Marginal and Conditional Posterior Predictive p-values in Bayesian SEM

A Thesis presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

---

by

ELLEN FITZSIMMONS

Dr. Edgar C. Merkle, Thesis Supervisor

MAY 2021

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled:

Marginal and Conditional Posterior Predictive p-values in Bayesian SEM

---

presented by Ellen Fitzsimmons, a candidate for the degree of Master of Arts and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Edgar Merkle

Dr. Jeffrey Johnson

Dr. Wesley Bonifay

Acknowledgements

I would like to express my deepest gratitude to Dr. Ed Merkle for helping me make the work presented in this thesis possible. In particular, I would like to thank him for always helping me find resources to figure anything out and answering every question I ever had multiple times as I grappled with new material. I would like to thank Ben Graves and Ron Flores for their support and friendship during my time here and being available to chat about research ideas and give feedback on presentations. Finally, I would like to thank my friend Bre Boss for her encouragement and friendship. She was always excited to talk about what I was working on and reminded me to make time for myself via a healthy dose of Korean dramas.

Contents

Table captions

Figure captions

Abstract

The posterior predictive p-value (ppp-value) is currently the primary measure of fit for Bayesian SEM. It is a measure of discrepancy between observed data and a posited model, comparing an observed likelihood ratio test (LRT) statistic to the posterior distribution of LRT statistics under a fitted model. However, the LRT statistic requires a likelihood, and multiple likelihoods are available for a given SEM: we can use a marginal likelihood that integrates out the latent variable(s), or we can use a conditional likelihood that conditions on the latent variable(s). A ppp-value based on conditional likelihoods is unexplored in the SEM literature, so the goal of this project is to study its performance alongside the marginal ppp-value. We present comparisons of the marginal and conditional ppp-values using real and simulated data, leading to recommendations on uses of the metrics in practice.

Keywords. Posterior predictive p-value, Bayesian SEM, Confirmatory Factor Analysis Models, Latent Growth Models, model fit metrics

Introduction

Until recently, applications of Bayesian analysis were limited outside of the field of statistics, presumably because it was difficult due to complex statistical specifications required (B. Muthen & Asparouhov, 2012). The development of Bayesian analysis in Mplus provided a less technical software with convenient defaults, effectively contributing to expanding the accessibility of Bayesian analysis in other fields of study. With this increased accessibility, Bayesian methods have rapidly grown in popularity and are often applied to many of the models found under the umbrella of structural equation modeling (SEM). Some benefits to Bayesian analysis are: it produces an analysis that better reflects substantive theories, it is useful for measurement aspects of latent variable modeling (e.g., Confirmatory Factor Analysis and measurement part of SEM), it is shown to perform well with both non-informative and informative priors, and when a model is misspecified Bayesian estimation with informative priors can outperform ML estimation (B. Muthen & Asparouhov, 2012; Cain & Zhang, 2019; Lee, Cai, & Kuhfeld, 2016; Garnier-Villarreal & Jorgensen, 2019; Gelman, Carlin, et al., 2013).

Despite this progress in the adoption of Bayesian methods into other fields of study, there remain valuable expansions in methodology to investigate. For instance, model fit and appraisal diagnostics have been understudied compared to model building and estimation. Further, within the study of model fit and appraisal diagnostics, there seems to be a focus on model comparison over model fit metrics. In fact, popular model comparison criteria (e.g., AIC, DIC, WAIC) are often used alongside or instead of model fit metrics in the literature (Bozorgzadeh & Bathurst, 2019; Levy, 2011; Cain & Zhang, 2019).

Model comparison has its merits, but we are interested here in expanding knowledge about and encouraging the use of metrics to judge the absolute fit of a model. There are multiple avenues one could travel to explore model fit metrics in the

Bayesian framework. To start, Bayesian model fit metrics can largely be categorized as prior predictive or posterior predictive model checking. Prior predictive model checking does not require the model to be fit to the current data set and is completely dependent on the choice of the prior. Posterior predictive model checking uses a posterior distribution based on the model after the data have been incorporated and therefore is more influenced by the data (Levy, 2011). Further, there are even more choices for selecting model fit metrics within these categories of prior predictive and posterior predictive model checking. However, we will be focusing on a metric that falls within posterior predictive model checking, which seems to be more popular than prior predictive model checking (Lee et al., 2016; Levy, 2011). Specifically, we will be focusing on the posterior predictive p-value.

Meng (1994) promoted use of the posterior predictive p-value (ppp-value), which is currently the primary measure of fit for Bayesian SEM. The ppp-value should not be confused with the classical (frequentist) p-value and the formal hypothesis testing and ad hoc methods associated with it. In many practical situations in the classical setting, nuisance parameters interfere with calculating p-values. Nuisance parameters refer to unknown parameters, which in our case would be the latent variables in our models. In the classical setting, there is a two-level dependence on these nuisance parameters, with the first-level dependence being a dependence of a discrepancy variable on the nuisance parameters and the second-level dependence being a dependence of the sampling distribution on the nuisance parameters. Solutions that account for these nuisance parameters produce p-values that are not the tail-area probabilities the classical approach intended (Meng, 1994; Lancaster, 2000; Woutersen, 2001). Bayesian formulation better defines and evaluates the tail-area probability underlying the classical setting (Meng, 1994). For this reason, Meng (1994) suggested that the ppp-value could serve as a Bayes/non-Bayes compromise that could promote Bayesian methods of statistical inference in areas of study where

full Bayesian analyses are not yet accepted as a standard approach.

Use of the ppp-value and Bayesian methods in general have gained popularity since Meng's (1994) article in areas outside of statistics, like the social sciences. This expansion into other areas has resulted in a recent boost in articles addressing topics such as model fit assessment for Bayesian SEM (Asparouhov & Bengt, 2019; Garnier-Villarreal & Jorgensen, 2019; Levy, 2011), Bayesian model comparison metrics (Bozorgzadeh & Bathurst, 2019; Cain & Zhang, 2019; Gelman, Hwang, & Vehtari, 2013), and more flexible and efficient methodology for model estimation and model checking (Lee et al., 2016). These articles usually offer advice on how to use a given procedure or present a newer, more efficient way to carry out widely used procedures. Our project is a little more pointed than many of these articles because we will be investigating the behavior of ppp-values calculated in different manners. Although there are multiple statistics that can be used for calculating ppp-values, the likelihood ratio statistic has been widely used. We will be comparing this standard ppp-value calculated with a marginal likelihood against a ppp-value calculated with a conditional likelihood. In the spirit of the earlier mentioned articles, we will compare the standard option against the alternative and then offer recommendations for use.

In our study, the ppp-value is a measure of discrepancy between the observed data and posited assumptions that compares the observed likelihood ratio test (LRT) statistic to the posterior distribution of LRT statistics under a fitted model. If the model fits the observed data well, then the observed LRT statistic should fall in the middle of the posterior distribution of LRT statistics. The proportion of times that the posterior predictive LRT statistic is greater than the observed LRT statistic is the ppp-value. This proportion can be any value from 0 to 1, with values close to 0.5 indicating good model fit and values approaching 0 indicating misfit. Other than the extreme values just mentioned, there are not established cutoffs for ppp-values. Instead, they can be thought of as the probability of the fitted model producing a

data set like the observed data set (Gelman & Shalizi, 2013).

Meng's (1994) procedure for calculating ppp-values can be condensed into the following steps with LRT statistics computed for each posterior draw ν^s (s = 1 ...S):

1. Compute the observed LRT statistic, LRT(Y, ν^s)
2. Simulate data Y^{rep} from the model
3. Compute the posterior predictive LRT statistic, LRT(Y^{rep}, ν^s)
4. Record whether or not LRT$(Y^{rep}, \nu^s) >$ LRT(Y, ν^s)

Marginal and Conditional ppp-values

An LRT statistic can be calculated using either a marginal likelihood or a conditional likelihood, with the difference between the two being their respective treatments of latent variables f_i in a given model. When calculating the marginal ppp-value, f_i does not count as a model parameter and is integrated out, as shown in (1). In contrast, when calculating the conditional ppp-value f_i does count as a model parameter, with $\boldsymbol{x}_i$ conditioning on f_i, as shown in (2) and (3). More specifically, all three of these equations are depicting the marginal or conditional data needed for calculating the marginal and conditional likelihoods for a Confirmatory Factor Analysis model with one latent variable and three manifest variables. These equations can also be used for a latent growth model.

$$\boldsymbol{x}_i \sim N(\boldsymbol{\nu}, \boldsymbol{\lambda}\sigma_f^2\boldsymbol{\lambda}' + \boldsymbol{\psi}) \tag{1}$$

In (1), $\boldsymbol{x}_i$ represents marginal data, ν represents the means for the manifest variables, λ represents the factor loadings, σ_f^2 represents variances for manifest variables, and ψ represents variances for the latent variables.

$$\boldsymbol{x}_i | f_i \sim N(\boldsymbol{\nu} + \boldsymbol{\lambda} f_i, \boldsymbol{\psi}) \tag{2}$$

with

$$\boldsymbol{\psi} = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix} \tag{3}$$

In equation (2), $\boldsymbol{x}_i|f_i$ represents the conditional data (note that the data is conditioned on the latent variable f_i), ν represents the means for the manifest variables, λ represents the factor loadings, and ψ represents a covariance matrix for the manifest variables. Equation (3) shows that the covariance matrix ψ consists of a main diagonal of σ^2 values representing the variances for each manifest variable and zeros for all off-diagonal values. For these equations a single factor model is being depicted, but multifactor models will have an additional variable α representing the mean for the factors.

Practically speaking, the marginal method of calculating the ppp-value is related to the population behind the data, while the conditional method is related to the specific data being used (i.e., conditional is more influenced by the specific sample than the marginal) (e.g., E. C. Merkle, Furr, & Rabe-Hesketh, 2019). In the context of the parameters described earlier, the conditional likelihood has parameters f_i that are specific to each observation in the data, meaning that the conditional likelihood is more tailored to the observations at hand. For the marginal likelihood we average the f_i parameters, so the marginal likelihood is more about the population.

So far, ppp-values calculated with the marginal likelihood have been the standard and ppp-values calculated with the conditional likelihood have been unexplored in the SEM literature. For this reason, we are interested in comparing marginal and conditional ppp-values. Because calculating ppp-values with conditional likelihoods is not standard practice, it was necessary to write code that calculates ppp-values with conditional likelihoods.

In the sections below, we will examine ppp-values in the context of confirmatory factor analysis (CFA) and latent growth models. First, we will estimate models for simulated and real data for multiple CFA models. Then we will do the same for a group of latent growth models. We described the general process for calculating ppp-values earlier, but the process specific to the model types we will be using involves extracting parameters from our estimated models to simulate data sets for calculating marginal and conditional log likelihoods and saturated log likelihoods. The marginal and conditional log likelihoods and saturated log likelihoods, in turn, are used to calculate marginal and conditional observed and posterior predictive LRT statistics. This process repeats for a number of simulations, eventually resulting in a ppp-value produced by computing the proportion of times that the posterior predictive LRT statistic is greater than the observed LRT statistic.

We will first use simulated data to understand the behavior of the ppp-values, followed by applications of the ppp-values to real data. Then we will summarize our conclusions from these comparisons of marginal and conditional ppp-values and offer recommendations for use of the metrics in practice.

Chapter One: Confirmatory Factor Analysis

Data for the CFA models were based on the Holzinger and Swineford data set available in R (Holzinger & Swineford, 1939). These models had one latent variable and 3 to 4 manifest variables. CFA models were analyzed with the bcfa function in *blavaan* (E. Merkle & Rosseel, 2018). After estimating the model, parameters were extracted and used for simulating data sets and calculating likelihoods for ppp-values, as described earlier in Meng's (1994) 4-step procedure.

Simulated Data

Methods. Our simulated data set is a multivariate normal data set with three manifest variables and 100 observations. The model constructed for these simulated data is depicted in Figure 1. This model has one latent variable f, influencing the values for the three manifest variables x_1, x_2, and x_3. The values for the variables found in Figure 1 are depicted in (4)–(8) below. These equations show how the parameters listed earlier for extraction can be used to estimate $\boldsymbol{x}_i$, f_i, and the error for each manifest variable. These estimations contribute to simulating data for ppp-value calculation.

$$\boldsymbol{x}_i = \boldsymbol{\nu} + \boldsymbol{\lambda} \times f_i + \boldsymbol{e}_i \tag{4}$$

$$f_i \sim N(0, \sigma_f^2) \tag{5}$$

$$e_{1i} \sim N(0, \sigma_1^2) \tag{6}$$

$$e_{2i} \sim N(0, \sigma_2^2) \tag{7}$$

$$e_{3i} \sim N(0, \sigma_3^2) \tag{8}$$

The values for this data set are based on estimates from real data, with the means for the manifest variables (ν) set to 0, 5, and 10, respectively. The three factor loadings (λ) were set to 1, 6, and 11, respectively. Variances for manifest variables (θ)

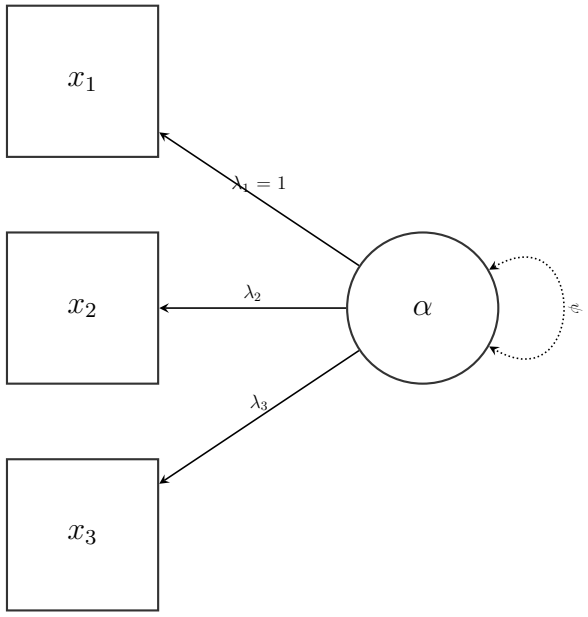

Figure 1. CFA Model 1 Path Diagram

were set to 4, 9, and 16, respectively. The variance for the latent variable (ψ) and the first factor loading for these models were all set to 1. The matrix of values for latent variables (f_i) was calculated using the provided mean (mean of α) and variance (ψ) for the latent variable. In this case, mean of α is 0 and ψ is 1.

Results. The model using the simulated data had a conditional ppp-value of 0.5 and a marginal ppp-value of 0.5. Both values were around 0.5, which indicates a good model fit. These results indicate that our code works because we generated data directly from the model, which was built with the true population values in mind (we know these values because we simulated the data). These findings gave us confidence that the code we had written to calculate marginal and conditional ppp-values worked as intended for CFA models and that we could test with real data.

Applied Data

Methods. For applied data, we used the Holzinger and Swineford data set (N = 301) that can be found in *lavaan* (Rosseel, 2012). This data set consists of mental

ability scores for seventh- and eighth-grade students from two schools in Chicago, IL, in 1939. For our project, we will be using the manifest variables labeled x1, x2, x3, x4, x8, and x9. Variables x1–x3 contain participants' scores on three spatial ability tests. Variable x4 contains participants' score on a verbal ability measure. Variables x8 and x9 contain participants' scores on two mental speed tests.

We built three models for the Holzinger and Swineford data set that had varying degrees of fit. These models each had one latent variable and three to four manifest variables. The first model and the one that is supposed to have the best fit uses variables x1–x3, i.e., a model that reflects participants' spatial abilities. The structure to this model is identical to that used with the simulated data, so Figure 1 can also represent this model.

The first model, in theory, should have good fit because the manifest variables all measure the same construct. This is a good start, but it is more interesting to see how marginal and conditional ppp-values may differ as model fit decreases. For this reason, we built the second model with variables x1–x4. This model has variables reflecting both spatial and verbal abilities, so this should lead to worse model fit. The path diagram in Figure 2 represents this second model.

Our third CFA model uses variables x1, x2, x8, and x9. With half of the manifest variables representing spatial abilities and the other half representing mental speed, this model should be the worst fitting of the three models. The path diagram in Figure 3 represents the final CFA model.

Results.

CFA Model 1 Results. The first model we built for the Holzinger and Swineford data set used manifest variables x1–x3 and is depicted in Figure 1. These three manifest variables contained participants' scores in three spatial ability tests. Summary statistics for model 1 can be found in Table 1 in the Appendix. All estimates had R-hat values of 1, indicating convergence (this will be the case for all

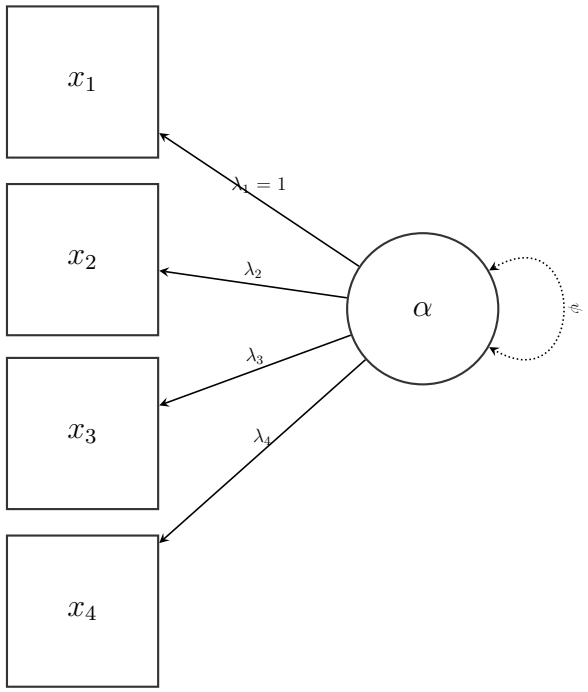

Figure 2. CFA Model 2 Path Diagram

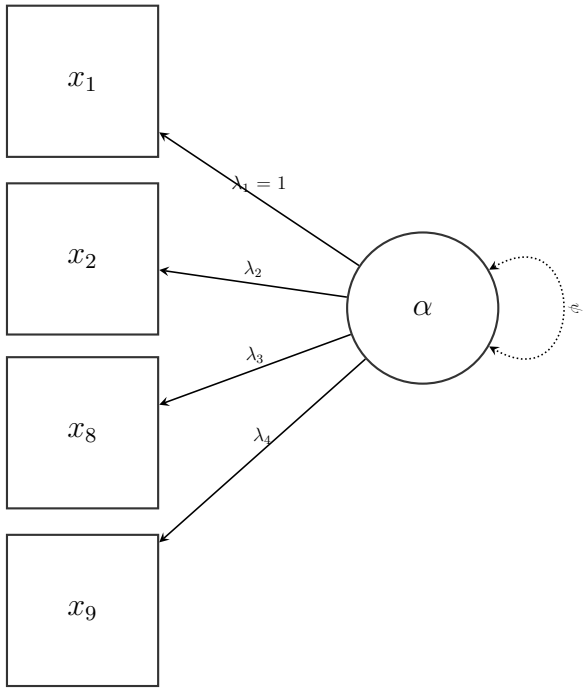

Figure 3. CFA Model 3 Path Diagram

models described). The marginal ppp-value was 0.47, and the conditional ppp-value was 0.54. Both of these values stay close to 0.5, indicating that there is a good model fit.

CFA Model 2 Results. Our second model built to test the Holzinger and Swineford data set used variables x1–x4, and its structure can be found in the path diagram in Figure 2. This model used manifest variables containing participants' scores on three spatial ability tests (variables x1–x3) and one verbal ability test (variable x4). We included variable x4 in model 2 to decrease model fit and record any differences that may occur between the marginal and conditional ppp-values.

The summary statistics for model 2 can be found in Table 2 in the Appendix. The R-hat values for all the estimates are around 1, indicating convergence. For model 1, the marginal and conditional ppp-values were close in value, but the addition of variable x4 in model 2 increased the difference between the two ppp-values. For this model, the marginal ppp-value was 0.09, and the conditional ppp-value was 0.45. In this case, the marginal ppp-value indicates poor model fit, and the conditional ppp-value remains close to 0.5 and indicates good model fit. These results were intriguing because they indicated the potential for disagreement between the marginal and conditional ppp-values on model fit as models became more ill-fitting.

CFA Model 3 Results. With the two previous models discussed, the conditional ppp-value stayed close to 0.5. We were unsure if these results indicated that conditional ppp-values always indicate good model fit or if they are inclined to indicate better model fit than marginal ppp-values. For this reason, the third CFA model built for the Holzinger and Swineford data set incorporated a mix of test score types in hopes that it would have poor enough fit to budge the conditional ppp-value away from 0.5. This model used variables x1, x2, x8, and x9, with half of these variables containing participants' scores on spatial ability tests (variables x1 and x2) and the other half containing participants' scores on mental speed tests (variables x8

and x9). The structure of this model can be found in the path diagram in Figure 3.

The summary statistics for this model can be found in Table 3 in the Appendix. As was the case with the previous models, R-hat values were around 1, indicating convergence. The marginal ppp-value was 0.05 and the conditional ppp-value was 0.18. As was the case for models 1 and 2, the conditional ppp-value indicated a better model fit than the marginal ppp-value, but for this model the conditional ppp-value had moved away from 0.5.

CFA Models Discussion

In this chapter, we presented four CFA models. One model was built for simulated data and ensured the code we had written to calculate marginal and conditional ppp-values produced expected results. After confirming that our code produced expected results for CFA models, we examined three models with decreasing model fit for the Holzinger and Swineford data set. The graph in Figure 4 plots the marginal and conditional ppp-values for the applied data set models. This graph shows a trend of the conditional ppp-values having larger values than their marginal counterparts. In other words, we found that the conditional ppp-values indicated better model fit than the marginal ppp-values for the CFA models we tested.

Although the results we found are promising for indicating a pattern in the relationship between marginal and conditional ppp-values within CFA models, there are some difficulties that must be addressed. For one, we would have to compare many more models to be able to generalize the results we found. Ideally, we would have many more models for more data sets (simulated and real) estimated for comparison. With the handful of models and data sets we used, the pattern of conditional ppp-values indicating better fit than the marginal could be a fluke dependent on the specific models or data we used. The only way to address this and be more certain the pattern exists is by comparing more models and data sets. Another concern that

comes up is that our findings may be limited to this specific type of model or data (i.e., CFA models or the type of data used in CFA models). To address this concern, the next chapter considers the same issues in the context of latent growth models.

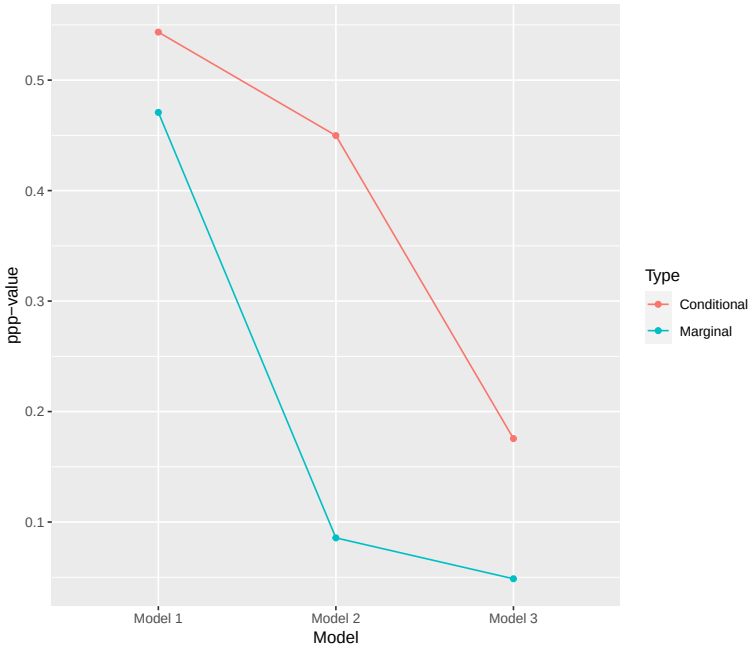

Figure 4. CFA Model ppp-values Comparison

Chapter Two: Latent Growth Models

The procedure for the Latent Growth model testing mirrors that of the CFA models, except for some extra parameters that were described earlier that will be accounted for. This difference in the parameters extracted is a reflection of the fact that the models proposed here will have two latent variables instead of one. The recommended minimum number of occurrences for a Latent Growth model varies depending on the complexity of the model, with 3 or 4 time points being the recommended number of occurrences for a linear model (Whittaker & Khojasteh, 2017; Curran, Obeidat, & Losardo, 2010; L. Muthen, 1999). For this reason, these proposed models will also have more manifest variables (4–5 instead of 3–4) than the CFA models to ensure that enough parameters are identified in the models. Latent Growth models will be analyzed with the bgrowth function in *blavaan* (E. Merkle & Rosseel, 2018). After running the model, parameters will be extracted and used for simulating data sets that will be used to calculate likelihoods for ppp-value calculation.

Simulated Data

Methods. When we simulated data for the CFA models, we created a data set from scratch. Creating a data set from scratch that can be used for Latent Growth models is more complicated, so we used the sleep study data set (N = 18) found in the *lme4* package in R (Bates, Mächler, Bolker, & Walker, 2015) to simulate data. This data set contains the average reaction time of 18 truck drivers over the course of 9 days. During this time, participants had their normal amount of sleep on day 0 and were limited to 3 hours of sleep per night on days 1–9. Each day, the participants were given a series of tests to measure average reaction time for that day. This data set has 180 observations for three variables: average reaction time in milliseconds ("Reaction"), number of days of sleep deprivation ("Days"), and participant

("Subject").

Simulating data from the sleep study data set involved building a model for the sleep study data set so that we could extract parameters for simulating the data. Our simulated data set had 2000 observations and four manifest variables. As can be seen in Figure 5, the model we built to test the simulated data set has two latent variables and four manifest variables.

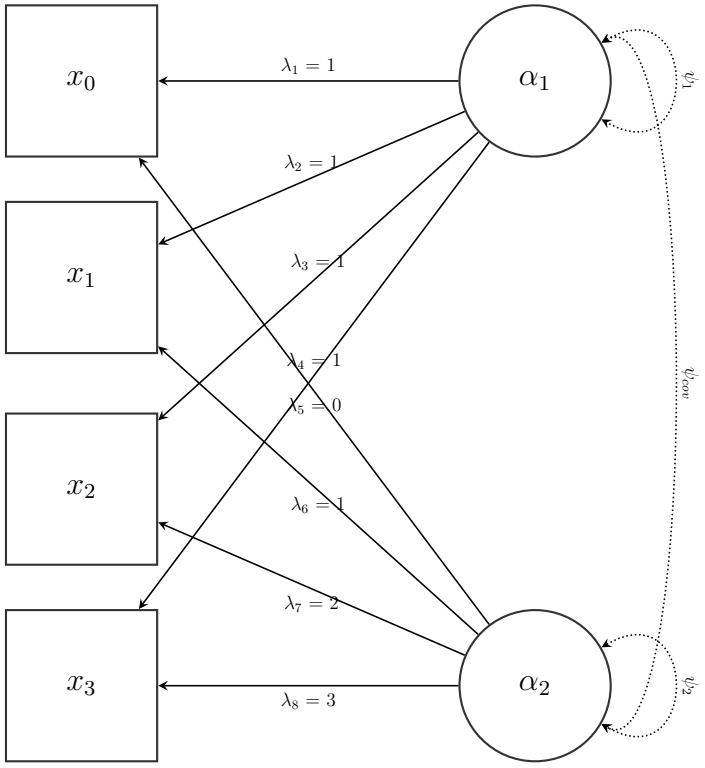

Figure 5. Latent Growth Model 1 Path Diagram

Results. The model we built for this data did not wholly produce the ppp-values we expected and had gotten for the simulated data for the CFA models. The marginal ppp-value for this model was higher than we expected, with a value of 0.85, and the conditional ppp-value was what we had expected with a value of 0.52. Although the marginal ppp-value was higher than the expected value of 0.5, it did equal the marginal ppp-value provided by the blavaan package, which suggests that

the marginal ppp-value is accurate. This may indicate that there is something about the data we created that is causing this strange marginal ppp-value rather than the code we created not working. Whenever we repeated this procedure with a new generated data set, we found results that were more in line with what we had seen before. With the new generated data set the marginal ppp-value for the model was 0.43 and the conditional ppp-value was 0.51. After this, we carried on with analyzing the models for the sleep study data set.

Applied Data

Methods. The proposed models for this portion of the project each have two factors (the intercept and slope) and 4–5 manifest variables, with each variable representing one day in the sleep study. The first proposed model will use the four manifest variables containing participants' average reaction times on days 0 through 3 of the study. This model has the same structure as the one used for the simulated data, so the path diagram in Figure 5 also depicts this model. The second model will use participants' reaction times on days 0, 2, 4, 7, and 9 of the study. The structure of this model is depicted in Figure 6. Finally, the third model will use participants' reaction times on days 0, 4, 5, and 9 of the study. Figure 7 depicts the structure of this last model.

Results.

Latent Growth Model 1 Results. Model 1 is represented in the path diagram found in Figure 5, and the estimates for this model can be found in Table 4 in the Appendix (the loadings for the model are included at the top of the table). This model had R-hat values of 1, so it did converge. For this model, the marginal ppp-value was 0.31, indicating a mediocre model fit. In contrast, the conditional ppp-value was 0.44, indicating good model fit.

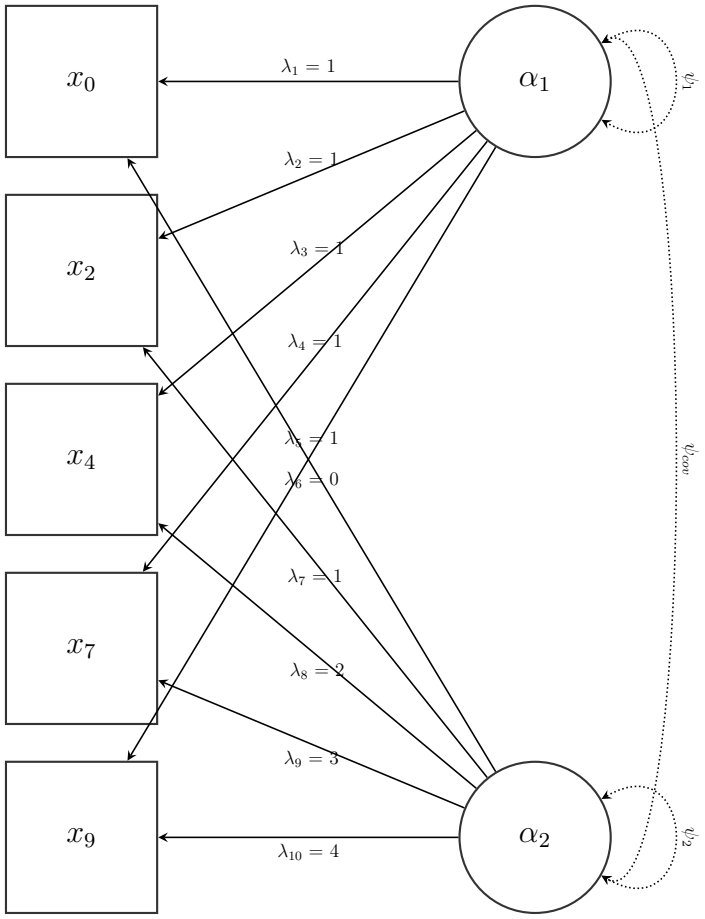

Figure 6. Latent Growth Model 2 Path Diagram

Latent Growth Model 2 Results. Model 2 is depicted in Figure 6, and its loadings and estimates are in Table 5 in the Appendix. The marginal ppp-value was 0.34, and the conditional ppp-value for this model was 0.41. As was the case with the first model, the marginal ppp-value indicates a mediocre model fit and the conditional ppp-value indicates a good model fit.

Latent Growth Model 3 Results. Model 3 is depicted in Figure 7, and this model's loadings and estimates are in Table 6 in the Appendix. As was the case for all the other models, this model had R-hat values at 1, indicating that the model converge. For this model, the marginal ppp-value is 0.29, and the conditional

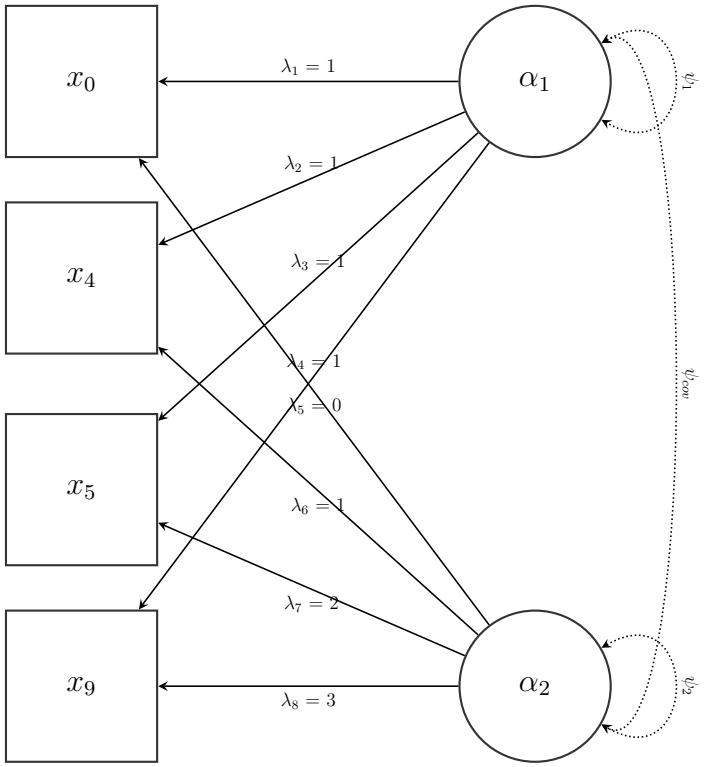

Figure 7. Latent Growth Model 3 Path Diagram

ppp-value is 0.39. Similar to the earlier models, the marginal ppp-value indicates mediocre model fit, and the conditional ppp-value indicates fair model fit.

Latent Growth Models Discussion

Our model for the simulated data produced a conditional ppp-value indicating excellent model fit, which is expected when using simulated data, but the marginal ppp-value had a high value indicating that it was ordered differently than the other ppp-values. It is possible the sleep study data set was a peculiar data set, resulting in the unusual marginal ppp-value produced. The fact that the marginal ppp-value matched the marginal ppp-value provided by preexisting software (blavaan) gives us confidence in the accuracy of our results.

As can be seen in Figure 8, the applied models had mediocre to good model fit,

with the conditional ppp-values indicating a better model fit than the marginal. These results were on par with what we had expected and mirrored the trend we noticed with the CFA models.

There were more difficulties involved with this portion of the project compared to the portion for CFA models. It was more difficult to write the code for calculating ppp-values compared to the CFA models because we were adapting the existing CFA code to this more complex model, which had the extra alpha parameter and multiple latent variables. Additionally, it is trickier to simulate data for this type of model, which further contributed to making it difficult to check that our code calculated reasonable ppp-values for our models. Finally, as was the case with the CFA models, many more models and data sets being used for these comparisons would be useful in determining if we can generalize the trends we found with the relationship between marginal and conditional ppp-values.

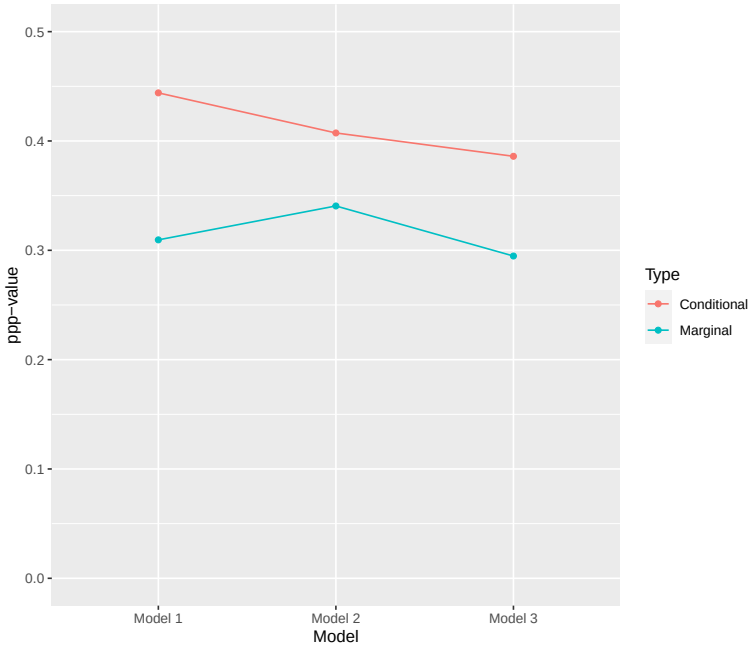

Figure 8. Latent Growth Model ppp-values Comparison

Conclusions

For this project, we explored the relationship between marginal and conditional ppp-values by comparing them in the context of Bayesian SEM. We built multiple CFA and Latent Growth models and studied them using simulated and applied data sets. For all of these models, we recorded a trend of conditional ppp-values being closer to 0.5 and therefore indicating better model fit than their marginal counterparts. These findings make sense for two reasons. The first reason is that posterior checking double-uses actual observations by testing the actual observations with a predictive distribution learned from the actual observations. This double-use of the observations results in an optimistic bias in which ppp-values are concentrated around 0.5 rather than being uniformly distributed between 0 and 1 (Qiu, Feng, & Li, 2017; Marshall & Spiegelhalter, 2007). The second reason these findings make sense is because conditional ppp-values are related to a model's fit to the specific people who were observed in the data, whereas the marginal ppp-values are related to the model's fit to the population of interest (for more discussion of these differences, see E. C. Merkle et al., 2019). The combination of ppp-values having an optimistic bias towards 0.5 anyway because of double-use of data and that conditional ppp-values are more related to their specific data than the marginal could be contributing to the conditional ppp-values being larger than their marginal counterparts.

Although the conditional ppp-values exhibited a pattern of indicating better model fit than the marginal ppp-values, the size of the difference between the conditional and marginal ppp-values for a given model varied. For instance, the conditional ppp-value for the second applied CFA model is 0.41 and the marginal ppp-value is 0.34, which has a difference of 0.07. These values have a large enough difference that, in this case, the conditional and marginal ppp-values indicate different degrees of model fit for the same model. This instance presents a potential problem when using conditional and marginal ppp-values: there is the possibility that for some models the

conditional and marginal ppp-value may disagree on model fit.

With the current trend that we have recorded, users should be aware that there is the potential for abuse of the conditional ppp-value by researchers who want their models to fit their data and who would prefer statistics that indicate good fit. However, the differences in marginal and conditional ppp-values may also be informative about the degree and sources of misfit in a dataset.

Limitations

As mentioned before, one of the study's limitations includes the limited number of data sets and models tested, making it difficult to determine what results are caused by quirks in the data or specific models tested and what results can be generalized to represent the general behavior of conditional and marginal ppp-values. Based on our findings, it seems likely that the conditional ppp-value will generally indicate better fit — or at least be closer to 0.5 — than its marginal counterpart for CFA and Latent Growth models. However, these results cannot be generalized to other model types until more model types and data sets have been examined. The last limitation that will be mentioned here is that there may be more efficient ways of determining model fit. Looping through the same model hundreds or even thousands of times can add up, especially for more complex models. For our code, the loops are not dependent on each other, so our code's efficiency may improve if we can successfully parallelize our for-loops. This may be possible in R by rewriting the for-loop as an lapply call or by using foreach loops in conjunction with either the package doMC or doSNOW, depending on which OS is being used (Microsoft & Weston, 2020; Wallig, Analytics, & Weston, 2020; Corporation & Weston, 2020). Some other options in R for parallel processing are the parallel and future packages (R Core Team, 2020; Bengtsson, 2020). The parallel package is part of the core distribution of R and uses multiple cores on the user's machine. This package is supposed to be fairly simple to use

(compared to other options for parallel processing). The future package implements the Future API for parallel processing and R code can be evaluated on a local machine, in parallel a set of local machines, or distributed on a mix of local and remote machines. Another thing to note is that the parallel and future packages are not restricted to an OS like some of the other packages mentioned are.

Future Directions

The work that has been described so far has set a foundation for how we think the degree of model fit provided by marginal and conditional ppp-values compare to each other, but more can be done to increase confidence in the conclusions we have drawn. First, the ppp-values for many more models with varying data sets could be calculated and compared. For instance, a simulation study could be set up that catalogs the marginal and conditional ppp-values for hundreds of models. If this gave similar results to what we found, we could be more confident in our conclusions. Second, our conclusions would be even more compelling if we could back up our findings with a proof-type solution demonstrating that conditional ppp-values are inclined to indicate better model fit than marginal ppp-values.

In addition to corroborating what we already found, future research paths also include examining the behavior of conditional and marginal ppp-values for an even broader range of model types and data. Part of this exploration could involve using other metrics in conjunction with marginal and conditional ppp-values to establish a better idea of what true model fit is when marginal and conditional ppp-values disagree. Garnier-Villarreal and Jorgensen (2019) propose multiple BSEM counterparts to frequentist chi-square-based SEM fit indices that could be used in conjunction with the marginal and conditional ppp-values. In particular, $\text{BRMSEA}^{\text{PPMC}}$ and $\text{BRMSEA}^{\text{DevM}}$ would be interesting to look into in conjunction with marginal and conditional ppp-values. When there are large samples in which

ppp-values would reject models with even minor misspecification, BRMSEA^{PPMC} complements ppp-values because it will indicate approximately well-fitting models are acceptable (Garnier-Villarreal & Jorgensen, 2019). BRMSEA^{DevM} is similar to BRMSEA^{PPMC}, but it more closely estimates the same quantity that RMSEA estimates in a frequentist framework. Because of the complementary nature of the relationship between ppp-values and BRMSEA metrics, BRMSEA may offer insight into the cases when marginal and conditional ppp-values disagree.

Not only could we incorporate other metrics to expand on the type of work we have conducted in this paper, but there are also many other uses for ppp-values that could be explored. Some examples that have been found in the literature but could be explored further is using ppp-values to identify outliers or influential cases and to assess and compare different priors and models (Bozorgzadeh & Bathurst, 2019; Hjort, Dahl, & Steinbakk, 2006). Rather than limiting the use of the ppp-value as an absolute model fit metric, it can be used to test essentially any component of the model. Using the ppp-value in this fashion could improve model building and model selection. On trend with the rest of this project it would be interesting to examine this use of the ppp-value in the context of marginal and conditional ppp-values.

As stated in the beginning, there is still much to be explored in using Bayesian SEM in Psychology. Our findings indicate that keeping marginal ppp-values as the standard may be best considering most researchers are interested in generalizable models. However, it seems that conditional ppp-values could be valuable in assessing model fit for researchers interested more in the sample than the overall population.

References

Asparouhov, T., & Bengt, M. (2019). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bengtsson, H. (2020, aug). *A unifying framework for parallel and distributed processing in r using futures.* Retrieved from https://arxiv.org/abs/2008.00553

Bozorgzadeh, N., & Bathurst, R. J. (2019). Bayesian model checking, comparison and selection with emphasis on outlier detection for geotechnical reliability-based design. *Computers and Geotechnics*, *116*. doi: 103181

Cain, M. K., & Zhang, Z. (2019). Fit for a Bayesian: An evaluation of PPP and DIC for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 39-50.

Corporation, M., & Weston, S. (2020). dosnow: Foreach parallel adaptor for the 'snow' package [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=doSNOW (R package version 1.0.19)

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*(2), 121-136. Retrieved from https://doi.org/10.1080/15248371003699969 (PMID: 21743795) doi: 10.1080/15248371003699969

Garnier-Villarreal, M., & Jorgensen, T. D. (2019). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological methods*.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013).

Bayesian data analysis: Texts in statistical science (3rd ed.). London: CRC Press. (ISBN 978-1439840955)

Gelman, A., Hwang, J., & Vehtari, A. (2013, 07). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*. doi: 10.1007/s11222-013-9416-2

Gelman, A., & Shalizi, C. R. (2013). Philosophy and practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 8-38. doi: 10.1111/j.2044-8317.2011.02037.x

Hjort, N. L., Dahl, F. A., & Steinbakk, G. H. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, *101*(475), 1157-1174. doi: 10.1198/016214505000001393

Holzinger, K., & Swineford, F. (1939). A study in factor analysis: The stability of a bifactor solution. *Supplementary Educational Monograph*(48).

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, *95*(2), 391 - 413. doi: https://doi.org/10.1016/S0304-4076(99)00044-5

Lee, T., Cai, L., & Kuhfeld, M. (2016). A poor person's posterior predictive checking of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 206-220.

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(4), 663-685.

Marshall, E., & Spiegelhalter, D. (2007). Identifying outliers in Bayesian hierarhical models: A simulation-based approach. *Bayesian Analysis*, *2*(2), 409-444.

Meng, X. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142-1160.

Merkle, E., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4). doi:

10.18637/jss.v085.i04

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, *84*, 802–829.

Microsoft, & Weston, S. (2020). foreach: Provides foreach looping construct [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=foreach (R package version 1.5.1)

Muthen, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313 - 335. doi: 10.1037/a0026802

Muthen, L. (1999). *Mplus discussion: Growth modeling of longitudinal data.* http://www.statmodel.com/discussion/messages/14/20.html. (Accessed: 2020-12-03)

Qiu, S., Feng, C., & Li, L. (2017). Approximating cross-validatory predictive p-values with integrated is for disease mapping models. *Statistics in Medicine*.

R Core Team. (2020). 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 (2)*, 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Wallig, M., Analytics, R., & Weston, S. (2020). domc: Foreach parallel adaptor for 'parallel' [Computer software manual]. Retrieved from https://cran.r-project.org/package=doMC (R package version 1.3.7)

Whittaker, T. A., & Khojasteh, J. (2017). Detecting appropriate trajectories of growth in latent growth models: The performance of information-based criteria. *The Journal of Experimental Education*, *85*(2), 215-230. Retrieved from

https://doi.org/10.1080/00220973.2015.1123669 doi: 10.1080/00220973.2015.1123669

Woutersen, T. (2001). *Robustness against incidental parameters and mixing distributions* (Research Report No. 2001-10). Retrieved from http://hdl.handle.net/10419/70386

Appendix

The tables here contain the summary statistics for the applied models described earlier. In these tables, "Latent Variables" corresponds to summary statistics for factor loadings, "Intercepts" corresponds to summary statistics for intercept values for manifest or latent variables, "Variances" corresponds to summary statistics for variances for manifest or latent variables, and for the Latent Growth models "Covariances" corresponds to summary statistics for the covariance of the latent variables.

Table 1

CFA Model 1 Results

	Estimate	Post SD	Lower	Upper
Latent Variables				
x1	1.000			
x2	0.813	0.159	0.532	1.153
x3	1.261	0.324	0.799	2.083
Intercepts				
x1	4.936	0.067	4.804	5.067
x2	6.087	0.068	5.955	6.22
x3	2.250	0.066	2.123	2.381
f	0.000			
Variances				
x1	0.886	0.128	0.64	1.142
x2	1.095	0.113	0.887	1.328
x3	0.581	0.175	0.114	0.864
f	0.480	0.136	0.24	0.767

Table 2

CFA Model 2 Results

	Estimate	Post SD	Lower	Upper
Latent Variables				
x1	1.000			
x2	0.582	0.138	0.334	0.871
x3	0.750	0.154	0.473	1.086
x4	0.552	0.103	0.362	0.771
Intercepts				
x1	4.934	0.068	4.799	5.066
x2	6.087	0.067	5.953	6.22
x3	2.249	0.065	2.121	2.379
x4	3.060	0.068	2.926	3.192
f	0.000			
Variances				
x1	0.564	0.161	0.191	0.85
x2	1.140	0.112	0.931	1.368
x3	0.860	0.109	0.651	1.08
x4	1.134	0.104	0.943	1.353
f	0.802	0.187	0.486	1.223

Table 3

CFA Model 3 Results

	Estimate	Post SD	Lower	Upper
Latent Variables				
x1	1.000			
x2	0.616	0.170	0.308	0.981
x8	1.034	0.205	0.711	1.501
x9	1.704	0.417	1.082	2.675
Intercepts				
x1	4.938	0.066	4.811	5.066
x2	6.089	0.068	5.956	6.222
x8	5.528	0.059	5.413	5.644
x9	5.376	0.058	5.262	5.487
f	0.000			
Variances				
x1	1.079	0.108	0.88	1.299
x2	1.298	0.110	1.094	1.525
x8	0.759	0.077	0.615	0.914
x9	0.288	0.139	0.007	0.537
f	0.280	0.093	0.125	0.483

Table 4

Latent Growth Model 1 Results

	Estimate	Post SD	Lower	Upper
Latent Variables				
Int Reaction 0	1.000			
Int Reaction 1	1.000			
Int Reaction 2	1.000			
Int Reaction 3	1.000			
Slope Reaction 0	0.000			
Slope Reaction 1	1.000			
Slope Reaction 2	2.000			
Slope Reaction 3	3.000			
Intercepts				
Int	255.248	9.027	237.213	273.313
Slope	7.972	3.554	0.977	15.083
Variances				
Reaction Times	229.415	61.026	139.954	372.189
Int	1292.962	689.683	483.872	3019.274
Slope	183.472	109.351	53.937	457.7
Covariances				
Int:Slope	-157.892	203.129	-623.812	139.498

Table 5

Latent Growth Model 2 Results

	Estimate	Post SD	Lower	Upper
Latent Variables				
Int Reaction 0	1.000			
Int Reaction 2	1.000			
Int Reaction 4	1.000			
Int Reaction 7	1.000			
Int Reaction 9	1.000			
Slope Reaction 0	0.000			
Slope Reaction 2	1.000			
Slope Reaction 4	2.000			
Slope Reaction 7	3.000			
Slope Reaction 9	4.000			
Intercepts				
Int	247.691	7.979	231.795	263.61
Slope	24.218	4.096	16.224	32.468
Variances				
Reaction Times	795.597	152.327	548.26	1141.186
Int	669.973	503.655	92.463	1951.709
Slope	220.876	136.819	61.15	574.368
Covariances				
Int:Slope	71.713	170.137	-293.996	401.243

Table 6

Latent Growth Model 3 Results

	Estimate	Post SD	Lower	Upper
Latent Variables				
Int Reaction 0	1.000			
Int Reaction 4	1.000			
Int Reaction 5	1.000			
Int Reaction 9	1.000			
Slope Reaction 0	0.000			
Slope Reaction 4	1.000			
Slope Reaction 5	2.000			
Slope Reaction 9	3.000			
Intercepts				
Int	255.832	8.596	238.651	272.88
Slope	30.256	5.595	19.043	41.272
Variances				
Reaction Times	632.052	145.069	403.205	976.099
Int	886.065	605.396	180.749	2431.901
Slope	428.979	246.279	132.99	1054.364
Covariances				
Int:Slope	224.239	254.420	-254.041	777.492