Original Article

Pushing the Limits

The Performance of Maximum Likelihood and Bayesian Estimation With Small and Unbalanced Samples in a Latent Growth Model

Mariëlle Zondervan-Zwijnenburg,[1] Sarah Depaoli,[2] Margot Peeters,[3] and Rens van de Schoot[1,4]

[1]Department of Methodology and Statistics, Utrecht University, The Netherlands
[2]Department of Psychological Sciences, University of California, Merced, CA, USA
[3]Department of Child and Adolescent Studies, Utrecht University, The Netherlands
[4]Optentia Research Focus Area, North-West University, South Africa

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Abstract: Longitudinal developmental research is often focused on patterns of change or growth across different (sub)groups of individuals. Particular to some research contexts, developmental inquiries may involve one or more (sub)groups that are small in nature and therefore difficult to properly capture through statistical analysis. The current study explores the lower-bound limits of subsample sizes in a multiple group latent growth modeling by means of a simulation study. We particularly focus on how the maximum likelihood (ML) and Bayesian estimation approaches differ when (sub)sample sizes are small. The results show that Bayesian estimation resolves computational issues that occur with ML estimation and that the addition of prior information can be the key to detect a difference between groups when sample and effect sizes are expected to be limited. The acquisition of prior information with respect to the smaller group is especially influential in this context.

Keywords: latent growth model, ML estimation, Bayesian estimation, informative priors

Many researchers in the social and behavioral sciences use latent growth modeling (LGM) to study development of individuals over time (e.g., Little, 2013). Within LGM, it is also possible to compare growth and the impact of variables on growth between different groups of individuals, for example, between a focal (i.e., small) group and a reference group. Researchers with this objective, however, often encounter two difficulties. In particular, the comparisons they want to make are between groups: (1) that have relatively different sample sizes, or (2) of which at least one is considered to be very small according to common guidelines for implementing the statistical model.

From the literature, we know that with traditional maximum likelihood (ML) estimation, the consequences of small sample sizes can include biased point estimates (Boomsma & Hoogland, 2001; Depaoli, 2013; Lee & Song, 2004; Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Meuleman & Billiet, 2009; van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & Van Loey, 2015), inadmissible estimates (Boomsma & Hoogland, 2001; Can, van de Schoot, & Hox, 2015; Hox & Maas, 2001; Meuleman & Billiet, 2009; Tolvanen, 2000), convergence issues (Boomsma & Hoogland, 2001; Hochweber & Hartig, 2017; Hox, Moerbeek, Kluytmans, & van de Schoot, 2014; Lüdtke et al., 2011), and inflated Type-I error rates (Boomsma & Hoogland, 2001; Hox & Maas, 2001; Hox et al., 2014; Lee & Song, 2004; Meuleman & Billiet, 2009).

There is, however, little known about the consequences of unbalanced samples (i.e., where sample sizes vary drastically across the subgroups being examined, e.g., 10 participants in the focal group vs. 500 in the reference group), especially when latent growth models are being implemented. We only know that unbalanced samples in LGM often result in low statistical power (Muthén & Curran, 1997), but its specific effect on coverage, biased point estimates, and estimation problems has not been thoroughly examined in the literature. Altogether, these issues may deter researchers from comparing the development of focal and reference groups in latent growth models.

Bayesian estimation is an alternative estimation method gaining in popularity (Kruschke, 2011; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). In Bayesian statistics, prior information is combined with the data in the analysis, resulting in a posterior distribution. The posterior distribution reflects probable parameter values given the prior information and the data. From the

© 2018 Hogrefe Publishing

https://doi.org/10.1027/1614-2241/a000162

posterior distribution, a measure of central tendency (i.e., the mean, median, or mode) is usually taken as a point estimate for the parameter of interest. Additionally, a 95% (credible) interval can be derived from the posterior distribution containing the most probable values for the parameter given the data. The frequentist 95% confidence interval, in contrast, will contain the true population value in 95% of the intervals over a long run of trials. To readers interested in a gentle introduction into Bayesian statistics for social scientists, we recommend Kruschke (2014) and van de Schoot et al. (2013).

In the current paper, we conduct a simulation study to evaluate the performance of maximum likelihood estimation and Bayesian estimation for latent growth models with small and unbalanced samples. The goal of the simulation is to highlight best practice when dealing with subgroup sizes that are quite different from one another.

Background on Sample Size Limits in LGM With ML and Bayesian Estimation

Muthén and Curran (1997) investigated the effect of unbalanced sample sizes in experimental designs on statistical power in LGM with sample size ratios varying from 1:1 (balanced) to 1:10. In general, Muthén and Curran (1997) found that the more extreme the sample size ratios were, the lower the statistical power to detect a difference between groups with ML estimation. When the ratio was more extreme than 1:5, even samples with 1,000 participants in total showed less than desirable power (< .80) to detect a small effect (Cohen's d = .20). Due to their focus on experimental designs, Muthén and Curran (1997) do not cover very small sample sizes, extreme sample size ratios, or the inclusion of covariates to limit the impact of confounders. No literature was found that covered aspects other than power under unbalanced sample sizes in LGM.

With respect to estimation in relation to total sample size for one group, estimates from ML estimation with a sample size as low as 50 do not substantially deviate from the population value (i.e., relatively unbiased) for means and factor loadings in LGM and related multilevel models (Hox & Maas, 2001; Maas & Hox, 2005; McNeish, 2016a, 2016b; Meuleman & Billiet, 2009; Tolvanen, 2000). Statistical power, however, is generally insufficient with samples smaller than 100 for the types of effect sizes commonly seen in empirical studies, and convergence issues also arise (Boomsma & Hoogland, 2001; Hochweber & Hartig, 2017; Hox & Maas, 2001; Lüdtke et al., 2011; Maas & Hox, 2005; McNeish, 2016a; Meuleman & Billiet, 2009; Tolvanen, 2000). Bayesian estimation does not have the same issues with small samples as ML estimation for two reasons. First, in Bayesian estimation, the results are determined by more than the data: Prior information is also included by means of prior distributions. Prior distributions can be based on information that a researcher has about parameters in the model *a priori*. When no information is available, so-called uninformative distributions can be adopted, which typically specify a very wide range of values for the parameter as probable. The more prior mass surrounding the population value, the better the model estimate will represent this value. Consequently, the non-null detection rate[1] is higher, and inference errors are less likely to occur (Depaoli, 2013; Lee & Song, 2004; van de Schoot et al., 2015).

The second reason Bayesian estimation does not have the same issues with small samples is that Bayesian estimation does not rely on asymptotic assumptions about sampling distributions akin to ML estimation (Asparouhov & Muthén, 2010). Depaoli (2013) shows in a growth mixture model that the use of uninformative priors as compared to ML estimation results in fewer problematically biased parameter estimates (i.e., bias $\geq$ 10%). When Bayesian estimation is used with an uninformative prior, a sample size of 20 already results in accurate estimates in a multilevel model (Hox, van de Schoot, & Matthijse, 2012; Hox et al., 2014). In addition, the coverage of the population value was better with Bayesian estimation, a result confirmed by van de Schoot et al. (2015) for repeated-measures analyses.

The Current Investigation

In order to ensure conditions were applicable to real data situations, the simulation study is inspired by an empirical dataset on the development rate of working memory in young heavy cannabis users versus their non-using peers. The data originate from 268 young adolescents enrolled in special education due to behavioral problems (Peeters, Monshouwer, Janssen, Wiers, & Vollebergh, 2014). To improve on the notion of causality, the development of both groups was corrected (by means of a time-invariant covariate) for the impact of quantity and frequency of alcohol use at the start of the study, as recommended by Jacobus, Bava, Cohen-Zion, Mahmood, and Tapert (2009). With this simulation set up, we aimed to compare and establish sample size requirements to evaluate a small difference in development between groups for ML and Bayesian estimation when one of the groups has a sample size below 50.

By means of the simulation, we compare the sample size requirements to evaluate a small difference in development between groups for ML and Bayesian estimation. Regarding Bayesian estimation, the balance between sample size

[1] Statistical power is a frequentist term that involves the null hypothesis. Since the null hypothesis does not exist in Bayesian statistics, we refer to the non-null detection rate instead.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

 © 2018 Hogrefe Publishing

requirements and the required specificity of prior information is investigated as well. Additionally, we explore how the results are affected when a substantial amount of prior information can be found for the reference group but not for the focal group. It can be expected that prior information with respect to a focal group is harder to obtain.

Method

To compare the performance of ML estimation and Bayesian estimation in the evaluation of small and unbalanced samples in a latent growth model, we conducted a simulation study. In this section, we elaborate on the model of interest, the main characteristics of the simulation study, and the evaluation criteria.

The Latent Growth Model

Figure 1 displays the latent growth model as applied in this study. The model has four observed variables ($y_1^g - y_4^g$) representing repeated measures of the same construct. In the empirical data, this construct is performance on a working memory task expressed in percentages. The repeated measures are the indicators for the intercept, linear slope, and quadratic slope latent variables. The linear growth factor in this model represents the growth rate at one time point (typically the first time point). The model has one covariate representing an observed time-invariant predictor, which is a measure of alcohol use quantity and frequency at the start of the study in the empirical data. As a result, the latent time variables technically have intercepts instead of means. However, to avoid confusion between the intercept growth factor and the intercepts of the latent growth factors, the latter will be referred to as being "means" throughout the paper.

In order to assess the growth rate difference between groups, a new parameter (denoted by Δα) was constructed by subtracting the linear slope mean of group 2 (i.e., the focal group) from that of group 1 (i.e., the reference group). A Monte Carlo study was conducted in M*plus* version 7.11 (Muthén & Muthén, 1998–2012) directed by the R-package MplusAutomation (Hallquist, 2013) in R 3.0.1 (R Core Team, 2015). To promote transparency and replicability, population syntax files are provided in Appendix A, and all input and output files are available at the project page https://osf.io/gjzu8.

Simulation Study Design

The population parameters originated from multiple group latent growth analyses (see Appendix A and https://osf.io/ttybt) on empirical data. The difference between the linear

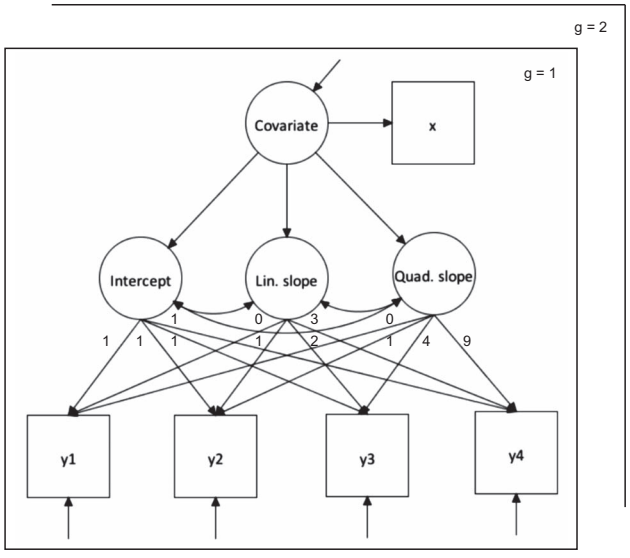

Figure 1. Multiple group latent growth model with one covariate and groups indicated by g. y_1^g, y_2^g, y_3^g, and y_4^g represent four assessments of a developing construct with residual error variances. x^g is a time-invariant predictor of growth that represents the latent variable Covariate^g without measurement error. The regressions of the latent growth factors Intercept^g, Lin. slope^g, and Quad. slope on the Covariate^g are equal over groups.

slope factors, Δα, was set at 1.60, while the disturbance of the linear slope factors was 64.00 in order to represent a small effect size ($\frac{1.60}{64.00}$ = .20 Cohen's *d*; Cohen, 1988), which is considered a realistic effect size for this parameter (see, for instance, Jacobus et al., 2009).

For this population, we varied the sample sizes in the reference group, the sample sizes in the focal group, and the estimation settings. The sample sizes for the reference group were ∈ {50, 100, 200, 500, 1,000, 2,000, 5,000, 10,000}, which represents a wide range of sample sizes commonly specified in the empirical and methodological literature. The sample sizes for the focal group were 5, 10, 25, and 50. Consequently, the sample size ratios ranged from 1:1 to 1:2,000. The estimation methods were ML estimation and Bayesian estimation.

ML estimation was applied with standard errors robust to non-normality and nonindependence of observations (MLR), which suits analyses with repeated measures. M*plus* uses accelerated expectation maximization (EMA) to obtain the ML estimates (Muthén & Muthén, 1998–2012). The ML output shows one extra parameter compared to the exact same Bayesian specification. This "knownclass" parameter, however, is not estimated. Therefore, we consider the models to be exactly equal.

Bayesian estimation was implemented with seven different prior distribution settings for the means of the latent growth factors. Normally distributed informative priors were specified for the latent growth factor means, because it was considered most likely that researchers would have

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

knowledge about these parameters before analyzing their data. Theoretically, however, prior information can be found for all parameters. The more appropriate the information being included in the prior is, the more accurate the parameter estimates will be. All user-specified priors were normally distributed with mean μ_0 and variance σ_o^2. The population values of the growth factor means[2] were used as prior means to understand the upper-bound performance of Bayesian methods under these modeling circumstances. The prior variances σ_o^2 ranged from 0.1 (i.e., very informative) to 10^{10} (i.e., uninformative). Specifically, $\sigma_o^2 \in \{0.1, 0.3, 0.5, 1.0, 2.0, 5.0, 10^{10}\}$. M*plus* default priors were used for the other parameters in the model. Specifically:

- A normal distribution with a mean of 0 and variance of 10^{10} for the mean of the covariate and the regression coefficients,
- An improper inverse gamma with the shape parameter set at −1, and the scale at 0 for the variance of the covariate and the residuals of the observed variables,
- An improper inverse Wishart with 0 forming the scale matrix, and −4 degrees of freedom for the covariances and disturbances of the growth factors.

Furthermore, 22 Markov chains were used for the Bayesian analyses to capture the impact of many different starting values. Note that 22 chains may be excessive in other modeling contexts due to the length of time it would take to obtain convergence. We were able to have the large number due to the computational capacity that was available to us. It is important to note that methods and results described here using these 22 chains are generalizable to situations requiring fewer chains. In order to assess convergence, it is recommended that at least two chains are used (Gelman & Rubin, 1992). The minimum number of iterations (or samples) in the chain was set at 5,000, and the maximum was set at 100,000. The first half of the chain was discarded as burn-in, and the second half was used to construct the posterior (Muthén & Muthén, 1998–2012).

Convergence was assessed using the Gelman–Rubin potential scale reduction factor (PSRF; Gelman & Rubin, 1992). When the PSRF was less than 0.05 points away from 1 for all parameters in the second half of the iterations, the model was considered to be converged. Syntax for the analyses is provided in Appendix B and at https://osf.io/qwf3r. Altogether, the number of cells in the simulation study was 4 (focal group sample sizes) × 8 (reference group sample sizes) × 8 (estimation settings: 1 × ML + 7 × Bayes with varying σ_o^2) = 256.

The simulation was extended with additional Bayesian analyses to investigate what would happen if a substantial amount of prior information (specified as having a variance hyperparameter of $\sigma_o^2 = 0.1$, indicating a great deal of precision in the prior) could only be obtained for the reference group, but not for the focal group (with a variance hyperparameter of $\sigma_o^2 = 10.0$, indicating less precision in the normal prior). In the focal group, σ_o^2 was set at 10.0 instead of 10^{10} (the M*plus* default) because, even when prior information is hard to find, researchers and experts are generally able to estimate its value to some extent. We investigated the effects of these conditions for the largest (i.e., best performing) focal group (n = 50). The sample size of the reference group was again manipulated for this additional condition examined. Input for this analysis is located at https://osf.io/xm3v5/.

Evaluation

Since the main interest in multiple group LGM is to compare development between groups, the growth rate difference parameter Δα was the parameter of interest in the simulation study. For the Bayesian cells in the design, the median of the posterior distribution was interpreted as the point estimate. Credible intervals were obtained by the equal tail method, having tails on both sides that each contains 2.5% of the posterior distribution (Muthén & Muthén, 1998–2012).

The difference parameter Δα was evaluated in terms of proportion of bias, coverage, statistical power or non-null detection rates, and estimation problems. The proportional bias was calculated by dividing the average bias over the analyzed datasets by the value of the population estimate. A proportional bias lower than .10 was considered acceptable (Muthén & Muthén, 2002). Coverage is the rate of 95% confidence intervals (frequentist statistics captured through the ML estimation condition) or credible intervals (Bayesian statistics) that cover the population parameter estimate. For a 95% confidence or credible interval, coverage should be around the advocated 95%. In the current study, a minimum level of .90 was considered acceptable. Statistical power and non-null detection rates were calculated as the percentage of replications in which the 95% interval for Δα did not include zero. The acceptable minimum level of statistical power or the non-null detection rates was considered to be .80 (Muthén & Muthén, 2002). The last criterion concerned estimation problems. Estimation problems arise when the following occur: (1) negative variances, (2) correlations larger than one, (3) linear dependencies among more than two latent variables are

[2] That is, 73.05, 71.54, 8.13, 6.53, and −2.16 for $\text{Intercept}_{\text{non-users}}$, $\text{Intercept}_{\text{users}}$, Lin. slope, Lin. $\text{slope}_{\text{non-users}}$, and Quad. slope, respectively.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

 © 2018 Hogrefe Publishing

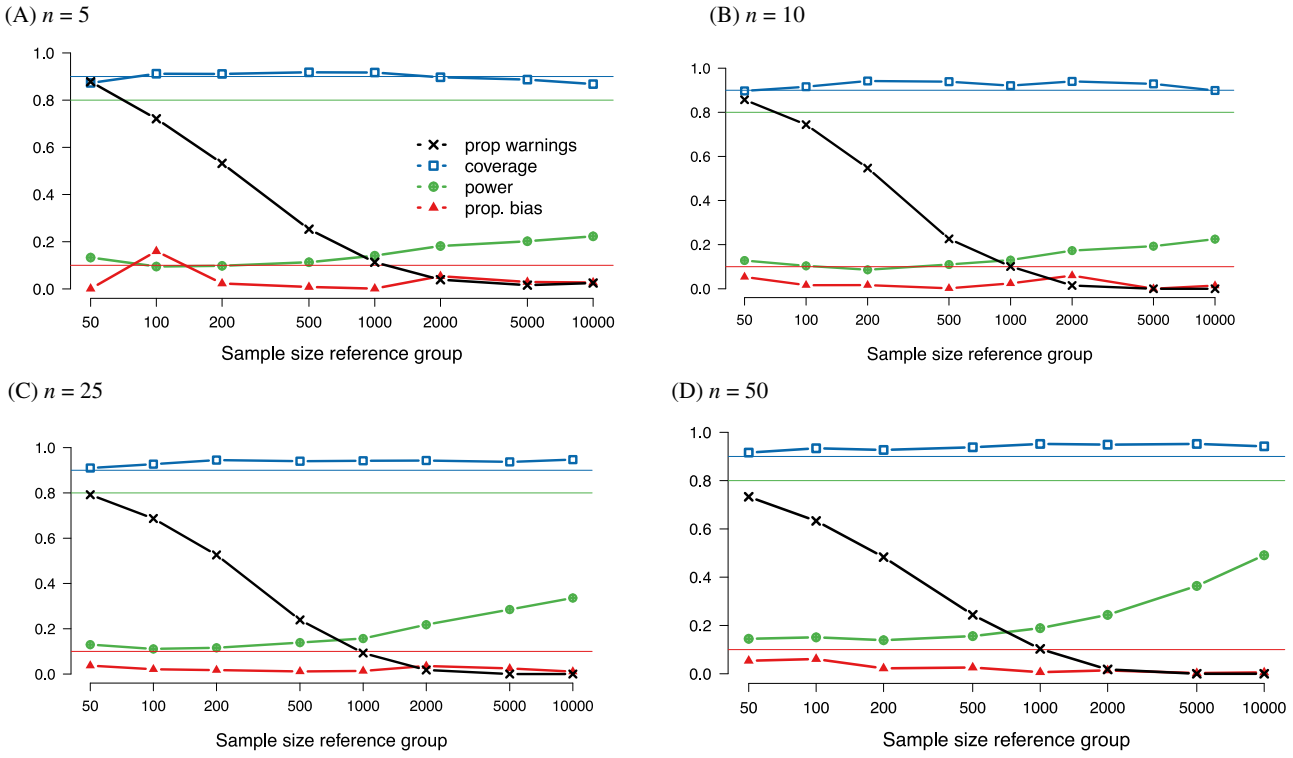

Figure 2. Results for ML estimation by focal group sample size. On the *x*-axis, the size of the reference group increases. From top to bottom, the static horizontal lines represent (1) the minimum acceptable value for coverage (i.e., .90), (2) the minimum acceptable value for statistical power (i.e., 0.80), and (3) the maximum acceptable value for proportional bias (i.e., 0.10).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

estimated, or (4) when the model does not converge. When using ML estimation, M*plus* notifies the user when one of these problems occurred. The proportion of datasets for which M*plus* produced warnings in this respect was used as an evaluation of estimation problems. Bayesian estimation cannot result in illegitimate estimates with the prior distributions used in this study. However, non-convergence can occur, and this can be detected by warnings and/or by visual inspection of the trace plots. Therefore, for every cell in the simulation design, two sets of trace plots were randomly selected and inspected for potential issues with convergence.

Results

Maximum Likelihood Estimation

Figure 2 shows the ML results in terms of proportion of warnings, coverage, statistical power, and proportional bias for the four focal group sample sizes separately. As can be seen, the proportion of bias was adequate for all combinations of sample sizes, except for a focal group sample size of 5 combined with a reference sample of 100 (Figure 2A). Coverage was in general lower than .95, but always sufficient when the focal sample contained at least 25 participants (Figures 2C and 2D). With sample sizes in the focal group of 5 and 10, reference group sample sizes at both extreme ends did not cover the population value often enough in the 95% confidence intervals (coverage < .90), even though the average relative bias over datasets was acceptable (Figures 2A and 2B). Truly worrisome, however, were the statistical power and the proportion of warnings. Even with 10,000 participants in the reference group, the power to detect a small effect was lower than .50 for all focal groups, while a minimum of .80 is pursued. The proportion of warnings with a reference group sample size of 50 ranged from .73 to .88[3]. These warnings concerned illegitimate estimates, which make the results of the analysis unreliable. Examples of warnings that were obtained for ML models with estimation issues were as follows:

[3] A check with the lavaan R-package (Rosseel, 2012) instead of M*plus* for the focal group with 5 participants resulted in even more convergence issues.

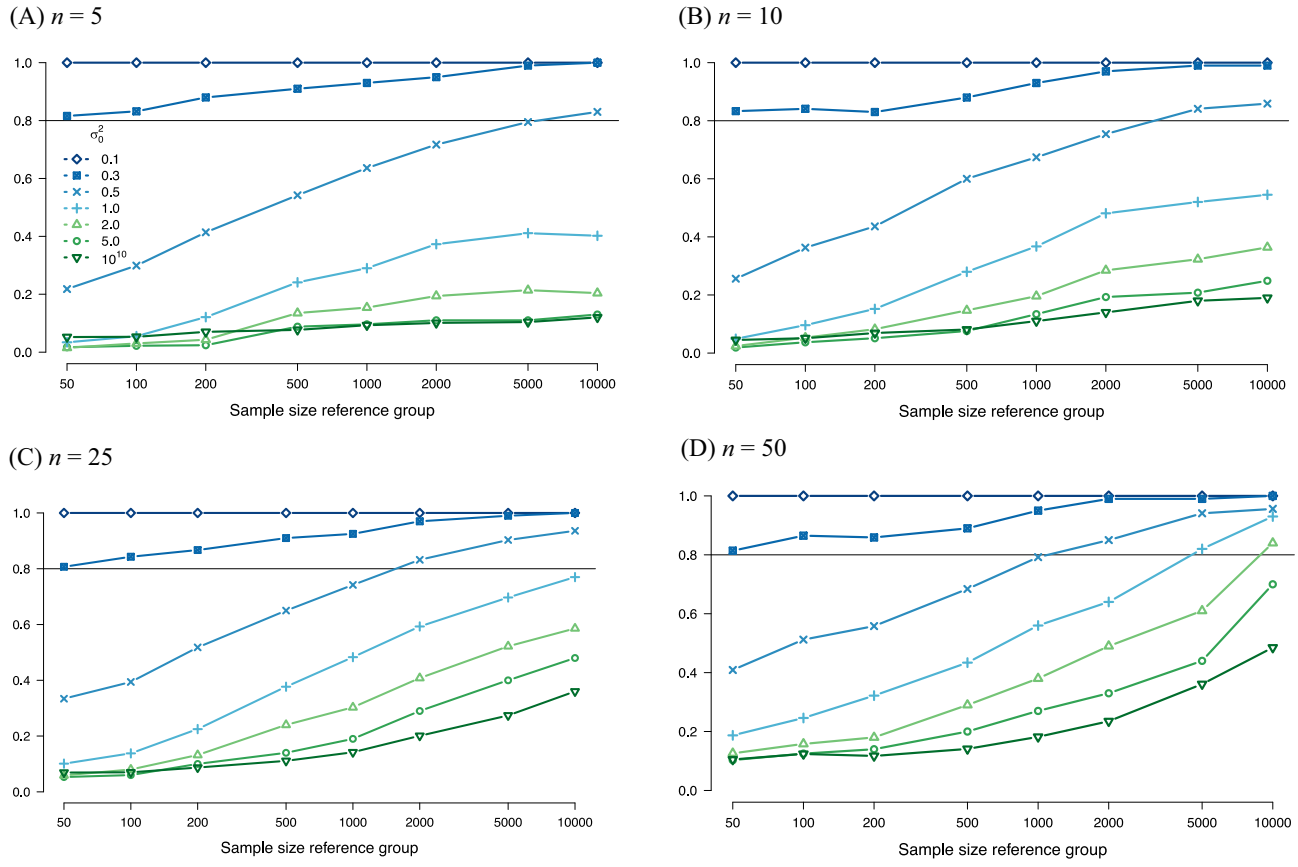

Figure 3. Non-null detection rate for Bayesian estimation by focal group sample size. On the *x*-axis, the size of the reference group increases. The *y*-axis represents the non-null detection rate. The static horizontal line represents the minimum acceptable value for (i.e., 0.80). The remaining lines reflect the results for varying σ_0^2.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

```
THE MODEL ESTIMATION TERMINATED NORMALLY

WARNING: THE RESIDUAL COVARIANCE MATRIX
(THETA) IS NOT POSITIVE DEFINITE. THIS
COULD INDICATE A NEGATIVE VARIANCE/RESI-
DUAL VARIANCE FOR AN OBSERVED VARIABLE,
A CORRELATION GREATER OR EQUAL TO ONE
BETWEEN TWO OBSERVED VARIABLES, OR A LIN-
EAR DEPENDENCY AMONG MORE THAN TWO
OBSERVED VARIABLES. CHECK THE RESULTS
SECTION FOR MORE INFORMATION.

WARNING: THE LATENT VARIABLE COVARIANCE
MATRIX (PSI) IS NOT POSITIVE DEFINITE.
THIS COULD INDICATE A NEGATIVE VARIANCE/
RESIDUAL VARIANCE FOR A LATENT VARIABLE,
A CORRELATION GREATER OR EQUAL TO ONE
BETWEEN TWO LATENT VARIABLES, OR A LINEAR
DEPENDENCY AMONG MORE THAN TWO LATENT
VARIABLES. CHECK THE TECH4 OUTPUT FOR
MORE INFORMATION.
```

Bayesian Estimation

With Bayesian estimation, bias and coverage were acceptable for every cell of the simulation design. Plots for all cells can be found at https://osf.io/s59cz. In addition, Bayesian estimation showed decent convergence. As a result, the remaining aspect of interest was statistical power. Figure 3 shows for all four focal group sample sizes (i.e., n = 5, 10, 25, and 50) how many participants are in the reference group and how much prior information is necessary to obtain satisfactory non-null detection rates. With uninformative priors imposed on all parameters (i.e., $\sigma_0^2 = 10^{10}$), non-null detection rates were insufficient, regardless of the sample size in the reference group. The same held when the variances of the priors for the latent growth factor means were decreased to 5.0. An exploration of the non-null detection rate with a focal group of 100 and the prior variance of the latent growth factor means at 5.0 showed an improvement in the non-null detection rate, but still about 10,000 participants in the reference group were needed, to acquire a non-null detection rate close to .80. Prior

 © 2018 Hogrefe Publishing

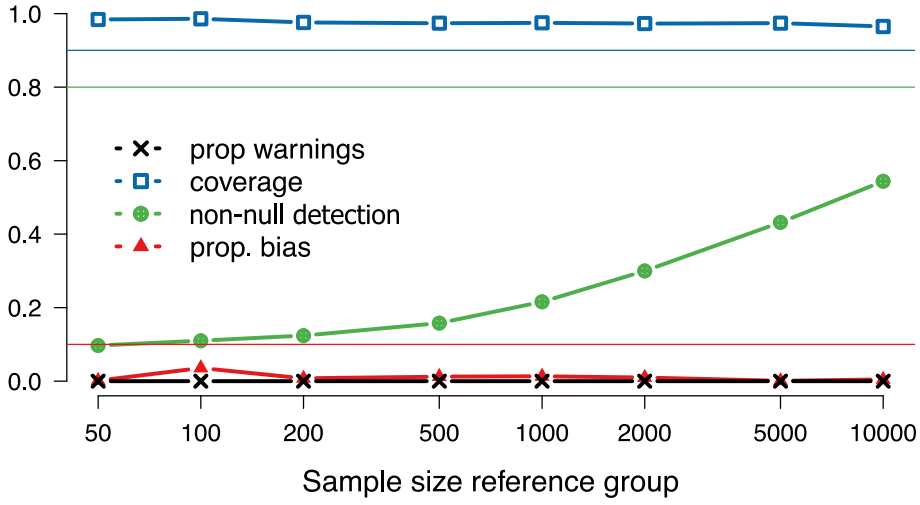

Figure 4. Results for Bayesian estimation with unbalanced prior information. σ_0^2 for latent growth factors in reference group = 0.1. σ_0^2 for latent growth factors in focal group = 10. Focal group n = 50.

variances as specific as 0.1, on the other hand, resulted in a non-null detection rate of 1.0 for every cell.

Unbalanced Prior Information

The simulation results presented in the previous section show that a focal group of 50 participants combined with a prior variance of 0.1 can lead to an optimal situation in all respects assessed (Figure 3). Figure 4 shows that when prior information is scarce for the focal group ($\sigma_0^2 = 10$), the non-null detection rate is an issue again. Additional analyses showed that no matter how much the prior variance in the reference group was decreased, a satisfactory non-null detection rate could not be achieved as long as the prior variance in the focal group was 10. Due to these clear results, the effect of unbalanced prior information was not further investigated for cells with focal groups smaller than 50.

Conclusion

The aim of the simulation study was to investigate lower-bound sample size issues in a multigroup LGM context, especially when one group is much smaller than the others. We set up the simulation in this way in order to compare and establish sample size requirements to evaluate a small difference in development between groups for ML and Bayesian estimation when one of the groups has a sample size not larger than 50.

The results showed that ML estimation has issues with statistical power when at least one of the groups is not larger than 50. Moreover, with ML estimation, analyses based on small sample datasets generally cannot be properly interpreted because of nonpositive definite matrices that yield inadmissible estimates.

By adopting Bayesian estimation, the issue of non-interpretable output disappears and consequently smaller samples can be analyzed. Bayesian inference with uninformative as well as minimally informative priors, however, has non-null detection rate issues similar to ML estimation. Specifically, even comparison groups with 10,000 participants do not yield satisfactory non-null detection rates for a small effect. To obtain a satisfactory non-null detection rate in the context of limited small and unbalanced sample sizes, Bayesian estimation is necessary in combination with the availability of very specific prior information. This may seem trivial to those who are familiar with the Bayesian concept, but the current simulation study provided additional insight to the effect of prior information by showing the consequences of specific degrees of informativeness. Note, however, that our use of an empirical model with empirical population values limits the direct applicability of the simulation results to other research situations. The simulation results are only directly indicative for other researchers under specific circumstances. The statistical model needs to be equal (e.g., a latent growth model including a time-invariant covariate, a multiple group confirmatory factor model with a covariate, or a multiple indicators multiple causes model with the groups as a covariate), the expected effect size small, and the growth rate difference needs to be comparable or proportional after taking the impact of the covariate into account. When the growth rate is proportional, the impact of the prior variances is proportional as well. If these circumstances do not hold, the presented simulation results are mainly useful as inspiration for new simulation efforts.

As was shown by the simulation study with unbalanced prior information, highly informative priors are particularly necessary for the focal group. To be able to specify such informative priors, the available prior information must be

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

very specific and convincing. This, however, may be seldom feasible because of the exceptionality of the group. In such a situation, we advise researchers to publish their updated estimates and data nevertheless. Such a publication provides a future researcher on the topic with more prior information, and over time, the amount of prior information can be sufficient to draw conclusions about the effect under study. Thus, when separate analyses cannot obtain sufficient power to make inferences, cumulative efforts of researchers can overcome the issue.

Cautionary Points Regarding Bayesian Estimation

To avoid misinterpretations of this study, we hereby provide a disclaimer. The goal of Bayesian analyses with informative priors is to make optimal use of all available information. Accordingly, the simulation study shows the relation between the amount of prior information and results in terms of estimation and the non-null detection rate. With this information, researchers can observe the relation between the specificity of prior information and other factors such as estimation problems, bias, non-null detection rate, and coverage. This paper is not a demonstration of how prior distributions should be manipulated to secure statistically significant results: This would not be an ethical use of the information, and the exact results may vary between study variables and models. As shown in Zondervan-Zwijnenburg, Peeters, Depaoli, and van de Schoot (2017), prior knowledge has to be acquired systematically and specifications of prior distributions have to be justified. Moreover, to promote transparency, we advise to demonstrate the impact of other priors on the results by means of a sensitivity analysis (see also Depaoli & van de Schoot, 2015). We believe that the manipulation of priors to obtain a "desirable" result would fall under unethical research practices.

Another cautionary note should be made on the use of default priors for variance parameters with small samples. Variance and disturbance parameters were not the focus of this study, but it has been shown, for example, by McNeish (2016a) and van de Schoot et al. (2015) that these estimates can be severely biased with uninformative priors.

Final Recommendations

Based on these findings, we recommend researchers with focal groups with fewer than 200 participants to conduct a simulation study in order to evaluate the impact of the small sample on estimation issues, bias, coverage, and non-null detection rate.

When maximum likelihood estimation cannot generate proper output under the circumstances of interest, we suggest to obtain prior information. Zondervan-Zwijnenburg et al. (2017) provide guidelines on collecting and including prior information. If sufficiently precise prior information can be acquired, the data can be analyzed. If the researcher is not able to meet the requirements, simpler models (e.g., descriptive statistics, case studies), waiting until more prior information and participants become available (e.g., by following Google Scholar Alerts, RSS feeds, and reapproaching schools in a new academic year), or conducting the analysis to contribute to cumulative science without making inferences, are alternative ways to deal with the data.

References

Asparouhov, T., & Muthén, B. O. (2010, September). *Bayesian analysis of latent variable models using Mplus*. Retrieved from https://www.statmodel.com/techappen.shtml

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, K. G. Jöreskog, & D. Sörbom (Eds.), *Structural equation models: Present and future. A festschrift in honor of Karl Jöreskog* (pp. 139–168). Lincolnwood, IL: Scientific Software International.

Can, S., van de Schoot, R., & Hox, J. (2015). Collinear latent variables in multilevel confirmatory factor analysis a comparison of maximum likelihood and Bayesian estimations. *Educational and Psychological Measurement, 75*, 406–427. https://doi.org/10.1177/0013164414547959

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum. https://doi.org/10.4324/9780203771587

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods, 18*, 186. https://doi.org/10.1037/a0031609

Depaoli, S., & van de Schoot, R. (2015). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods, 22*(2), 240–261. https://doi.org/10.1037/met0000065

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472. https://doi.org/10.1214/ss/1177011136

Hallquist, M. (2013, October). *MplusAutomation: Automating Mplus model estimation and interpretation. Package MplusAutomation*. Retrieved from https://cran.r-project.org/web/packages/MplusAutomation/MplusAutomation.pdf

Hochweber, J., & Hartig, J. (2017). Analyzing organizational growth in repeated cross-sectional designs using multilevel structural equation modeling. *Methodology, 13*, 83–97. https://doi.org/10.1027/1614-2241/a000133

Hox, J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8*, 157–174. https://doi.org/10.1207/S15328007SEM08021

Hox, J., Moerbeek, M., Kluytmans, A., & van de Schoot, R. (2014). Analyzing indirect effects in cluster randomized trials. The effect of estimation method, number of groups and group sizes on accuracy and power. *Frontiers in Psychology, 5*, 78. https://doi.org/10.3389/fpsyg.2014.00078

Hox, J., van de Schoot, R., & Matthijse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

 © 2018 Hogrefe Publishing

perspective. *Survey Research Association, 6*, 87–93. https://doi.org/10.18148/srm/2012.v6i2.5033
Jacobus, J., Bava, S., Cohen-Zion, M., Mahmood, O., & Tapert, S. (2009). Functional consequences of marijuana use in adolescents. *Pharmacology, Biochemistry and Behavior, 4*, 559–565. https://doi.org/10.1016/j.pbb.2009.04.001
Kruschke, J. K. (2011). Introduction to special section on Bayesian data analysis. *Perspectives on Psychological Science, 6*, 272–273. https://doi.org/10.1177/1745691611406926
Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). San Diego, CA: Academic Press.
Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*, 653–686. https://doi.org/10.1207/s15327906mbr3904_4
Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.
Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16*, 444. https://doi.org/10.1037/a0024376
Maas, C. J., & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86–92. https://doi.org/10.1027/1614-1881.1.3.86
McNeish, D. M. (2016a). On using Bayesian methods to address small sample problems. *Structural Equation Modeling, 23*, 750–773. https://doi.org/10.1080/10705511.2016.1186549
McNeish, D. M. (2016b). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using Mplus. *Journal of Educational and Behavioral Statistics, 41*, 27–56. https://doi.org/10.3102/1076998615621299
Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods, 3*, 45–58. https://doi.org/10.18148/srm/2009.v3i1.666
Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*, 371–402. https://doi.org/10.1037/1082-989X.2.4.371
Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599–620. https://doi.org/10.1207/S15328007SEM0904_8
Peeters, M., Monshouwer, K., Janssen, T., Wiers, R. W., & Vollebergh, W. A. (2014). Working memory and alcohol use in at-risk adolescents: A 2-year follow-up. *Alcoholism: Clinical and Experimental Research, 38*, 1176–1183. https://doi.org/10.1111/acer.12339
R Core Team. (2015). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36. Retrieved from https://www.jstatsoft.org/article/view/v048i02
Tolvanen, A. (2000). *Latenttien kasvukäyrä- ja simplex-mallien teoriaa ja sovelluksia pitkittäisaineistoissa kehityksen ja muutoksen analysointiin* [Latent growth and simplex models: Theory and applications in longitudinal models for analysis of development and change]. Jyväskylä, Finland: Department of Statistics, University of Jyväskylä.
van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & Van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology, 6*, 25216. https://doi.org/10.3402/ejpt.v6.25216
van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2013). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development, 85*, 842–860. https://doi.org/10.1111/cdev.12169
van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*, 217–239. https://doi.org/10.1037/met0000100
Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., & van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development, 14*, 305–320. https://doi.org/10.1080/15427609.2017.1370966

Received February 8, 2017
Revision received August 8, 2018
Accepted September 27, 2018
Published online December 12, 2018

Funding
The first author Marielle Zondervan-Zwijnenburg, has been supported by the Consortium Individual Development (CID), which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003). Rens van de Schoot has been supported by Grant NWO-VIDI-452-14-006 from the Netherlands Organization for Scientific Research (NWO).

Mariëlle Zondervan-Zwijnenburg
Department of Methods and Statistics
Utrecht University
Padualaan 14
3584 CH Utrecht
The Netherlands
m.a.j.zwijnenburg@uu.nl

Mariëlle Zondervan-Zwijnenburg is a PhD candidate at the Department of Social and Behavioural Sciences at Utrecht University. Her focus is especially on the dynamics of youth, and her additional interests are Bayesian statistics, longitudinal data analysis, and structural equation modeling.

Sarah Depaoli (PhD, 2010) is an associate professor of Quantitative Psychology at the University of California, Merced. Her research interests are largely focused on issues surrounding Bayesian estimation of latent variable models.

Margot Peeters (PhD, 2014) is an assistant professor at Utrecht University. Her research interests are adolescent development, behavioral control, risk behavior (alcohol, drug use, externalizing problems, gaming), addiction, social environment and risk behavior (peers, SES).

Rens van de Schoot (PhD) is an associate professor at Utrecht University and extraordinary professor at the Optentia research program, North-West University in South Africa. His interests are Bayesian statistics, longitudinal data analysis, Mplus, multilevel analysis, PTSD, and structural equation modeling.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix A

Population Parameters

The text below shows the input file used to generate the datasets for the simulation study for the specific case with 50 participants in the reference group and 5 in the focal group. For other group specifications, the nobs syntax was changed accordingly. The code is annotated with text after the exclamation mark.

The covariate is simulated as a count variable, because this fitted the empirical data best. It was analyzed as a normally distributed variable though, because (1) the scale of an exogenous variable does not affect the regression coefficients, and that is important, (2) the predictor itself was not the variable of interest, (3) Bayesian analysis in M*plus* (7.1) cannot handle count variables, and M*plus* provides a lot of possibilities for our analyses that are more important. (4) This is common practice in the social and behavioral sciences.

The variance of the covariate, however, was allowed to differ between groups, because this fitted the empirical data best. The empirical analysis including a quadratic factor had a better fit than without the quadratic factor (see the files named Bayes2group.out and Bayes2groupISonly.out, respectively, at https://osf.io/gjzu8; DIC = 6861.396 vs. 6892.445, BIC = 6948.434 vs. 6960.688). We constrained Q over groups so that the difference between groups is represented in the difference between the linear slopes.

```
MONTECARLO:
names = y1-y4 qft;                !variable names
count = qft;                      !count variable
generate = qft(c);                !create count variable
ngroups = 2;                      !2 groups
nobs = 50 5;                      !50 in reference group, 5 in focal
nreps = 1000;                     !produce 1000 datasets from the population input
seed = 4533;
repsave = all;
save = mc_5_50_*.dat;             !name for data files

ANALYSIS:
type = mixture;
algorithm = integration;
processors = 2;

MODEL POPULATION:
%OVERALL%                         !overall set up with group invariant and g=1 values
i s q | y1@0 y2@1 y3@2 y4@3;      !Intercept, Linear Slope, Quadratic Slope LGM syntax
i ON qft*-0.101;                  !Beta_141
s ON qft*-0.228;                  !Beta_24
q ON qft*0.131;                   !Beta_34
i WITH s*-53.669 q*12.342;        !covariance I with LS Psi_21, I with QS Psi_312
s WITH q*-14.052;                 !covariance LS with QS Psi_32
[qft*0.313];                      !Quantity frequency alcohol use, count parameter lambda3
[i*73.050 s*8.125 q*-2.161];      !means I (alpha_1^1), LS (alpha_2^1), QS (alpha_3)
i*67.887; s*64q*3.958;            !residual variances I (zeta_1), LS (zeta_2), QS (zeta_3)
y1*52.956 y2*64.049 y3*55.481 y4*19.390;
                                  !residual variances y_1^1 - y_1^1 (epsilon_1^1-epsilon_4^1)
%g#1%                             !values reference group (g=1)
[qft*0.313];
[i*73.050 s*8.125 q*-2.161];
%g#2%                             !values focal group (g=2), overwrite overall set up
[qft*2.704];
[i*71.541 s*6.525 q*-2.161];
```

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

 © 2018 Hogrefe Publishing

Population values for β are based on a Bayesian analysis with default settings of the latent growth model as depicted in Figure 1. The .inp syntax and .out output files named "Bayes equal q var regress" are provided at https://osf.io/gjzu8. Population values for the covariances, disturbances, and intercepts are based on a Bayesian analysis with default settings, but with the regression parameters estimated for both groups separately. The .inp syntax and .out output files named "Bayes equal q" and "equal var" are provided at https://osf.io/gjzu8.

Population values for the count variable are based on the results of a nonpositive definite ML analysis of the latent growth model, because only with these settings, M*plus* could estimate the values for a count variable. The .inp syntax and .out output files named "ML all par" are provided at https://osf.io/gjzu8.

Algorithm = integration was necessary to create the count data and to regress the latent variables on the count variable. With mixture (i.e., knownclass) analyses, M*plus* uses EMA optimization. With a grouping specification, M*plus* does not do this. Hence, the results can differ.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix B

Syntax Analyses

Both syntax files concern the simulated data for the cell with 5 participants in the focal group, and 5 participants in the reference group. Logically, syntax for other cells included different datafile lists.

ML Estimation

For ML estimation, `Algorithm = integration` was necessary to obtain convergence.

```
DATA:     FILE = "mc_5_50_list_1.dat";
          TYPE = MONTECARLO;
VARIABLE: NAMES = QFT Y1 Y2 Y3 Y4 G;
          CLASSES = cg(2);
          KNOWNCLASS is cg(g=1 g=2);
ANALYSIS: TYPE = mixture;
          ALGORITHM = integration;
          PROCESSORS = 4;
MODEL:
%OVERALL%
i s q | y1@0 y2@1 y3@2 y4@3;
i ON qft*-0.101;
s ON qft*-0.228;
q ON qft*0.131;
i with s*-53.669 q*12.342;
s with q*-14.052;
[qft*0.313];
qft;
[i*73.050 s*8.125 q*-2.161];
i*67.887; s*64 q*3.958;
y1*52.956 y2*64.049 y3*55.481 y4*19.390;
%cg#1%
i s q | y1@0 y2@1 y3@2 y4@3;
[qft*0.313];
qft;
[i*73.050 s*8.125 q*-2.161] (I1 S1 Qg);
%cg#2%
```

```
i s q | y1@0 y2@1 y3@2 y4@3;
[qft*2.704];
qft;
[i*71.541 s*6.525 q*-2.161] (I2 S2 Qg);
MODEL CONSTRAINT:
NEW(diff_s)*1.6;
diff_S = S1 - S2;
OUTPUT: TECH9;
```

Bayesian Estimation

The syntax below concerns the analyses in which the informative priors had a variance of 1.0. Other cells had a different value for the variance of the prior under MODEL PRIORS.

```
DATA:     FILE = "mc_5_50_list_1.dat";
          TYPE = MONTECARLO;
VARIABLE: NAMES = QFT Y1 Y2 Y3 Y4 G;
          CLASSES = cg(2);
          KNOWNCLASS is cg(g=1 g=2);
ANALYSIS: TYPE = mixture;
          ESTIMATOR = BAYES;
          BCONVERGENCE = .05;
          Chains=22;
          Processors=22;
          Biterations=(5000) 100000;
MODEL:
%OVERALL%
i s q | y1@0 y2@1 y3@2 y4@3;
i ON qft*-0.101;
s ON qft*-0.228;
q ON qft*0.131;
i with s*-53.669 q*12.342;
s with q*-14.052;
[qft*0.313];
qft;
[i*73.050 s*8.125 q*-2.161];
i*67.887; s*64 q*3.958;
y1*52.956 y2*64.049 y3*55.481 y4*19.390;
%cg#1%
i s q | y1@0 y2@1 y3@2 y4@3;
[qft*0.313];
qft;
[i*73.050 s*8.125 q*-2.161] (I1 S1 Qg);
%cg#2%
i s q | y1@0 y2@1 y3@2 y4@3;
[qft*2.704];
qft;
[i*71.541 s*6.525 q*-2.161] (I2 S2 Qg);
MODEL PRIORS:
I1~N(73.050, 1);
S1~N(8.125, 1);
I2~N(71.541, 1);
```

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

 © 2018 Hogrefe Publishing

```
S2~N(6.525, 1);
Qg~N(-2.161, 1);
MODEL CONSTRAINT:
NEW(diff_s)*1.6;
diff_S = S1 - S2;
OUTPUT: TECH9;
```

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.