model with different effect sizes (see, e.g., the results from Model 9, displayed later). Figure 6 displays the power and MSE of the between-level slopes of Model 6. Again, the pattern from Model 4 and 5 is repeated. The additional random logv has higher sample size requirements than Y and φ. This is evident in Figure 6c where not even the N = 200 case attains power above 0.8 for all considered total number of observations. Coverage is between 0.92 and 0.98 for Models 4 through 6 for all cases.

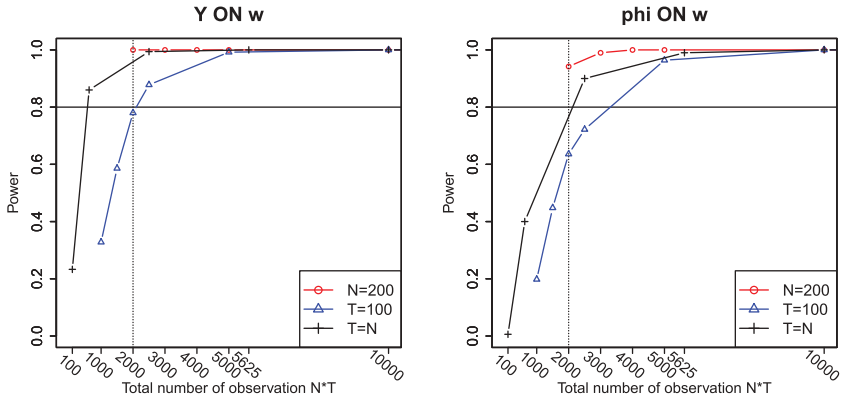
Models 7 through 9

Figures 7 through 9 and Figures 11 through 12 display power for the between- level slopes of Models 7 through 9. The power patterns of the random coefficients regressed on **W** all follow the same pattern as those of Models 4 through 6. In fact, the power and MSE performances are only slightly worse for these slopes when the regression of Z is added. This indicates that the estimation performance of the slopes in Models 4 through 6 does not suffer severely from adding a between-level dependent variable, such as Z.

Figure 7 displays the power and MSE of the between-level slopes of Model 7. The slopes of Z regressed on the random coefficients and **W** have substantial MSE for the $T = 100$ and $N = T$ cases until the total number of observations is larger than approximately 5,000. The MSE of the $N = 200$ case is small for all T. The power is also highest for the $N = 200$ case followed by $N = T$ and lowest for $T = 100$. For a total number of observations larger than 10,000, MSE is close to zero and power above 0.8 for all cases.

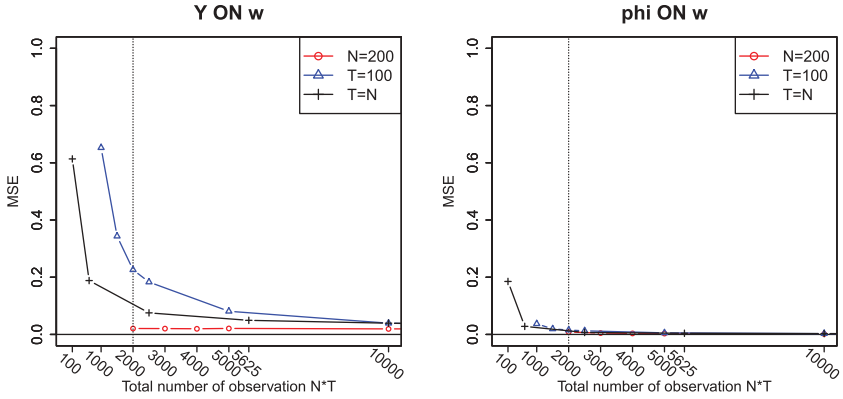
Figure 8 displays the power and MSE of the slopes of Z regressed on the random coefficients and **W** for Model 8. The power results for the slopes of Z regressed on the random coefficients and **W** have several interesting aspects. For a total number of observations between 2,000 and 10,000, the $N = 200$ case has the highest power followed by the $N = T$ case and last the $T = 100$ case. However, the power of $N = T$ and $T = 100$ crosses at $N = T = 100$ (10,000 on the x-axis) from where $T = 100$ has the second highest power after $N = 200$. At a total number of observations larger than 20,000, $T = 100$ has the highest power followed by $N = 200$ and last $N = T$. For a total number of observations larger than 20,000, all cases have MSE close to zero and power above 0.8. The MSE patterns of the three cases are more similar to each other than the power patterns for the same range of total number of observations. However they still have the same internal ordering, crossing at the same places. For a total number of observations larger than 20,000, all cases have MSE close to zero and power above 0.8. Note that the $N = 200$ case attains similar properties already at a total number of observations larger than 10,000 (i.e., $N = 200$, $T \geq 50$).

Figure 10 displays the relative bias of Z regressed on the random φ and **W** for Model 8 in the $N = 200$ case. The relative bias of φ is unstable for $T < 25$, indicating that the estimation of this model is not working properly for such



(a) Model 5

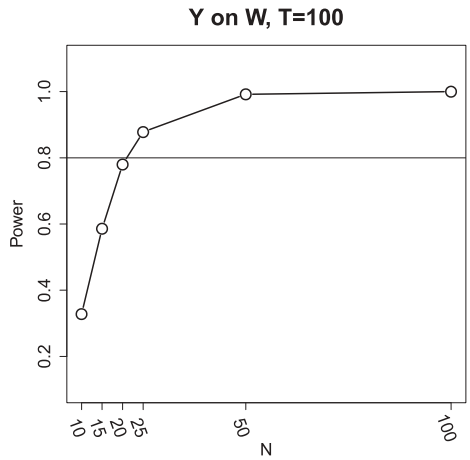
(b) Model 5



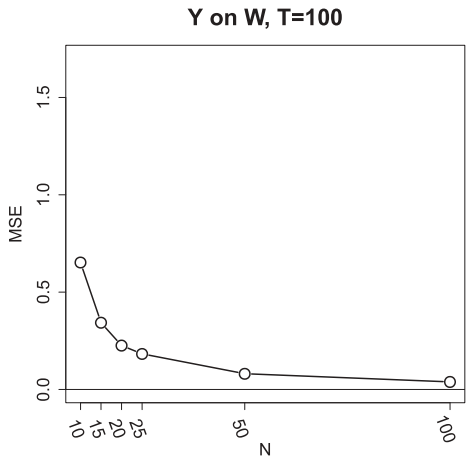
(c) Model 5

(d) Model 5

FIGURE 4 Power and mean squared error for between-level regression slopes of Model 5 for all three cases $N = 200$, $T = 100$ and $N = T$.



(a) Model 5

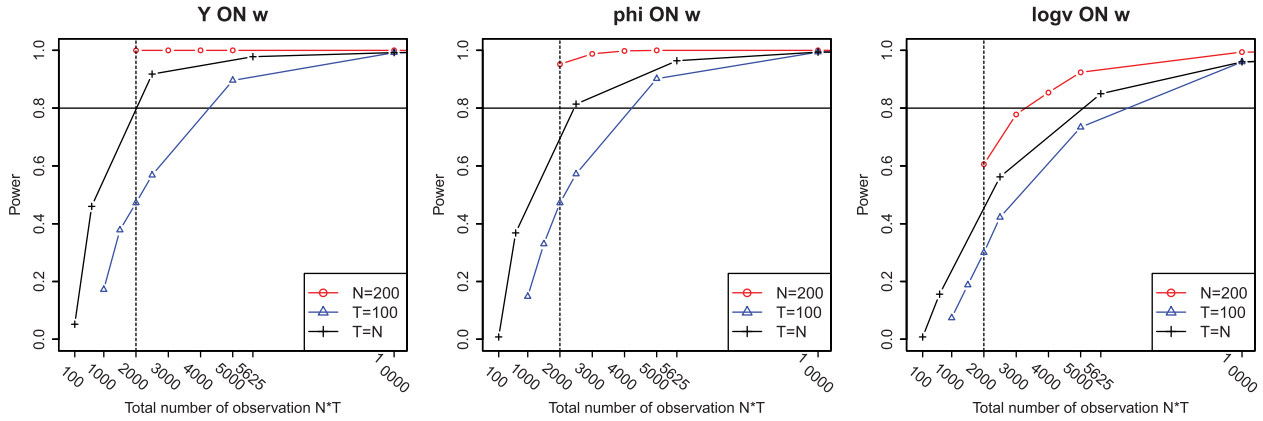


(b) Model 5

FIGURE 5 Power and mean squared error for between-level slope of the Y regressed on **W** for Model 5 with $T = 100$.

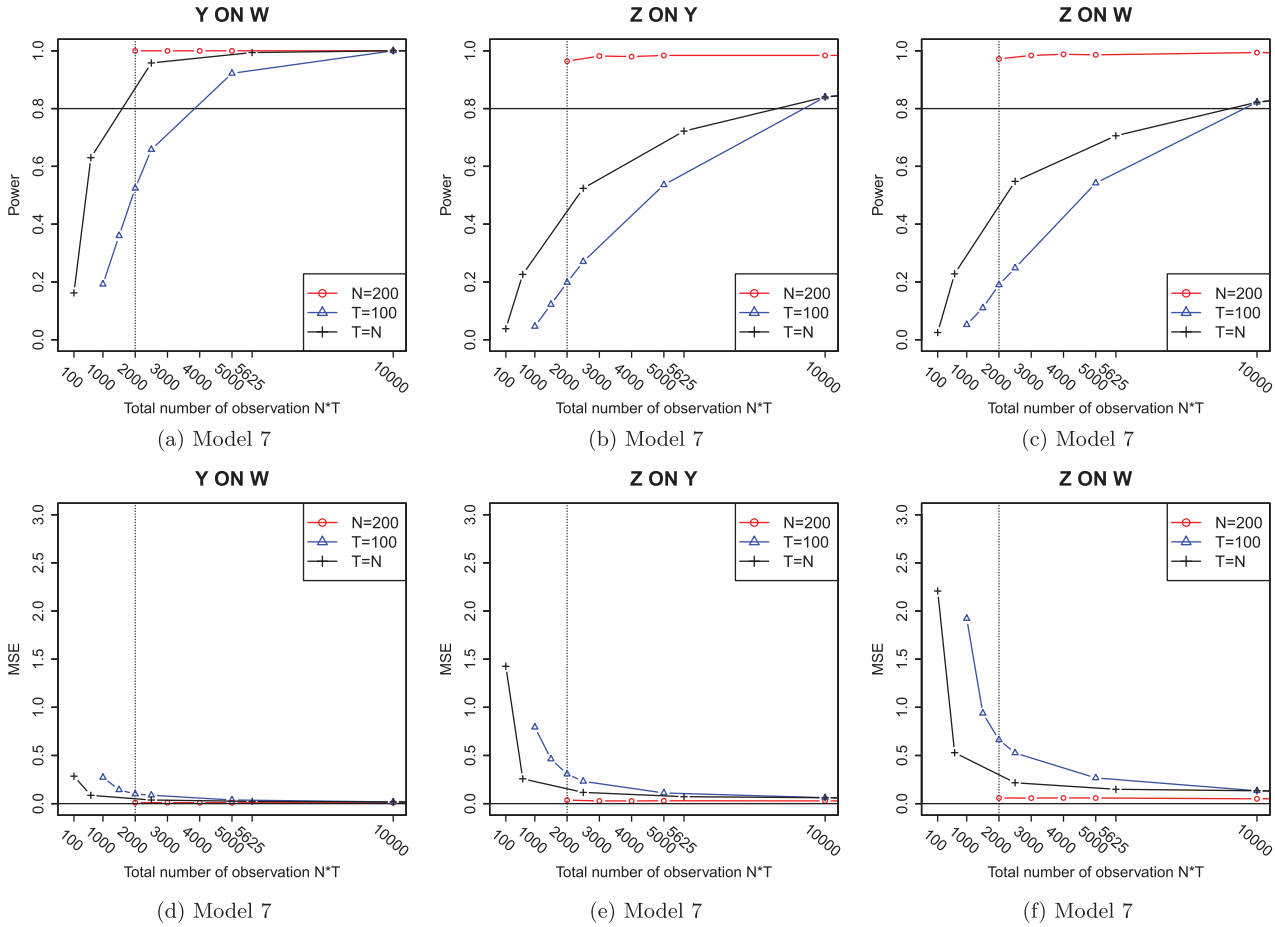
few time points. At $T \geq 50$ the relative bias stabilizes around one and the power comes close to 0.8. The *SE/SD* is close to one for $T \geq 15$.

Model 9 contains many parameters of potential interest. To give extra insight, two effect sizes are considered for this model, a weak effects model and a moderately strong effects



(a) Model 6 (b) Model 6 (c) Model 6

FIGURE 6 Power for between-level regression slopes of Model 6 for all three cases $N = 200$, $T = 100$, and $N = T$.



(a) Model 7 (b) Model 7 (c) Model 7

(d) Model 7 (e) Model 7 (f) Model 7

FIGURE 7 Power and mean squared error for between-level regression slopes of Model 6 for all three cases $N = 200$, $T = 100$, and $N = T$.

model. For Model 9, some sample size settings were excluded due to convergence issues. These settings required as many as 40,000 iterations to properly converge, thus running 500 replications was impractical due to computational time. This was a problem also for the moderately strong effects model, which indicates that this is probably not only due to the small effect sizes. Judging by the results of the included sample sizes here, the excluded

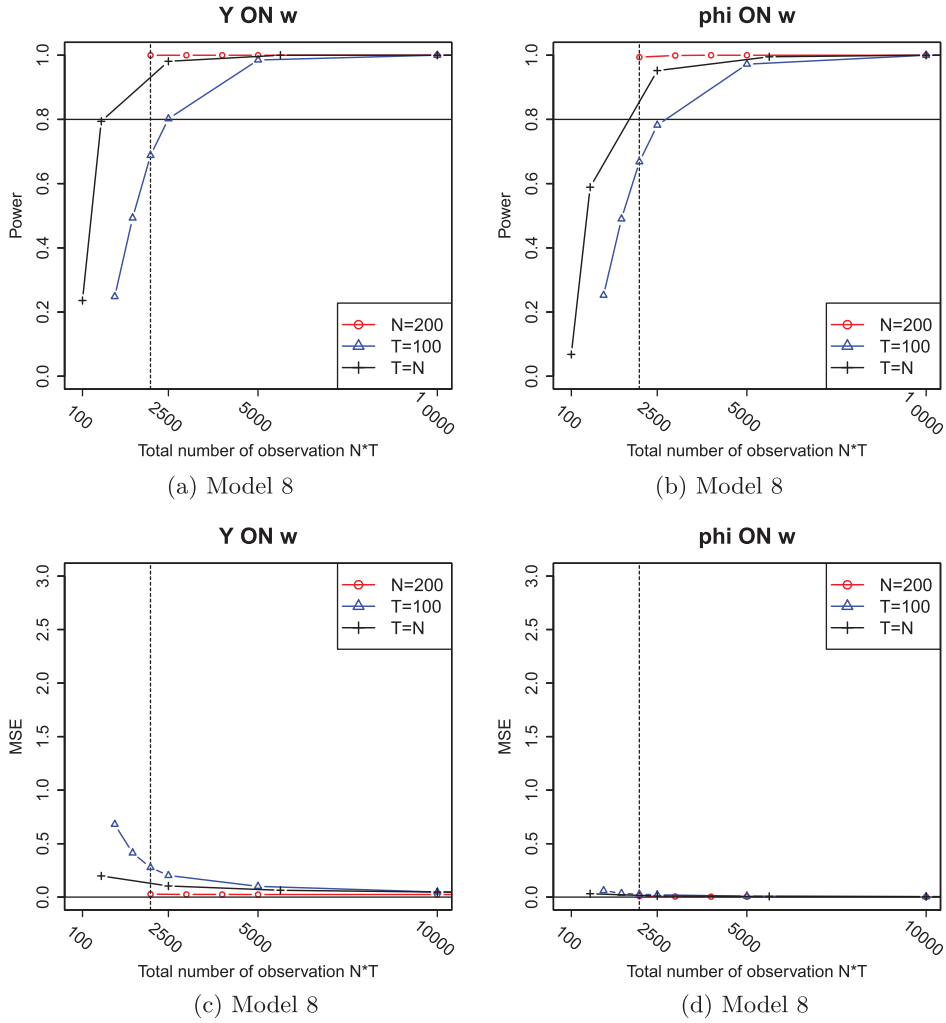


FIGURE 8 Power and mean squared error for between-level slopes of the random coefficients regressed on **W** in Model 8 with all three cases $N = 200$, $T = 100$, and $N = T$.

sample sizes are too small for any reasonable estimation properties regardless of convergence.

Figure 12 displays the power and MSE of the slopes of Z regressed on the random coefficients and **W** for Model 9. The power and MSE patterns are very similar to those of Model 8. One important difference is that the requirements for good properties are substantially higher. No parameters for any of the three cases attain power above 0.8 in the studied range of sample sizes.

The slope of Z regressed on **W** has power far below 0.8 for all cases. Because the effect sizes of the random coefficients on Z are between 0.20 and 0.23, whereas the corresponding size of **W** is only 0.15, the power is expected to be very low. It is clear that the power is consistently highest for the slope of the random mean, with power higher than 0.8 already for $T \geq 50$. The corresponding T requirement for the random φ is $T \geq 100$. The random logv does not attain 0.8 power for $T \leq 200$.

For Model 9 with moderately strong effects the results change substantially. Figure 13 displays the power of the slopes of Z regressed on the random coefficients and **W** for the moderately strong effects version of Model 9 in the $T = 100$ case. When the effects are stronger the sample size requirements drastically decline. For this case, $T = 100$, $N = 100$ (total number of observation = 10,000) give power above 0.8 for all slopes in the Z regression.

Figure 14 displays results from Model 9 with moderately strong effects in the $N = 200$ case. Note that the power is systematically lower for Z regressed on **W** than for Z regressed on logv even when the population effect size is slightly larger for the slope of **W** (0.30) than logv (0.26). *SE/SD* and coverage results are similar to the weak effects model.

Although the effect size in the moderately strong effects for Model 9 implies an implausibly large R^2 in the Z regression, around 80%, this setup serves to illustrate that

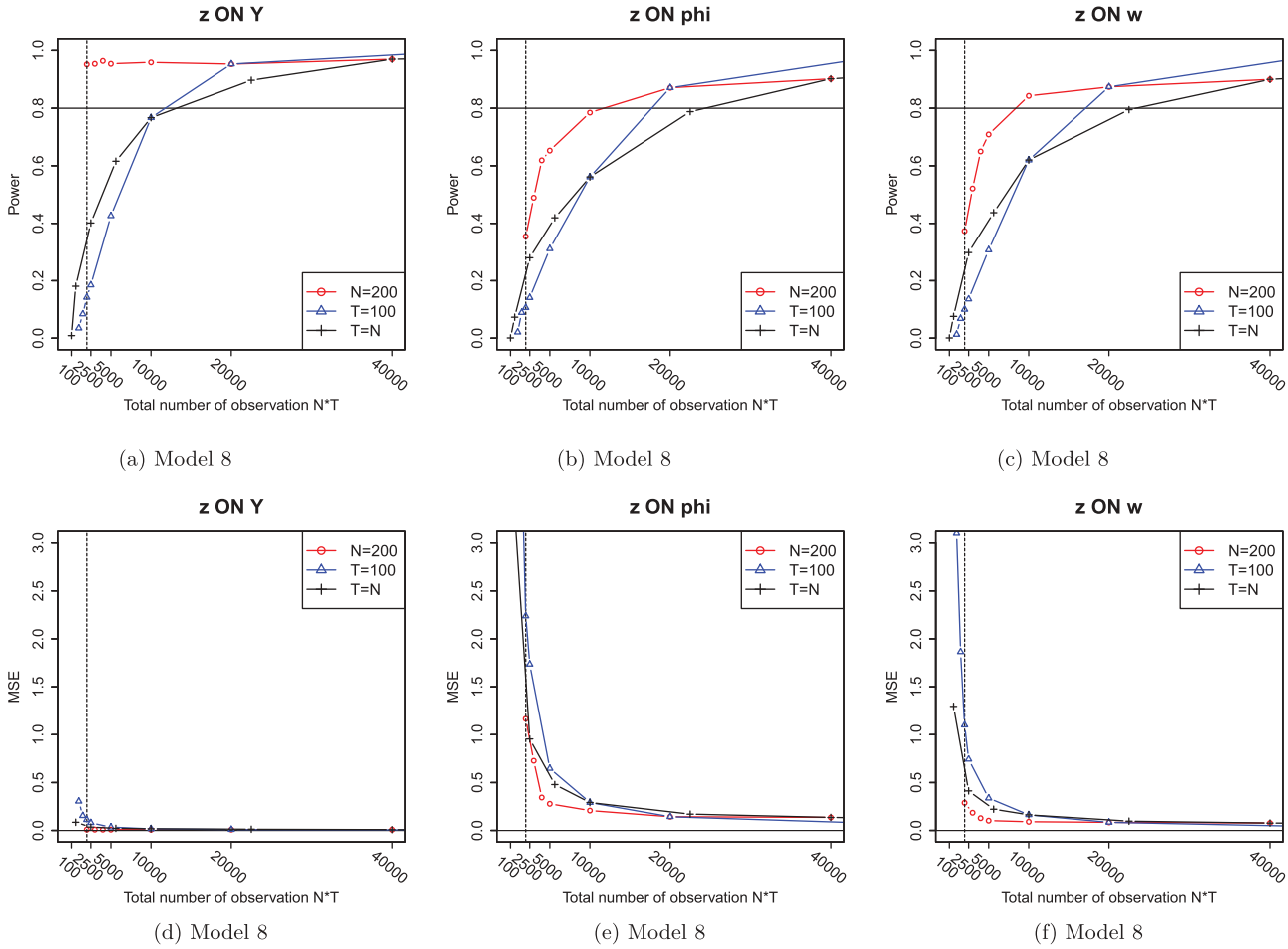


FIGURE 9 Power and mean squared error for between-level slopes of Z regressed on the random coefficients in Model 8 with all three cases $N = 200$, $T = 100$, and $N = T$.

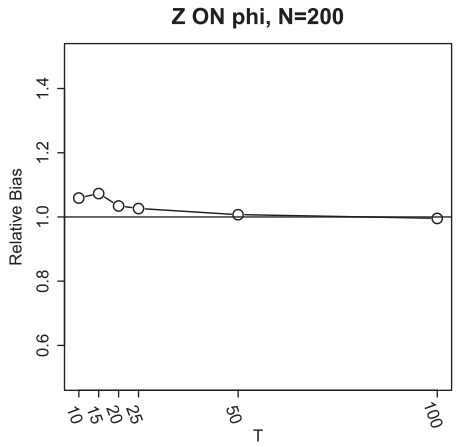


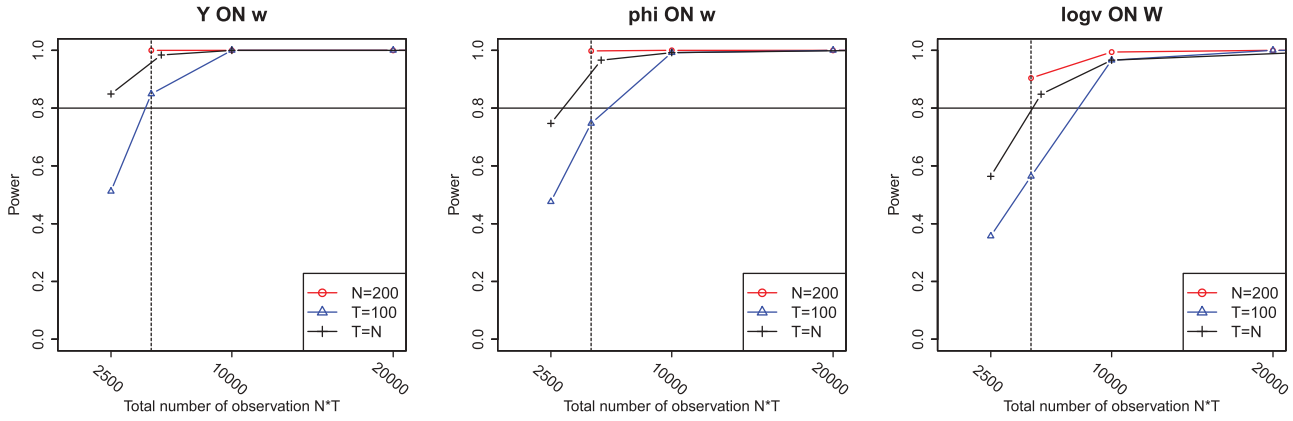
FIGURE 10 Relative bias for the slope of Z regressed on the random autoregressive coefficient φ in Model 8 with $N = 200$..

the pattern and differences between, for example, the slopes of Z regressed on logv and $\mathbf{W}$ are at least partly due to the choice of effect size. There is a substantial increase in the power of Z on logv, even though the effect size only increased from 0.224 to 0.257 between the weak and strong effects model. Coverage is between 0.92 and 0.98 for Models 7 through 9 with a few small deviations for Models 8 and 9.

Note that the simulations are performed under balanced samples; that is, with no missing data at any time point. Simulations with missing values (missing completely at random) on the within-level variable were performed for Model 3. The results showed that, beyond the expected decline in estimation quality due to a smaller effective sample, Model 3 showed little sensitivity to unbalance. That is, even if not all subjects are measured at each time point, as long as the observed sample size is still large enough, the model seems to perform well.

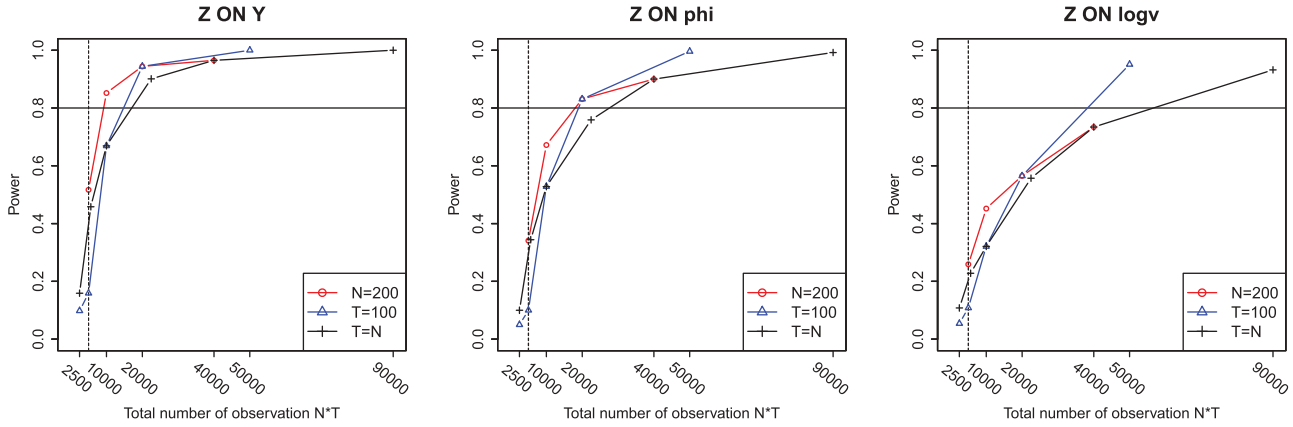
Summary

The comparisons of the three cases $N = 200$, $T = 100$, and $N = T$ give several interesting insights. For an especially small total number of observations, a large N is always better than a large T. This is true for practically all parameters.

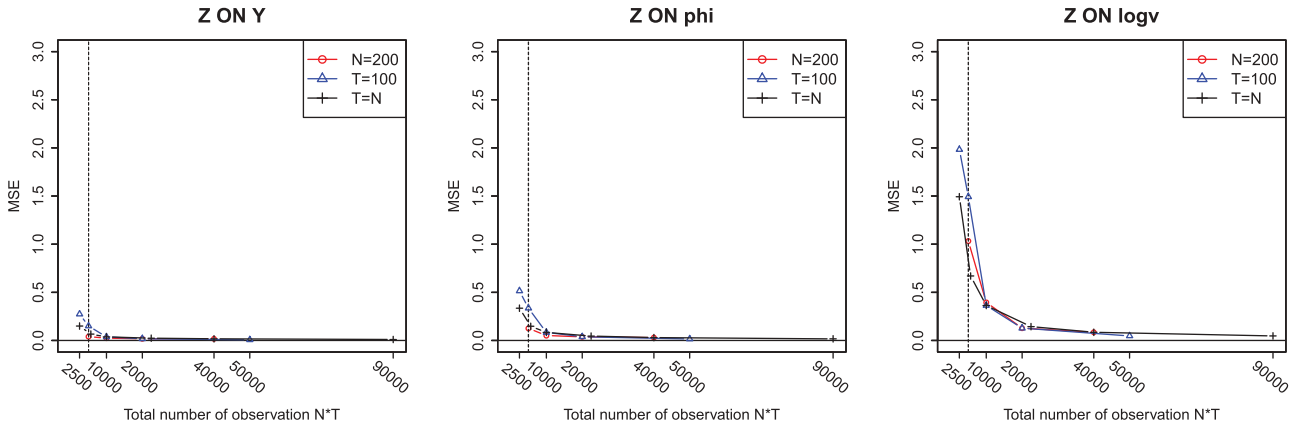


(a) Model 9 (b) Model 9 (c) Model 9

FIGURE 11 Power for between-level slopes of the random coefficients regressed on **W** in Model 9 with weak effects for all three cases $N = 200$, $T = 100$, and $N = T$.

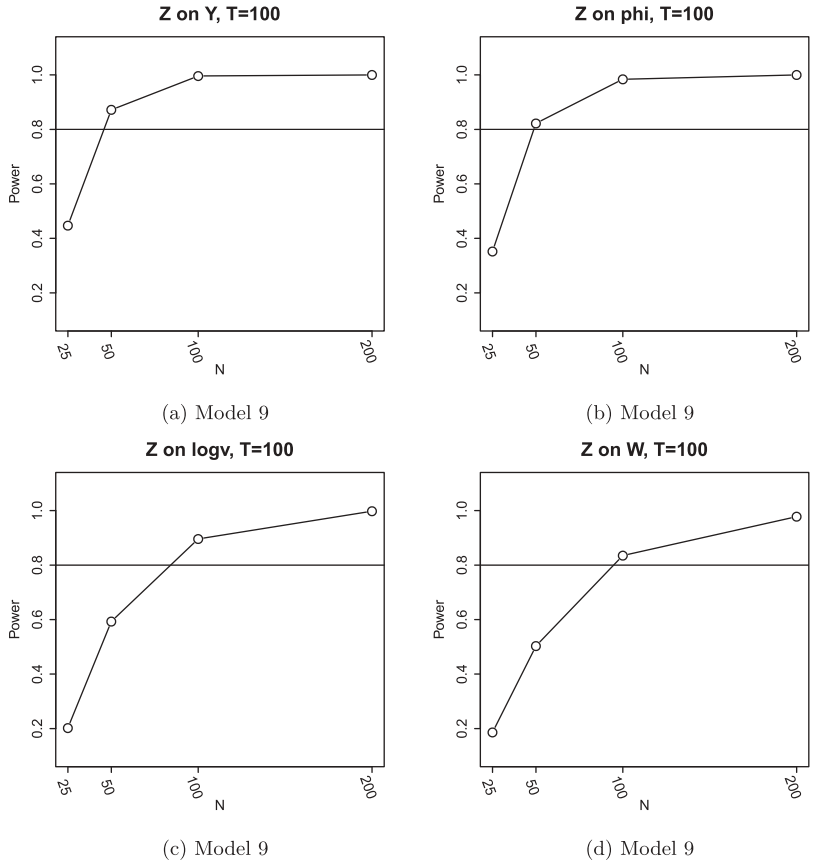


(a) Model 9 (b) Model 9 (c) Model 9



(d) Model 9 (e) Model 9 (f) Model 9

FIGURE 12 Power and mean squared error for between-level slopes of Z regressed on the random coefficients in Model 9 with weak effects for all three cases $N = 200$, $T = 100$, and $N = T$.



(a) Model 9 (b) Model 9

(c) Model 9 (d) Model 9

FIGURE 13 Power for between-level slopes of Z regressed on the random coefficients in Model 9 with weak effects and $T = 100$.

The power and MSE patterns in the regression of the random coefficients regressed on **W** are very similar for Models 4 through 6 and Models 7 through 9. The patterns of Z regressed on the random coefficients and **W** are also similar to those of the random coefficients regressed on **W**, but with the important difference that the sample size requirements for all cases increase substantially. $N = 200$ is best until the total number of observations is between 10,000 and 20,000, where $T = 100$ becomes better. $N = T$ is better than $T = 100$ up until $N = T = 100$, after which $T = 100$ is better than $N = T$. Note that this pattern is in line with the conclusion that large N is most important as after 20,000 total number of observations, the $T = 100$ case has larger N than the $N = 200$ case. The results of this section also point to the conclusion that the random mean requires a lot smaller sample than the random φ and logv, both as a dependent and as an independent variable.

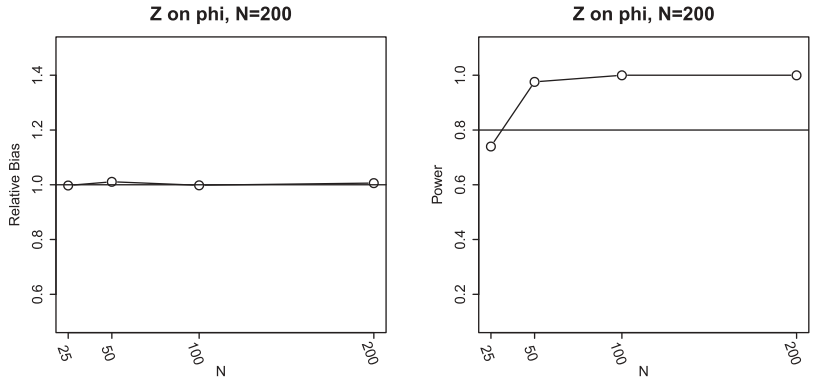
Comparing the $N = 200$ case against the $T = 100$ case gives interesting insights for designing studies of time ILD. The $N = 200$ results highlight the substantially different requirements across the nine models. $N = 200$ seems to give very good performance overall, even for T as small as 10 for some models. However, the most complex models, using random coefficients as predictors, require substantially larger T to attain the same properties. For all models there are indications that the random φ and logv are a lot more demanding than the random mean Y. The results from $T = 100$ indicate that even in the simplest empty two-level models, Models 1 through 3, the relative bias is substantial for very small N, which is not the case for small T in the $N = 200$ results.

Model Misspecification

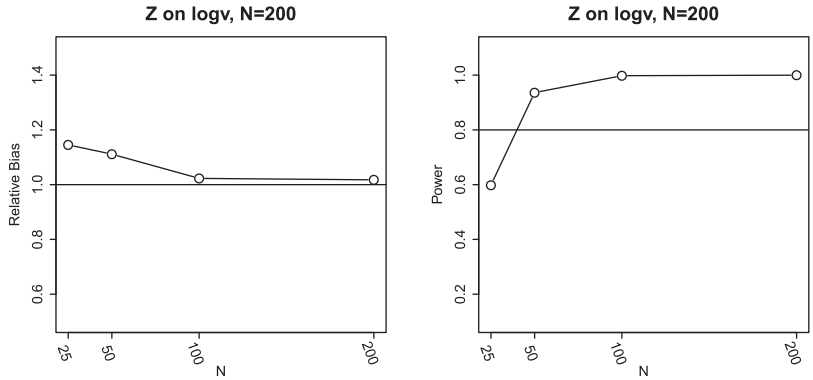
In this section, Model 6 is investigated further as three misspecification scenarios are analyzed. In Model 6, the mean, autoregressive coefficient, and logv are random and regressed on a between-level covariate **W**. In these scenarios φ, logv, or both φ and logv are incorrectly estimated as fixed, instead of the correct specification as random. The runs are performed with $N = T$. The x-axis of the plots in this section uses the total number of observations (i.e., $N*T$). For instance, 10,000 on the x-axis implies $N = T = 100$.

Incorrectly fixed phi

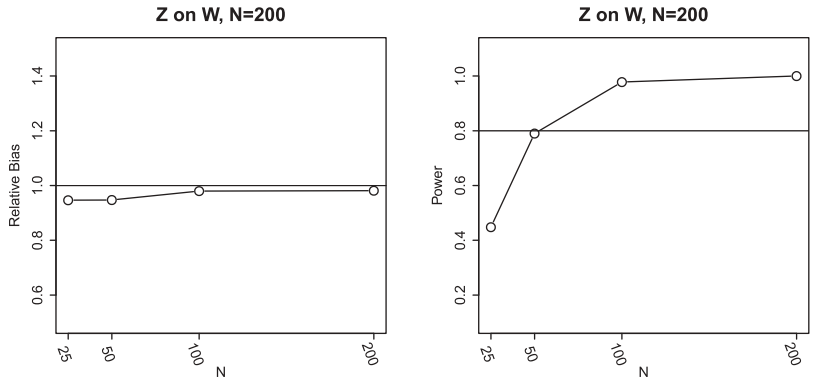
In this section, data generated from Model 6 are analyzed in two ways: with the correct model specification (i.e., random mean), φ and logv, and with incorrectly fixed φ.



(a) Model 9, mod. strong effects (b) Model 9, mod. strong effects



(c) Model 9, mod. strong effects (d) Model 9, mod. strong effects



(e) Model 9, mod. strong effects (f) Model 9, mod. strong effects

FIGURE 14 Relative bias and power for between-level slopes of Z regressed on the random coefficients in Model 9 with weak effects and $T = 100$.

Figure 15 displays the MSE of the slope of the random mean Y and the random residual variance logv regressed on **W**. The difference between the correctly specified model and the model with incorrectly fixed φ is not substantial.

Incorrectly fixed logv

In this section, data generated from Model 6 are analyzed in two ways: with the correct model specification (i.e., random mean), φ and logv, and with incorrectly fixed logv.

Figure 16 displays the MSE of the slope of the random mean Y and the random autoregressive coefficient φ regressed on **W**. The difference between the parameter estimates of the correctly and incorrectly specified model are similar to when φ is incorrectly fixed above (i.e., not substantial).

Incorrectly fixed φ and logv simultaneously

In this final section of misspecification sensitivity, data generated from Model 6 are analyzed in two ways: first,

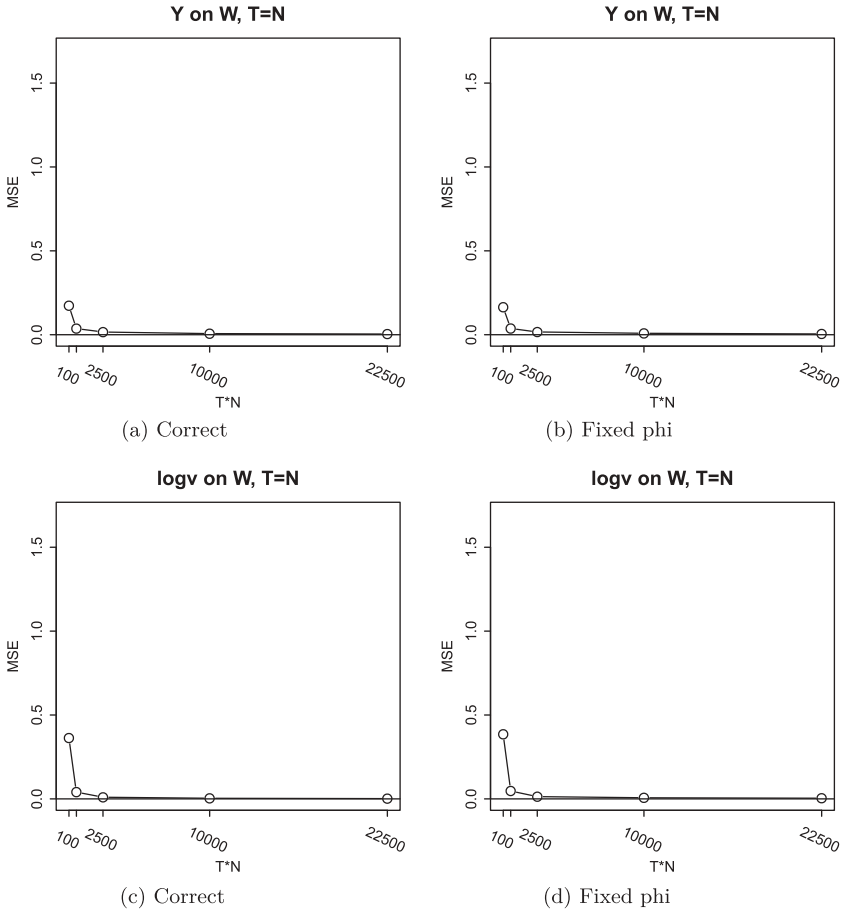


FIGURE 15 Mean squared error for the slope of **W** regressed on the random coefficients. Data generated from Model 6 and analyzed with model misspecification of the random coefficient φ.

with the correct model specification (i.e., random mean), φ and logv; second, with simultaneously incorrectly fixed φ and logv.

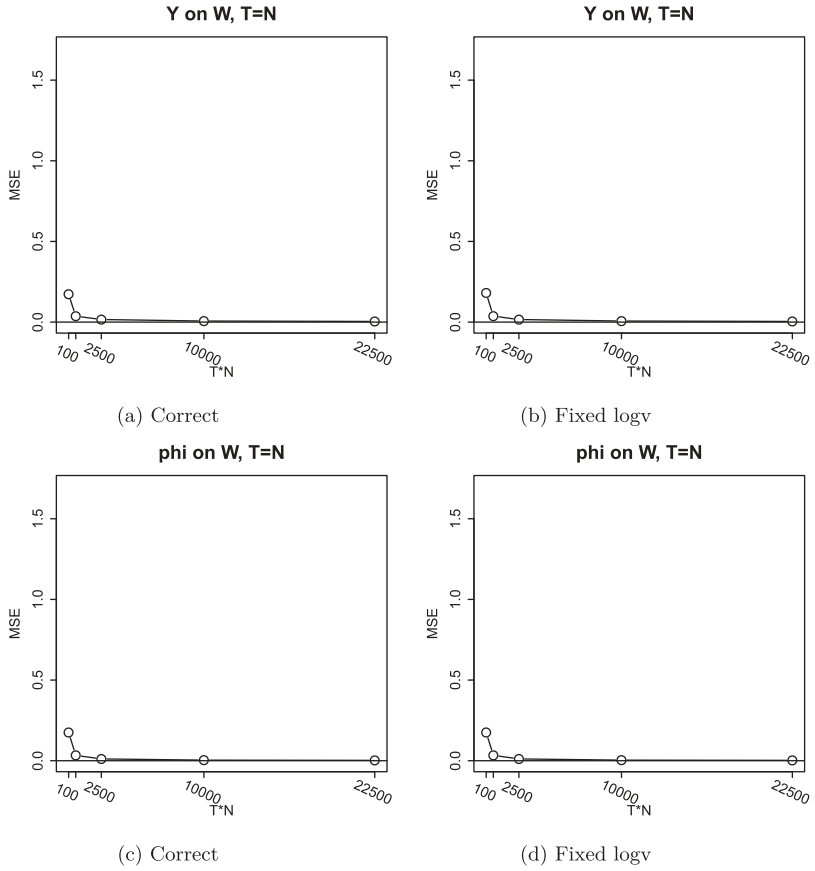
Figure 17 displays the results. In this case the MSE becomes very large when $N = T \geq 100$. The lower plots explain this pattern. The *SE* estimate becomes strongly underestimated as the total number of observations increases. This indicates that when both φ and logv are incorrectly fixed, the model specification is too far off and the model estimation breaks down.

Summary

The results of this small model misspecification sensitivity analysis indicate that if either the random φ or logv are incorrectly fixed, the model seems to be able to compensate for that to some extent. The slope of Y on **W** is practically not affected by the misspecification. When φ is incorrectly fixed, the residual variance of the random mean Y is overestimated by a factor of approximately 2. When logv is incorrectly fixed, the residual variance of the random mean is unbiased. However, when both φ and logv are incorrectly fixed, the model breaks down with strongly underestimated between-level *SE*s as a consequence. This in turn might cause Type I errors. The quadratic pattern in Figure 17b is likely explained by two different mechanisms. With small N and T there are estimation problems due to too little information. With larger N and T, on the other hand, the probability of sampling an individual with a deviating φ or logv increases with N. In addition, the probability that a large deviating realization occurs for such individuals increases with T. Therefore, the fixed parameters become increasingly unable to describe the diversity across subjects with increasing N and T.

Analysis Strategies

In exploratory analysis of two-level time-series data, there are a few important considerations. The results from the models with incorrectly fixed parameters earlier indicate that it is a good idea to always analyze two-level data with the residual variance specified as random, at least initially. There are two reasons for this: It is very often a reasonable



(a) Correct (b) Fixed logv

(c) Correct (d) Fixed logv

FIGURE 16 Mean squared error for the slope of **W** regressed on the random coefficients. Data generated from Model 6 and analyzed with model misspecification of the random coefficient logv.

assumption, as discussed in Jongerling et al. (2015), and it causes large relative biases in the random mean modeling if not correctly included. If, on the other hand, the residual variance is in fact not random, that will show up as a very small variance for the random residual variance. This suggests the somewhat unusual exploratory approach, starting with the most complex empty two-level model and then restricting the parameters with nonsubstantial variance.

CAVEATS WHEN SIMULATING TWO-LEVEL TIME-SERIES MODELS

This section collects things that are good to be aware of when working with two-level time-series models, especially when doing simulations.

Time-Series-Related Caveats

There are some features of the models considered in this article that call for extra caution in simulation studies. The most crucial component to keep an eye on in all the models considered is the random autoregressive coefficient φ. For an autoregressive process of order 1 (the only AR process considered in this article) to be stationary, $|\varphi|<1$ must hold. When generating the random φ from nontruncated normal distributions, values violating the stationarity assumption can of course be obtained. If, for example, the distribution for φ is set to be normal with $\mu = 0.2$ and $\sigma = 0.2$, the probability of drawing a value larger than 1 is small, around 1/30,000. With N = 200 and 500 random draws, $\varphi>1$ is expected to be generated for approximately three subjects. If a subject has a nonstationary process, it can explode (take very large values) due to the variance becoming nonfinite (see, e.g., Hamilton, 1994). If the outcome of the process becomes large enough, M*plus* Version 8 will discover the large numerical values in the data and skip that replication. If the values are not large enough,[1] however, estimation will proceed, usually with biased estimates that distort the simulation summary results. A two-step internal–external Monte Carlo simulation can be set up to study this (Muthén &

[1] M*plus* version 8 will automatically not run analysis on a data set if any variable in the generated data set has values larger than 10,000. An error message is printed if this happens (Muthén & Muthén, 2017).

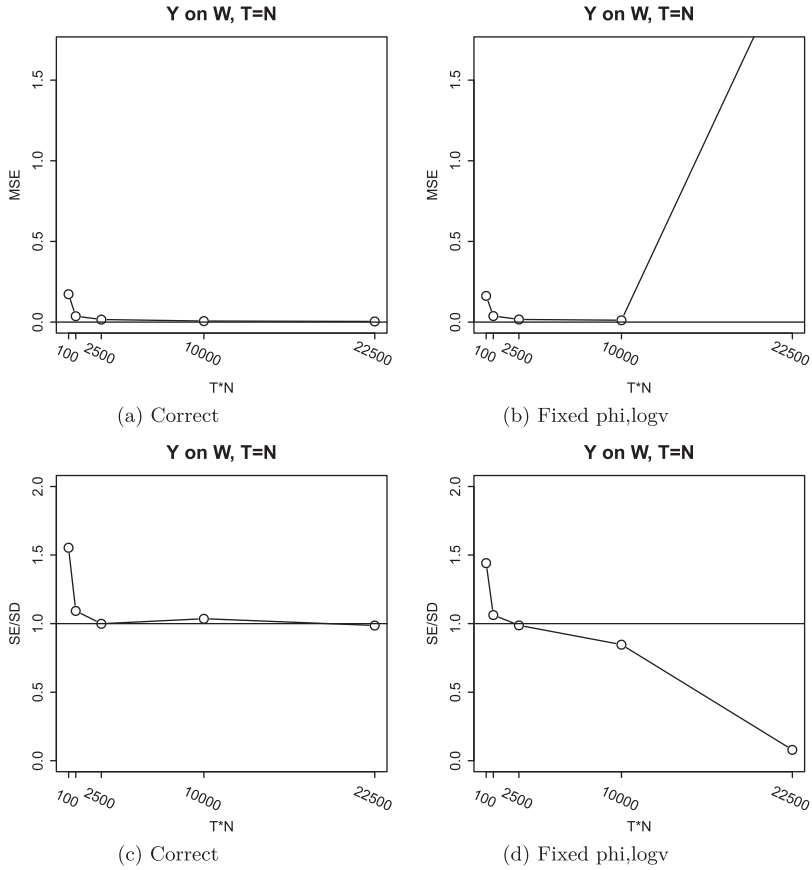


FIGURE 17 Mean squared error for the slope of **W** on the random coefficients. Data generated from Model 6 and analyzed with model misspecification of random coefficients.

Muthén, 2017). The data sets from the internal runs can be saved and inspected. The data sets with exploding series can then be deleted and the external Monte Carlo simulation run only with the remaining data sets. It can be hard to detect issues directly from data, whereas it usually becomes a lot clearer in the estimates.

Bayes-Related Caveats

Very few problems related to the usage of the Bayes estimator in M*plus* came up during this study. The only caution is the fact that the power and 95% coverage measures are based on the percentiles of the posterior distribution of the parameters. This means that if few samples are drawn, the high and low percentiles will be poorly estimated and the coverage and power estimates less reliable. The choice of using the default median estimate as point estimator is discussed in the Appendix.

Residual Variance of φ

Figure 18 displays a comparison of the MSE of slope of Z regressed on φ in Model 8, with different residual variances of φ. It might at first look counterintuitive that a smaller residual variance is associated with a higher MSE. This, however, has a simple explanation. If the variance of the random φ is too small it makes no sense to specify it as

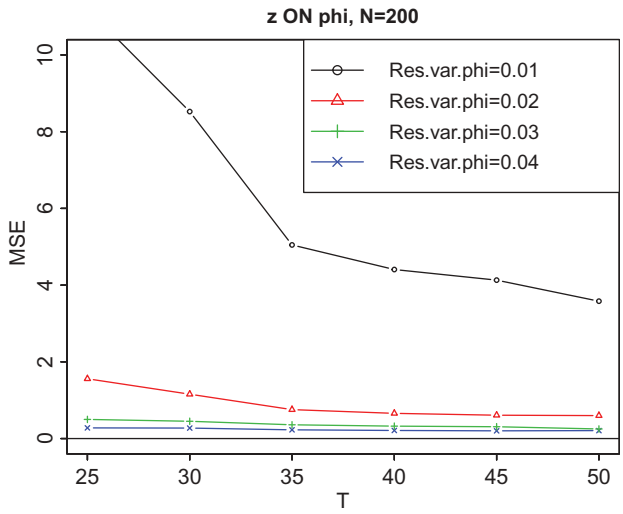


FIGURE 18 Mean squared error for the slope of Z on phi in Model 8 with two different population residual variances of phi.

random or to model the variation with a between-level covariate. However, even if φ has a reasonable variation there are some things to be aware of. Consider, for example, the case when φ is regressed on the between-level covariate **W** and the between-level dependent variable Z is regressed on both **W** and φ, as in Model 8. This model is saturated, which implies that the slope of Z regressed on φ is essentially identified by the residual variance of φ, not the full variation of φ. Because of this, if the residual variance is small, the slope of Z regressed on **W** is poorly identified. This makes sense because if **W** explains almost all variation in φ, it is redundant to regress Z on both **W** and φ. It can be thought of as a multicollinearity problem in the regression of Z. Of course, if this situation occurs the substantive question and natural temporal ordering should decide which model to use. Because the variation of φ is naturally small in a sample of stationary processes, it is good to keep this potential problem in mind.

Sample Size

Because the total number of observations (N*T) grows very quickly with N and T, it becomes impractical to use the same number of iterations for all cases. In the $N = T$ case, the total number of observations increases quadratically: $N = T = 10$ gives 10*10 = 100, and $N = T = 100$ gives 100*100 = 10,000. The estimation time increases approximately linearly with the total number of observations (an increase in T being somewhat more demanding than in N). Thus $N = T = 100$ takes about 100 times longer than $N = T = 10$. Even though the Bayes estimation is fast, the complex nature of some of the models covered in this article and the 500 replications make the simulations time consuming. Because convergence is generally faster (in terms of number of iterations) for larger samples, a large N causes no convergence issues in the study.

DISCUSSION

One of the most striking results of this simulation study is the performance of the random mean. Its performance is good in all models, with MSE close to zero and power close to 1 for most considered sample sizes. Even though the mean is expected to have good properties, it is reassuring to see that the novel feature of a slope for a between-level dependent variable regressed on the random mean also has very good power results. With $N \geq 200$, Model 7, which has a random mean modeled as a mediator on the between level, has power close to 1 for T as small as 10. Thus, using the random mean as dependent variable, independent variable, or both seems reasonable in many not-so-large-sample settings.

The second main finding is the sensitivity of the random autoregressive coefficient ≥ 200. It is sensitive in several ways. First, generating data with normally distributed random φ implies that it can take absolute values larger than 1, which in turn can give exploding series (very large values) for those subjects. In addition, as discussed earlier, the variance or residual variance of φ cannot be too small. For a saturated mediation model on the between level to make sense, the random φ must naturally have variation to be explained by the between-level covariates. Less obvious is the fact that the random φ must have substantial variation in the residual for the regression slope of Z on φ to be well identified. That is, the random φ must have unique variation beyond what is explained by **W**. This applies to all random coefficients modeled as mediators on the between level and is nothing new to mediation modeling. However, the variation of the random φ in stationary two-level time-series data must have quite small variation to begin with to fulfill the stationarity assumption of $|\varphi|<1$. These models are therefore perhaps more likely to suffer from these modeling features than mediation models in general.

Given DSEM's large span of features, it is not meaningful to talk about sample size requirements for DSEM models in general. Instead, requirements for specific features, such as type of model on the between level and different types of random coefficients covered by DSEM, are more informative. For large sample sizes such as N and $T \geq 100$, random mean, random φ, and random logv work well both as dependent variables and as predictors on the between level. In general, the sample size demands for good performance are lower for slopes of the random coefficients regressed on between-level covariates than a between-level dependent variable regressed on the random coefficients and between-level covariates. The random mean has consistently lower sample size demands than the random φ and logv, both as a dependent and independent variable. The random φ has in turn lower demands than the random logv, although the difference is smaller between these two than between the random mean and the random φ.

Natural questions after this summary are as follows: What is worse, having a lower N or a lower T? Can large N compensate for small T better than large T can compensate for small N? The answer seems to be clear: Large N is better. That is, large N seems able to compensate for small T, better than large T can compensate for small N. This means that for outcomes that are expensive or difficult to measure many times, one can get away with using many individuals and few time points. With that said, the random φ and logv do need fairly large T to be well estimated. If φ or logv is of substantive interest rather than just a heterogeneity feature to control for, many repeated measures will be needed.

It might seem obvious to some that the between-level parameters, describing between-subjects variation, benefit from many subjects (large N). However, this logic does not trivially hold in this case. The within-level model is an autoregressive model of order one; that is, an AR(1) time-series model. In the $N = 1$ time-series literature, the smallest recommended T for AR models is usually 50 or even 100

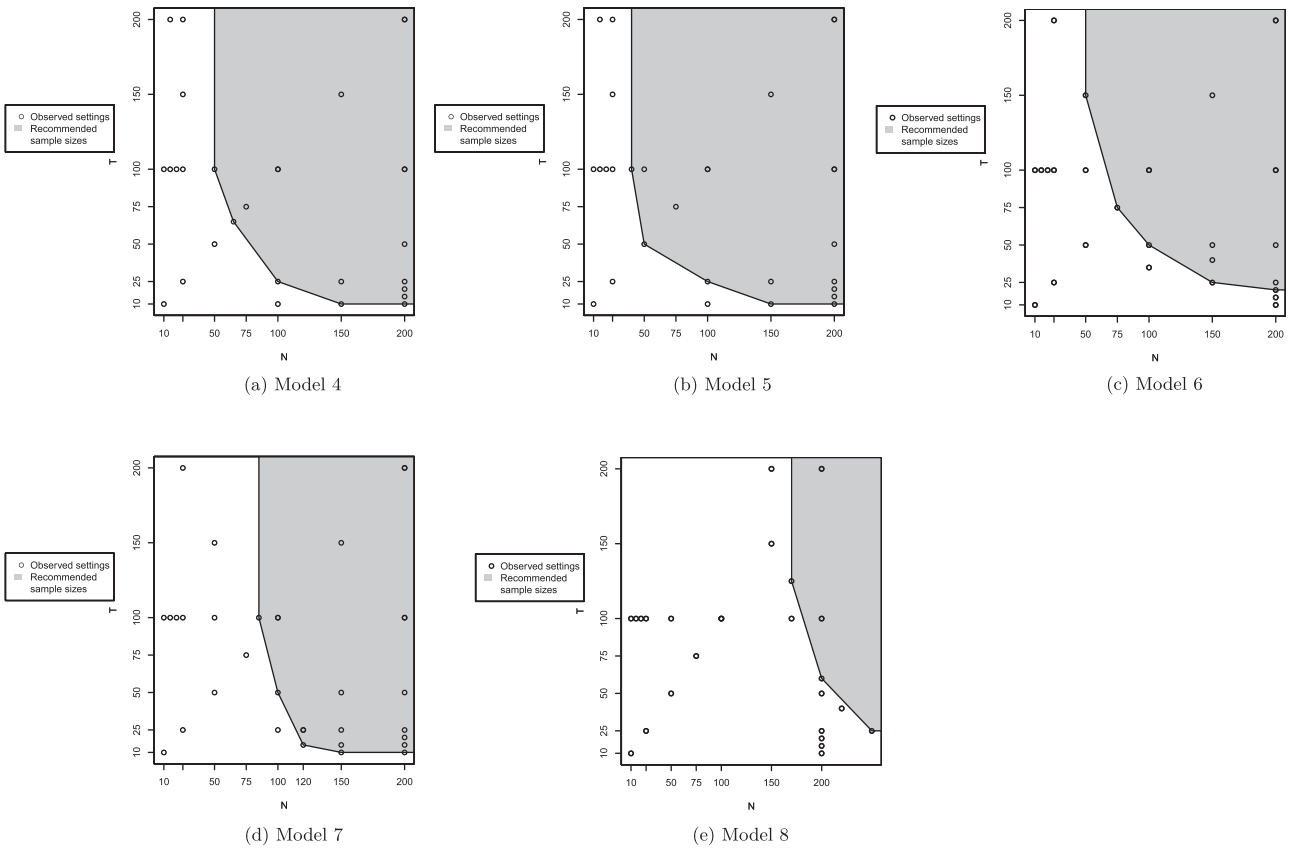
(see, e.g., Chatfield, 2016). This requirement is to get good estimation quality for the parameters from the analysis of one time series. Crude extrapolation to the DSEM models would suggest that time series of each subject would need 50 to 100 time points for the AR(1) model parameter estimates to be estimated with sufficient precision. This would matter also for the precision in the estimation of the between-level model because the estimates from the within level are used as variables. Approaching the same matter from the two-level modeling perspective gives other insights. If clusters share common parameters, the estimation of cluster-specific parameters benefits from a two-level modeling approach. This is because the subjects can "borrow" information from each other about the common parameter.

Taking both these perspectives into account, it is reasonable that for DSEM models with some fixed within-level parameters, the T requirements would be somewhat lower than the $N = 1$ time-series literature suggests. This reasoning does not, however, predict any large reductions in sample size requirements for the models where all three within-level parameters are random because subjects no longer have common parameters. Subjects do, however, still share the model assumption of the within-level AR(1) model with normally distributed random coefficients. It is not obvious from either the $N = 1$ time-series literature or the two-level literature how the subject-specific coefficients benefit from this common model assumption in a two-level time-series context. It is therefore striking that, if N is sufficiently large, this study suggests that T can be a lot smaller than 50 for models with all three coefficients random.

Guidelines Under the Correct Model

To simplify the results of this simulation study, a final set of summary guidelines are presented in the form of Figure 19 and Figure 20. These guidelines apply to models with weak to moderately strong effects and R^2 between 0.2 and 0.4, if nothing else is stated. These are recommendations for attaining the following estimation quality of the slopes of the between-level model: power above 0.8, relative biases less than 10% away from one, and *SE/SD* less than 15% away from one. In addition, the remaining parameters have MSE close to zero and coverage between 0.92 and 0.98.

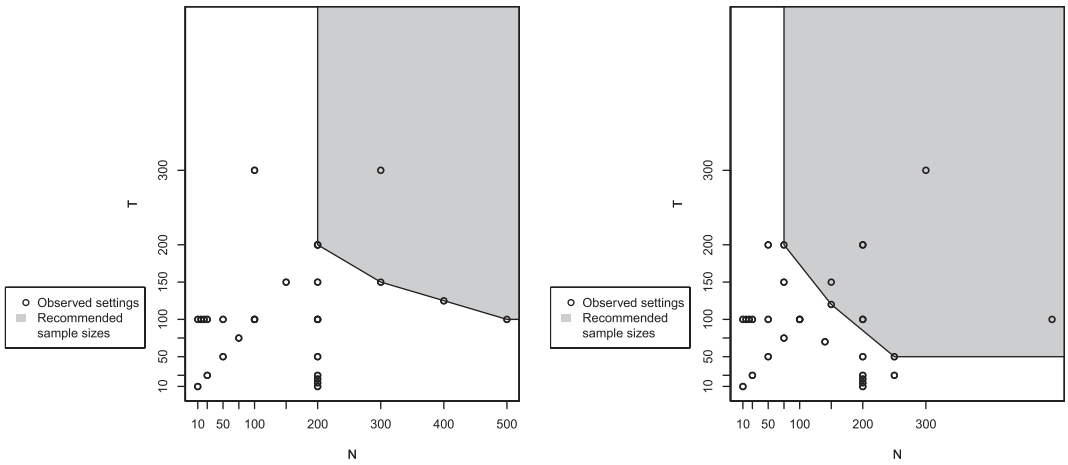
The shaded areas of Figures 19 and 20 show recommended sample sizes for Models 4 through 9 as a function of N and T. The small circles represent all combinations of N and T that have been investigated for the respective model. For instance, Figure 19c displays the recommendations for



(a) Model 4 (b) Model 5 (c) Model 6

(d) Model 7 (e) Model 8

FIGURE 19 Recommended sample sizes for the set of Models 4 through 8. The investigated combinations are marked with small circles. The shadowed areas are recommended sample sizes.



(a) Model 9, weak effects (b) Model 9, moderately strong effects

FIGURE 20 Recommended sample sizes for Model 9. The investigated combinations are marked with small circles. The shadowed areas are recommended sample sizes. The recommendations for the weak effects model do not include the slope of Z regressed on **W**.

Model 6. As can be seen in Figure 19c, good estimation quality can be attained by measuring 150 subjects 25 times, 75 subjects 75 times, or 50 subjects 150 times. For Model 9 with weak effects, displayed in Figure 20a, the slope of Z regressed on **W** did not attain these requirements for any of the considered sample sizes. The guidelines hold for all other parameters in Model 9. The guidelines for the moderately strong effects model, displayed in Figure 20b, do include the slope of Z regressed on **W.**

These plots confirm that a large N is more beneficial than a large T. This is clear from comparing, for example, Models 4 and 7, displayed in Figure 19a and Figure 19d. For Model 4, large T can compensate for N as small as 50. For the more complex Model 7, however, large T can no longer compensate for N smaller than 80. It is therefore interesting that a sufficiently large N allows for T as small as 10 for both these models. This pattern is also evident in general. For the plots of Figure 20, the shaded area tends to move to the right (larger N requirements) for more complex models, but only slightly upward (larger T requirements).

The effect sizes and R^2 in this study were chosen to give lower bounds for sample size in models with the smallest reasonable effects and R^2. This means that the guidelines are conservative but still apply for models with larger expected effect sizes and R^2. This is clear in the comparison of Model 9 with weak and moderately strong effects where the sample size requirements for the weak effect model give very good estimation quality for the corresponding moderately strong effects model.

ACKNOWLEDGMENTS

The authors gratefully acknowledge input and advice from Tihomir Asparouhov and Ellen Hamaker.

ORCID

Mårten Schultzberg http://orcid.org/0000-0002-1260-7737

REFERENCES

Asparouhov, T., Hamaker, E., & Muthén, B. (2017a). Dynamic latent class analysis. *Structural Equation Modeling*, *24*, 257–269. doi:10.1080/10705511.2016.1253479

Asparouhov, T., Hamaker, E., & Muthén, B. (2017b). *Dynamic structural equation models* (Tech. Rep.). Structural Equation Modeling. Advance online publication. doi: 10.1080/10705511.2017.1406803

Asparouhov, T., & Muthén, B. (2010a). *Bayesian analysis of latent variables models using Mplus* (pp. 1–60). Los Angeles, CA: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2010b). *Bayesian analysis using M*plus: *Technical implementation* (pp. 1–38). Los Angeles, CA: Muthén & Muthén.

Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford.

Bolger, N., & Schilling, E. A. (1991). Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. *Journal of Personality, 59*, 355–386. doi:10.1111/j.1467-6494.1991.tb00253.x

Chatfield, C. (2016). *The analysis of time series: An introduction (6th ed.)*. Boca Raton, FL: CRC.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cohen, L. H., Gunthert, K. C., Butler, A. C., Parrish, B. P., Wenze, S. J., & Beck, J. S. (2008). Negative affective spillover from daily events predicts early response to cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, *76*, 955–965.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). New York, NY: Taylor & Francis.

Hallquist, M. (2011). "MplusAutomation: Automating Mplus Model Estimation and Interpretation" R package. Retrived from https://github.com/michaelhallquist/MplusAutomation

Hamaker, E., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. O. (2017). DSEM of COGITO data. *Multivariate Behavior Research*. Retrieved from https://www.statmodel.com/TimeSeries.shtml

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

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. doi:10.1080/00273171.2014.1003772

Krone, T., Albers, C. J., & Timmerman, M. E. (2016). Comparison of estimation procedures for multilevel AR(1) models. *Frontiers in Psychology*, *7*(APR), 1–12. doi:10.3389/fpsyg.2016.00486

McAdams, D. P., & Constantian, C. A. (1983). Intimacy and affiliation motives in daily living: An experience sampling analysis. *Journal of Personality and Social Psychology*, *45*, 851–861. doi:10.1037/0022-3514.45.4.851

Muthén, B. (2010). *Bayesian analysis in M*plus*: A brief introduction*. Retrieved from www.statmodel.com

Muthén, L. K., & Muthén, B. O. (2017). *M*plus *user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Shiffman, S., & Waters, A. J. (2004). Negative affect and smoking lapses: A prospective analysis. *Journal of Consulting and Clinical Psychology*, *72*, 192–201. doi:10.1037/0022-006X.72.2.192

Trull, T. J., Solhan, M. B., Tragesser, S. L., Jahng, S., Wood, P. K., & Piasecki, T. M. (2008). Affective instability: Measuring a core feature of borderline personality disorder with ecological momentary assessment. *Journal of Abnormal Psychology*, *117*, 647–661. doi:10.1037/a0012532

Wang, L., & Preacher, K. J. (2015). Moderated mediation analysis using Bayesian methods. *Structural Equation Modeling*, *22*, 249–263.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301–322. doi:10.1037/a0016972

APPENDIX

Technical Details of the M*plus* Setup

All simulations are performed in M*plus* Version 8 using the built-in Monte Carlo procedure (Muthén & Muthén, 2017). The M*plus* input syntax for Model 9 is given in the next section.. The Bayes estimator is used with default noninformative priors (see Asparouhov & Muthén, 2010a, 2010b) for details. The parameter estimate is taken as the median of the posterior distribution of the parameters. Wang and Preacher (2015) found biases associated with the default median Bayesian point estimator in M*plus*. A comparison of several runs from this study raises no such problems for these models. The mean and median have very similar results and there is no apparent pattern in their small differences in favor of one of the estimators. Each cell of this study is replicated 500 times; using 1,000 replications did not change the results substantially. For all models at least 20 different combinations of N and T are investigated; these form the main focus of the study. For the set of Models 4 through 9, additional sample sizes are considered to better pinpoint the sample size requirements. For some models many more settings are investigated as well as different effect sizes, all described in detail in the article.

The MplusAutomation R package (Hallquist, 2011) was used to manage all simulations in this study. A minimum working example, with instructions on how to use the package, is available on request from the first author.

The biter = (5000); option is used if nothing else is stated. This option specifies that at least 5,000 iterations of the Gibbs algorithm will be performed. This means that 5,000 samples from the posterior distribution of the parameters will be generated. After 5,000 iterations, M*plus* will stop based on the Potential Scale Reduction (PSR) convergence criterion (Gelman et al., 2013). The first half is a so called burn-in phase and those iterations are discarded (Muthén, 2010) The second half of the iterations are used to calculate the estimates. With the default PROCESSORS = 2; option, which gives two MCMC chains, this means that 5,000 (2,500 + 2,500) draws are used to describe the posterior distribution. For some very small N and T values at least 10,000 are used. For T and N simultaneously larger than 200, at least 2,000 iterations are used. One potential risk with having too few iterations is the so-called premature stoppage problem, which occurs when the PSR becomes very close to one by chance rather than because the algorithm has converged. If more iterations are carried out, the PSR might suddenly increase again before converging. For all settings, multiple trace plots are inspected for sequences three times longer than used. No premature stoppage was detected. Sometimes replications are skipped due to large values in the time series in the generated data set, discussed further in the text. This occurred in at most 10 out of the 500 replications if nothing else is stated.

M*plus* Syntax for Model 9

```
TITLE: Model 9
MonteCarlo: NAMES ARE Y w Z;
            NOBS = 10000;
            NREP = 500;
            NCSIZES = 1;
            CSIZES = 200(50);
            LAGGED = Y(1);
            BETWEEN = w Z;
ANALYSIS: TYPE = TWOLEVEL RANDOM;
             ESTIMATOR = BAYES;
          PROCESSORS = 2;
          BITHER = (5000);
          BSEED = 9553;
MODEL POPULATION:
        %WITHIN%
    phi| Y on Y 1;
    logv| Y;

    %BETWEEN%
    w*0.119; [w*0];

    Y ON w*0.41;
```

```
    Y*0.07;[Y*0.5];

    phi ON w*0.31;
    phi*0.04;[phi*0.2];

    logv ON w*0.225;
    logv*0.02;[logv*-1.18];

    Z ON Y*0.5 phi*0.57 logv*0.75 w*0.15;
    Z*0.2; [Z*1];
MODEL:
    % WITHIN %
    phi| Y on Y 1;
    logv| Y;

    % BETWEEN %
    w*0.119; [w*0];

    Y ON w*0.41;
    Y*0.07;[Y*0.5];

    phi ON w*0.31;
    phi*0.04;[phi*0.2];

    logv ON w*0.225;
    logv*0.02;[logv*-1.18];

    Z ON Y*0.5 phi*0.57 logv*0.75 w*0.15;
    Z*0.2; [Z*1];

    logv WITH Y; logv WITH phi; phi WITH Y;
```