3

A TUTORIAL ON USING THE WAMBS CHECKLIST TO AVOID THE MISUSE OF BAYESIAN STATISTICS

Rens van de Schoot

DEPARTMENT OF METHODOLOGY AND STATISTICS, UTRECHT UNIVERSITY, UTRECHT, THE NETHERLANDS & OPTENTIA RESEARCH PROGRAM, FACULTY OF HUMANITIES, NORTH-WEST UNIVERSITY, VANDERBIJLPARK, SOUTH AFRICA

Duco Veen

DEPARTMENT OF METHODOLOGY AND STATISTICS, UTRECHT UNIVERSITY, UTRECHT, THE NETHERLANDS

Laurent Smeets

DEPARTMENT OF METHODOLOGY AND STATISTICS, UTRECHT UNIVERSITY, UTRECHT, THE NETHERLANDS

Sonja D. Winter

PSYCHOLOGICAL SCIENCES, UNIVERSITY OF CALIFORNIA, MERCED, CA, UNITED STATES OF AMERICA

Sarah Depaoli

PSYCHOLOGICAL SCIENCES, UNIVERSITY OF CALIFORNIA, MERCED, CA, UNITED STATES OF AMERICA

Introduction

The current chapter guides the reader through the steps of the When-to-Worry-and-How-to-Avoid-the-Misuse-of-Bayesian-Statistics checklist (the WAMBS checklist), in order to provide background for the other chapters in this book. New in comparison to the original WAMBS checklist is that we include prior and posterior predictive model checking. We also compare the performance of two popular Bayesian software packages: `RStan` (Carpenter et al., 2017) and `rjags` (Plummer, Stukalov, & Denwood, 2018) ran via `blavaan` (Merkle & Rosseel, 2018). We show why using the Hamiltonian Monte Carlo (HMC) procedure (Betancourt, 2017), available in `RStan`, is more efficient when sample size is small. Note that for a full explanation of each step we refer to the paper in which the checklist was published (Depaoli & Van de Schoot, 2017). For a more detailed introduction to

Bayesian modeling, we refer the novice reader to Chapter 1 (Miočević, Levy, & Van de Schoot), among many other resources. The checklist is extended in Chapter 4 (Veen & Egberts) with some additional tools and debugging options. All data and the annotated R code to reproduce the results are available on the Open Science Framework (https://osf.io/am7pr/).

Example data

The data we use throughout the chapter is based on a study of PhD delays (Van de Schoot, Yerkes, Mouw, & Sonneveld, 2013). Among many other questions, the researchers asked the PhD recipients how long it had taken them to finish their PhD thesis ($n = 333$). It appeared that PhD recipients took an average of 59.8 months (five years and four months) to complete their PhD trajectory. The variable of interest measures the difference between planned and actual project time in months ($\overline{delay} = 9.97$, $min/\max = -31/91$, $\sigma = 14.43$).

Let us assume we are interested in the question of whether age ($\overline{age} = 31.68$, $min/max = 26/69$) of the PhD recipients is related to delay in their project. Also, assume we expect this relation to be non-linear. So, in our model the gap between planned and actual project time is the dependent variable and age and age^2 are the predictors, resulting in a regression model with four parameters:

- the intercept denoted by $\beta_{intercept}$
- two regression parameters:
    - β_{age} or β_1 for the linear relation with age
    - β_{age^2} or β_2 for the quadratic relation with age
- variance of the residuals denoted by σ^2_ε

WAMBS checklist

Do you understand the priors?

Since we know that at least some degree of information is necessary to properly estimate small data (Smid, McNeish, Miočević, & Van de Schoot, 2019), the next question is: How to assess and use such information? There are many ways to specify subjective priors—for example, based on expert elicitation or previous data (Van de Schoot et al., 2018)—and none are inherently right or wrong. For more details on where to get the priors from, see Zondervan-Zwijnenburg, Peeters, Depaoli, & Van de Schoot (2017).

In the current chapter we propose to use background information to specify priors that cover a plausible parameter space. That is, we define a range of possible parameter values considered to be reasonable, thereby excluding impossible values and assigning only a limited density mass to implausible values. Note that

in the sensitivity analyses presented in Steps 7–9 of the checklist, we investigate the extent of "wiggle room" for these values. Stated differently, whether different specifications of the plausible parameter space lead to different conclusions.

Define parameter space

We developed a free online app that can help with specifying the plausible parameter space for the PhD delay example (see the OSF for source code (https://osf.io/am7pr/), and for the online version go to www.rensvandeschoot.com/pps or https://utrecht-university.shinyapps.io/priors_phd/)[1]; for a screenshot, see Figure 3.1. If useful background knowledge is available and you are therefore unsure about your prior beliefs, trying to infer a plausible parameter space is to be preferred over just relying on software defaults.

First, define what you believe to be a reasonable range for age (in years). Think about what you believe to be the youngest age someone can acquire a PhD (delay included) and what the oldest age might be. This yields an age range of, for example, 18–70. Then, define the delay (in months) you believe to be reasonable. A negative delay indicates that someone finished their PhD ahead of schedule. Think about how many months someone can finish ahead of schedule and what you believe to be the maximum time that someone can be delayed; for example, -25–120 (Figure 3.1).

Second, think about what you consider plausible estimates for the intercept, the linear effect, and the quadratic effect. The data is not centered, which means that the intercept represents the expected delay of a zero-year-old. The linear effect is the expected increase in delay (in months) over time. For example, a linear effect of 3 means that for a one-year increase in age, the expected delay increases by three months. The quadratic effect is the deviation from linearity. Let us assume we expect a positive linear increase of 2.5 starting at a delay of -35 months (note this is possible because it is the delay of a zero-year old PhD candidate) and a small negative quadratic effect of -.03, so that this effect would look like a negative parabola (n-shaped) with the maximum delay occurring around the fifties (Figure 3.1).

Priors for regression coefficients (or any prior for that matter) are never just a point estimate, but always follow a distribution. In this example, only normal distributions are used, but most Bayesian software will allow many different types of distributions. The variances of a normally distributed the prior, denoted by σ_0^2, resemble a measure of uncertainty; see also Chapter 1. It is important to note that these variances are measured on the same scale as the regression coefficients. A variance that is small for the intercept might be relatively large for the quadratic effect. This means that you always have to be careful with the default prior of any Bayesian software package. For our model, a small change in the variance of the quadratic effect has a large influence on the plausible parameter space. This becomes clear in the app because any small adjustment of the variance (note that the scales of variance sliders are different) for the quadratic effect leads to a large widening of the ribbon of the quadratic effect over time.

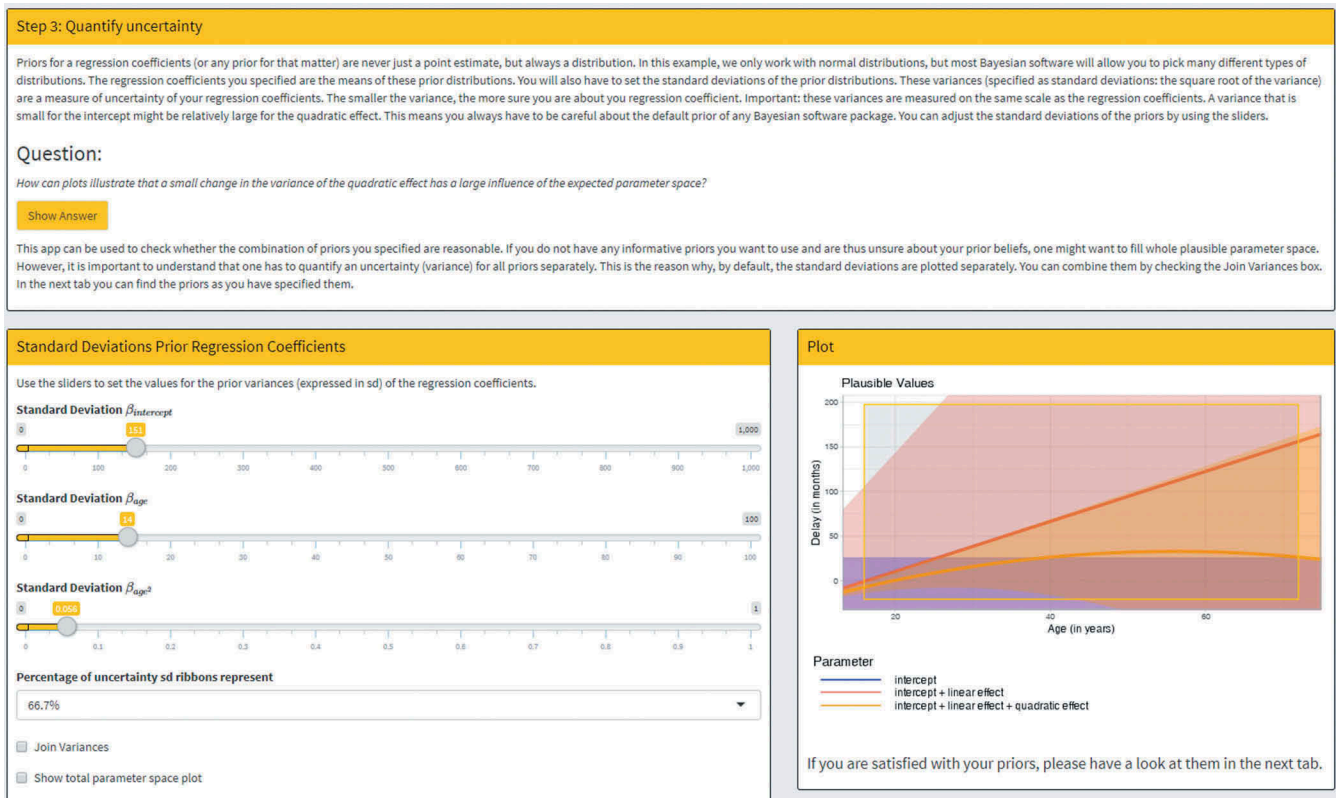

FIGURE 3.1 Screenshots of the app to specify the plausible parameter space for the PhD delay example

The following hyperparameters cover the entire plausible parameter space, with $\mathcal{N}(\mu_0,\ \sigma_0^2)$, and $IG(shape,\ scale)^2$:

- $\beta_{intercept} \sim \mathcal{N}(-35, 20)$
- $\beta_{age} \sim \mathcal{N}(.8, 5)$
- $\beta_{age^2} \sim \mathcal{N}(0, 10)$
- $\sigma_{\varepsilon}^2 \sim IG(.5, .5)$

Load data and define the model

Now that the hyperparameter values of the priors are specified, the data can be uploaded into `R` (or any other software for Bayesian estimation) and the statistical model can be specified. We ran the model using `RStan`. For an introduction, see for instance Carpenter et al. (2017). We also include some results obtained with `rjags` via `blavaan` (Merkle et al., 2019; Plummer et al., 2018) to show why using `RStan` might be preferred over `rjags` even though the syntax is more complicated. See the online supplementary materials for all annotated code (https://osf.io/am7pr/).

Prior predictive checking

Now that the model has been specified, we can investigate the priors further by computing prior predictive checks which allow for inspecting the implications of all univariate priors together. To get the prior predictive results we ignore the sample data. In the top panel of Figure 3.2, the 95% prior predictive intervals are shown for generated observations based on the priors for each individual, denoted by y_{rep}, and the observations from the sample, denoted by y. That is, values of y_{rep} are based on the prior specifications for each individual and represent possible values for PhD delay implied by the priors. In general, for all cases the prior intervals imply delays possible from approximately -100 to +100 months (with some extreme values up to $\pm$ 250) and the entire plausible parameter space (and more) is covered.

We can also look at the possible data sets generated by the priors. In the top panel of Figure 3.3, distributions of PhD delay are plotted based on the set of priors. It appears that a wide variety of data sets is plausible, though still ruling out delays larger or smaller than +300/-300. In general, we can at least be confident that when using our priors we do not exclude potential scenarios, but at the same time are able to rule out large parts of the parameters space, which is what is needed when sample sizes are small.

Does the trace-plot exhibit convergence?

To obtain estimates for the parameters in our model we make use of Monte Carlo simulations; see also Chapter 1. Traditionally a very successful and often-used algorithm is the Gibbs Sampler, a method of Markov chain Monte Carlo simulation

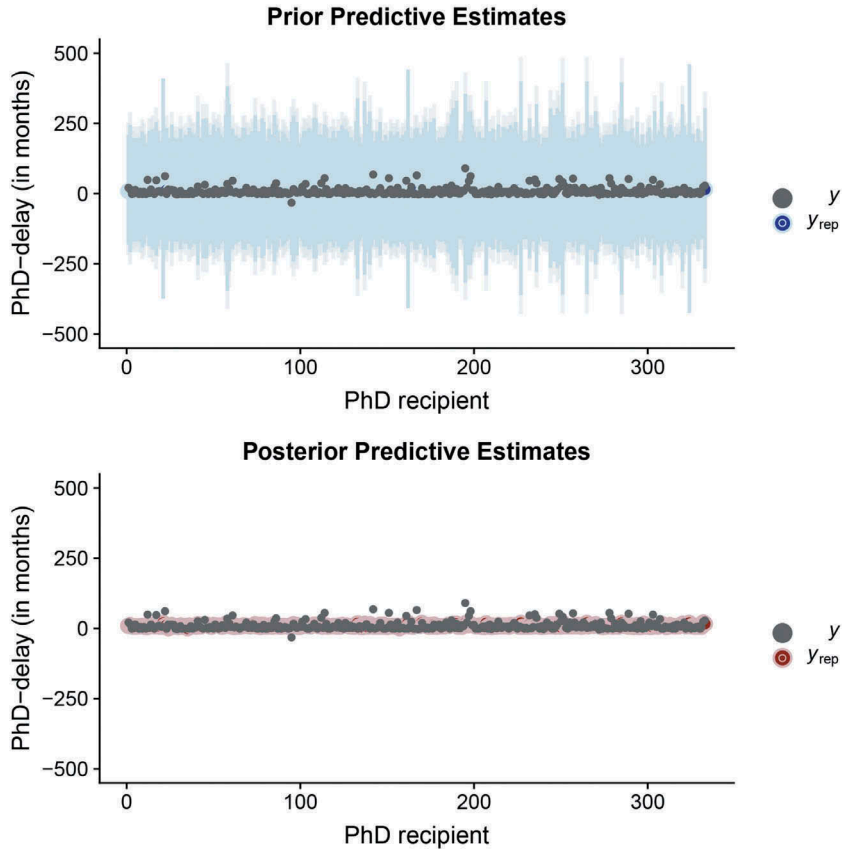

FIGURE 3.2 The 95% prior predictive intervals (top panel) and posterior predictive intervals (bottom panel) for each observation in the sample ($n = 333$)

(MCMC). This is the algorithm used by `rjags`. `RStan` uses a different MCMC algorithm, namely HMC and specifically the No-U-Turn-Sampler (Hoffman & Gelman, 2014); see also Chapter 4. For a conceptual introduction to HMC see Betancourt (2017) or the very accessible blog post by McElreath (2017). One of the benefits of this algorithm, and of the specific way it is implemented in `Stan`, is that these Monte Carlo samples suffer less from autocorrelation between the samples in the chains of samples (see Chapter 1 for an explanation of these terms). Thus, fewer Monte Carlo samples are needed to accurately describe the posterior distributions. In other words, the effective number of samples relative to our total amount of samples increases; see Chapter 4 for an extensive discussion on this topic. As a result, usually, convergence is obtained faster with the more efficient HMC.

To determine whether the sampling algorithm has converged, one should check the stability of the generated parameter values. A visual check of the

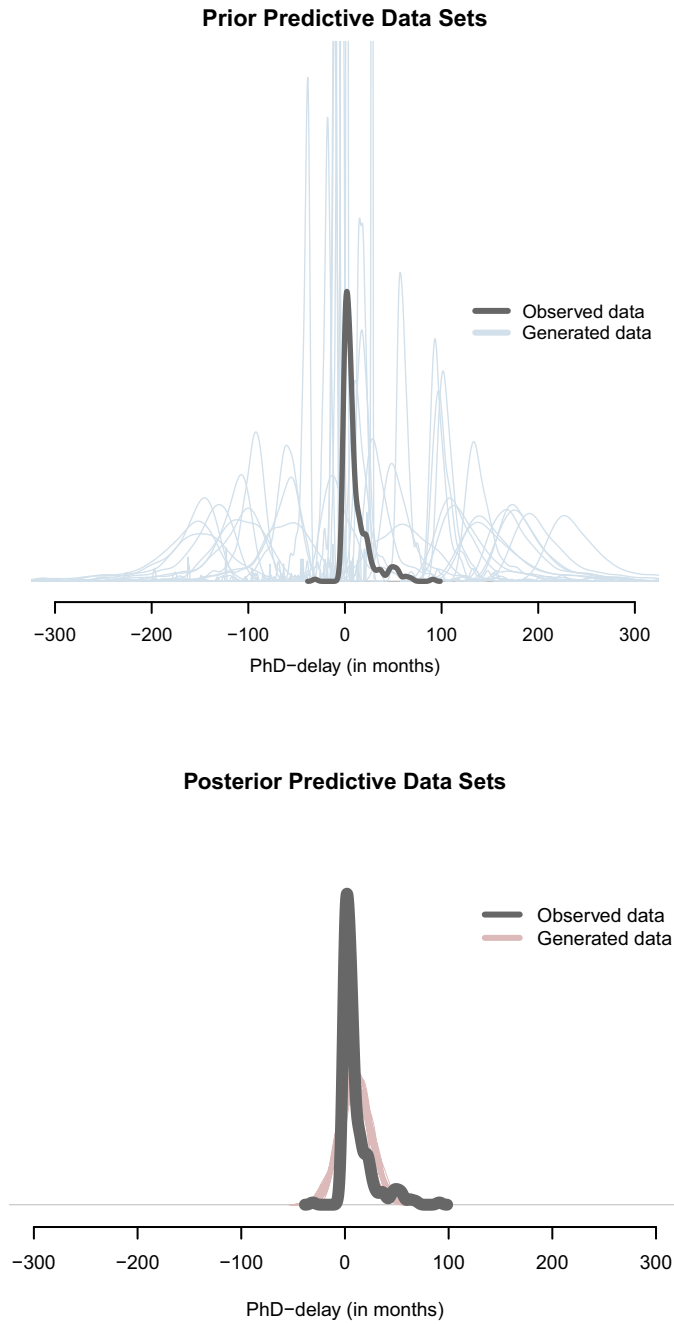

FIGURE 3.3 Generated data sets based on the prior (top panel) and posterior predictives (bottom panel)

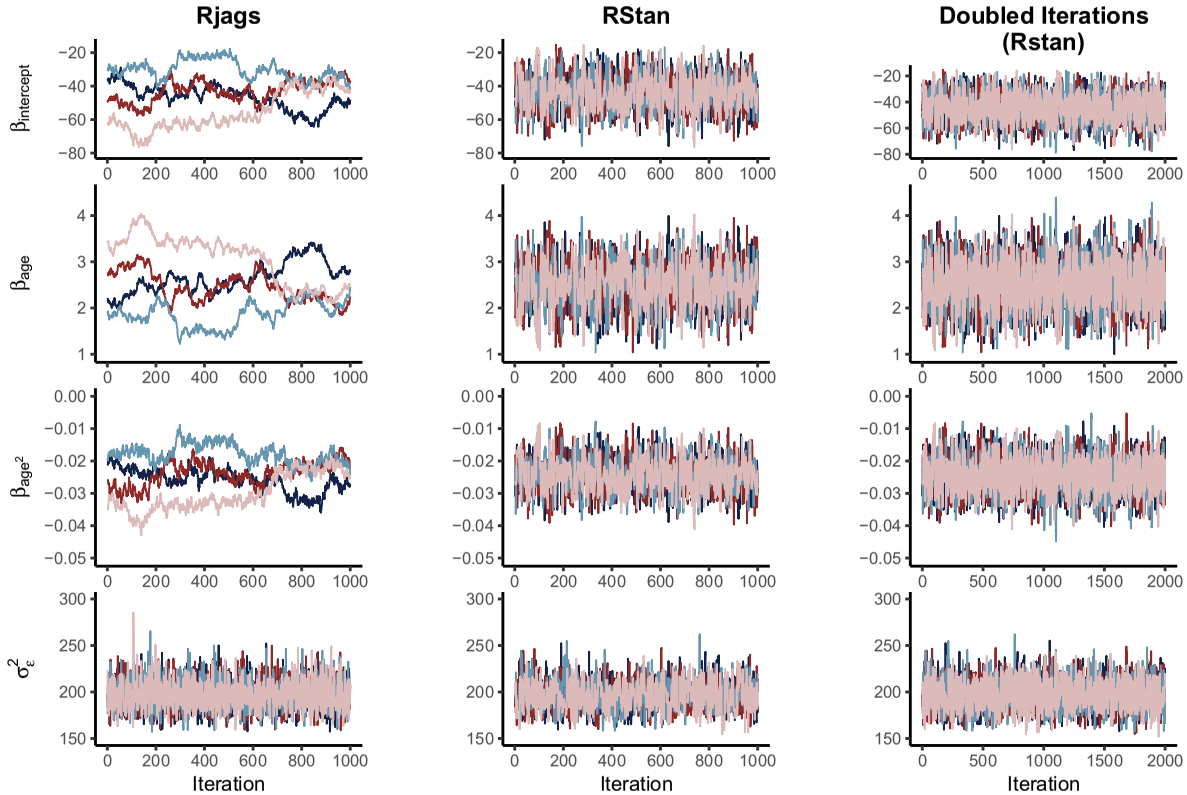

FIGURE 3.4 Trace plots for rjags and RStan with 1,000 and 2,000 iterations for each parameter in the regression model

stability of the generated parameter values implies estimating multiple chains and then plotting the results in so-called trace plots. Figure 3.4 shows the trace plots for all parameters obtained by `RStan` and `rjags`, based on four chains of 1,000 samples per chain for both samplers and 1,000 iterations burn-in. When comparing the results for `RStan` and `rjags`, it becomes clear that `RStan` is more efficient than `rjags`: we see in the plots that the chains of the MCMC sampler `rjags` move more slowly from one step to the next than in the HMC sampler `RStan`. This is caused by autocorrelation between the samples, as we show in Step 5 of the WAMBS checklist.

Next to inspecting trace plots there are several diagnostic tools to determine convergence. We discuss two completely different diagnostic tools.

First, the Gelman–Rubin statistic compares the amount of variance within the individual chains to the amount of variance between the chains up to the last iteration in the chains (Gelman & Rubin, 1992). If this ratio is close to 1—for example, if the value is smaller than 1.1 for all parameters (Gelman & Shirley, 2011)—we can be more confident that the chains describe the same distribution and that we have reached convergence. Figure 3.5 shows the development of the statistic as the number of samples increases using Gelman–Rubin diagnostic plots for both `rjags` and `RStan`.

Another convergence diagnostic is the Geweke diagnostic (Geweke, 1992), which is based on testing equality of means between the first 10% and last 50% parts of each chain. The test statistic is a standard *Z*-score: the difference between the two sample means divided by its estimated standard error. In

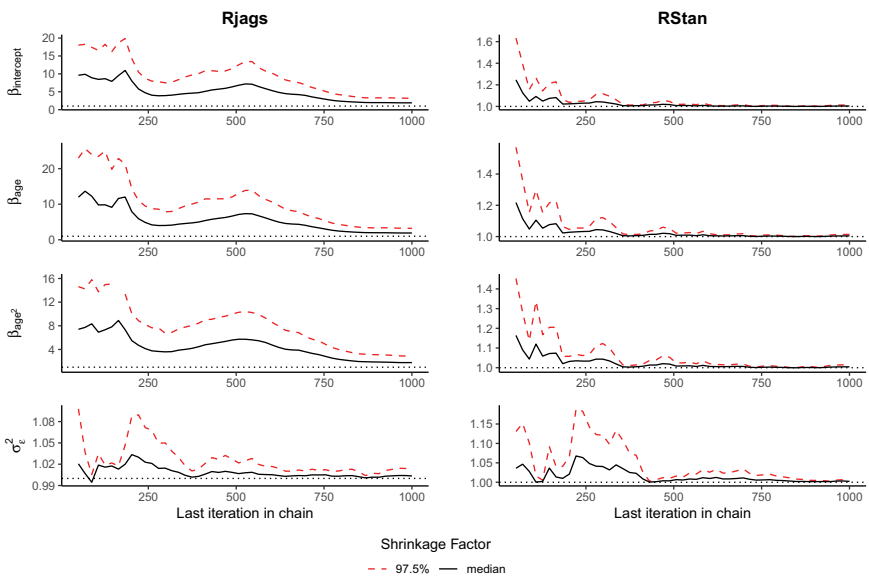

FIGURE 3.5 Gelman–Rubin statistics for `RStan` and `rjags`

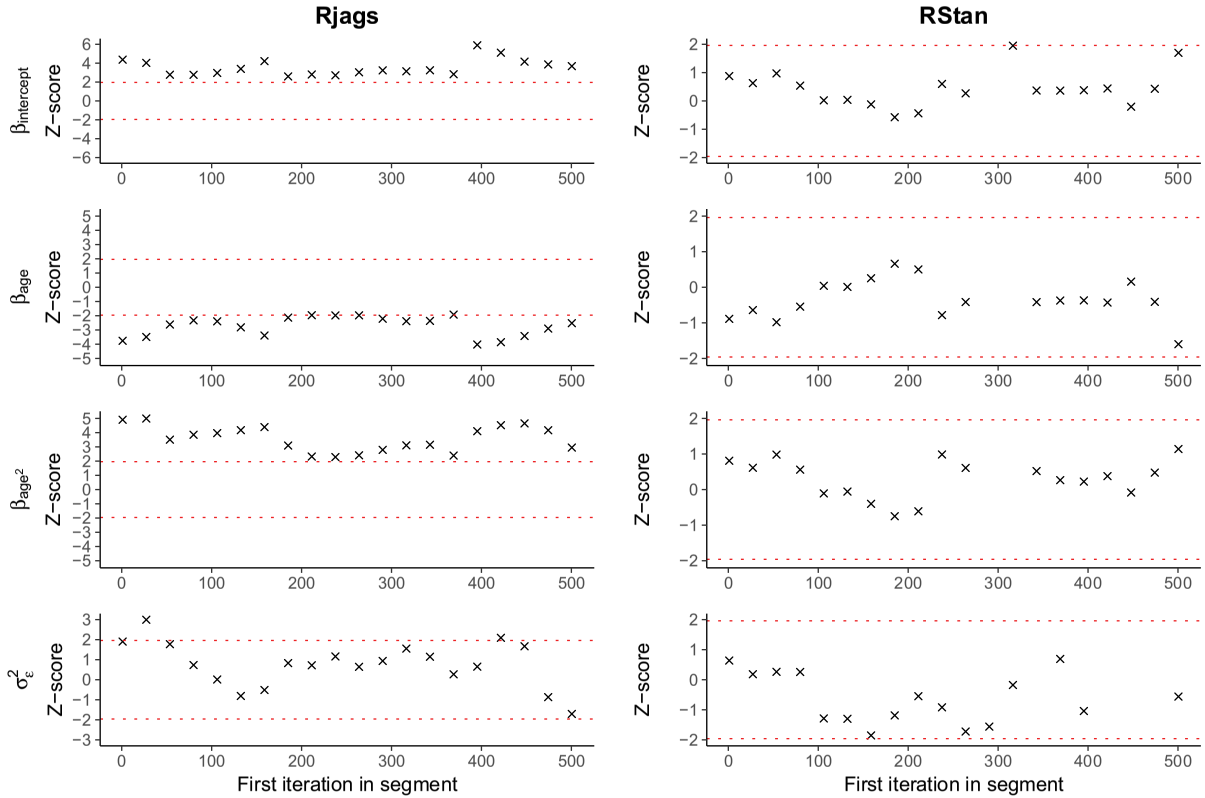

FIGURE 3.6 Geweke statistics for `RStan` and `rjags`

Figure 3.6, it could be checked how often values exceed the boundary lines of the *Z*-scores. Scores above 1.96 or below -1.96 indicate that the two portions of the chain differ significantly, and full chain convergence is not reached.

All results in Figures 3.4–3.6 point to convergence in the case of `RStan`, but not in the case of `rjags`. We continue using only `RStan`, with the exception of Step 5 in the checklist, where we compare the levels of autocorrelation between both packages to demonstrate the added value of `RStan` once more.

Does convergence remain after doubling the number of iterations?

As is recommended in the WAMBS checklist, we double the amount of iterations to check for local convergence. According to the checklist:

> Local convergence can be thought of as the case where convergence appears to be visually obtained – often with a smaller number of iterations – but when the chain is left to run longer, then the chain shifts and converges to another location.
>
> *(Depaoli & Van de Schoot, 2017)*

We re-ran the model with 2,000 samples per chain.

Next to inspecting the trace plots (see Figure 3.4) and the convergence diagnostics (available on the OSF) we can also compute the relative bias, in order to inspect if doubling the number of iterations influences the posterior parameter estimates. One can use the following equation by filling in a posterior estimate:

$$relative\ bias = 100 * \frac{|posterior\ estimate_{\ initial\ model}| - |posterior\ estimate_{\ new\ model}|}{|posterior\ estimate_{\ initial\ model}|}$$

If the relative bias is $> |5|\%$[3], then it is advised to rerun the initial model with four times the number of iterations, and again up till the relative bias is small enough; see also Chapter 4. As can be seen in the first column of Table 3.1, all

TABLE 3.1 Results of relative bias (in %) for different models

	Step 3: Double iterations	*Step 7: Different variance priors*	*Step 8: Non-informative priors*
$\beta_{intercept}$	-1.029	1.052	-6.349
β_{age}	-0.811	0.880	-5.312
β_{age^2}	-0.778	0.967	-5.578
σ^2_{ε}	-0.149	-0.345	0.248

values of relative bias are < 1.03%, which means doubling the number of iterations hardly changes the posterior estimates.

Does the histogram contain enough information?

The parameter estimates of all chains (after burn-in) can be plotted in a histogram. The amount of information, or smoothness, of the histogram should be checked to ensure that the posterior is represented using a large enough number of samples. There should be no gaps or other abnormalities in the histogram. The histograms in Figure 3.7 all look smooth, thus suggesting that adding more iterations is not necessary.

Do the chains exhibit a strong degree of autocorrelation?

The dependence between the samples of a Monte Carlo simulation can be summarized by autocorrelation. If samples are less correlated, we need fewer Monte Carlo samples to get an accurate description of our posterior distribution. High autocorrelation can be a sign that there was a problem with the functioning of the MCMC sampling algorithm or in the initial setup of the model. Also, if convergence is not obtained with an extreme number of iterations, then these issues can be indicative of a model specification problem, multicollinearity, *or* the sampling algorithm. In our case, the sampling algorithm itself solves the high amount of autocorrelation in the model. Compare

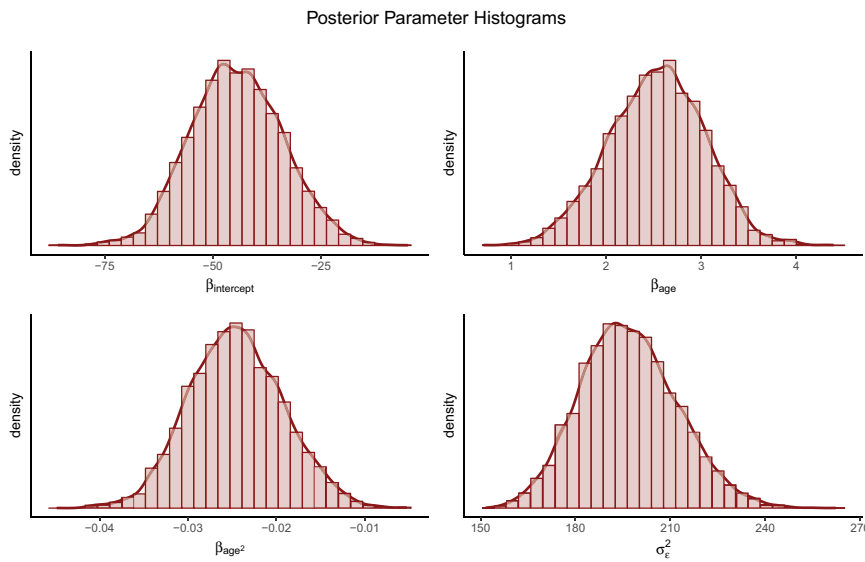

FIGURE 3.7 Plots with histograms

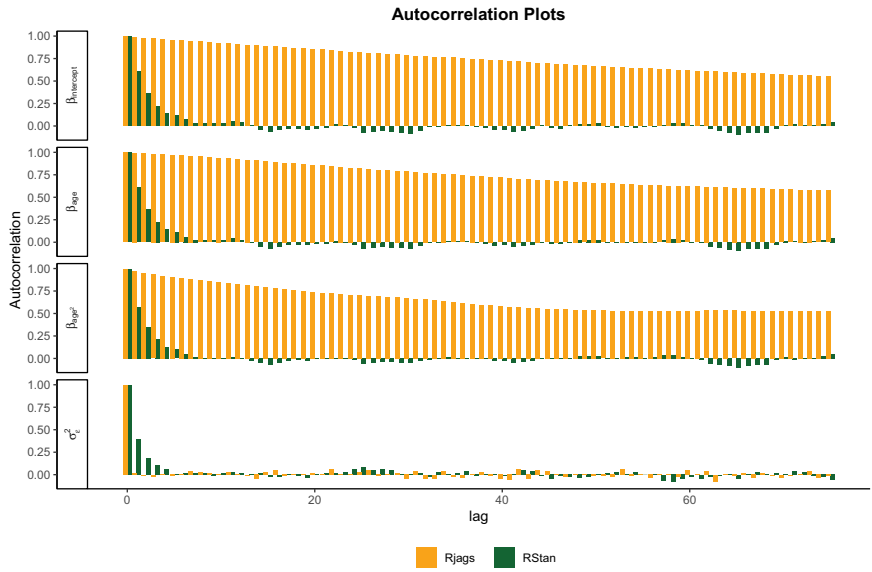

FIGURE 3.8 Plots with levels of autocorrelation for both `RStan` and `rjags`

the plots in Figure 3.8 showing high degrees of correlation across iterations obtained in `rjags` (orange) with those obtained in `RStan` (green). The results obtained in `RStan` show less dependency between iterations when compared to `rjags`.

Do the posterior distributions make substantive sense?

Plotting a smoothed line through the histogram can be used as an approximation of the posterior distribution. In Figure 3.7 we plotted these to check if they are unimodal (i.e., have one peak), are clearly centered around one value, give a realistic estimate, and make substantive sense compared to our prior beliefs. As can be seen in Figure 3.7, there are no such issues with the posterior distributions obtained for our parameters. The posterior distributions of our regression coefficients fall within the range we specified above, and the peak of our posterior distributions is within reasonable distance from the means of our prior specifications. Substantive interpretations of these posteriors will follow in Step 10 of the checklist.

Do different specifications of the priors for variance parameters influence the results?

To understand the influence of the priors as specified in Step 1, it is recommended to conduct a sensitivity analysis (Van Erp, Mulder, & Oberski,

2018). It is essential that researchers report results of a sensitivity analysis, even if there is no substantive impact on results. Do not despair if there are differences between posterior results! Such findings are actually very interesting (and even fun). In such situations, we recommend dedicating considerable space in the discussion section to the description of the discrepancy between results obtained using informative versus non-informative priors and the implications of this discrepancy. The discussion could then illustrate the mismatch between theory (i.e., priors should reflect the current state of affairs) and data, and it is up to the researcher to come up with an explanation for such a mismatch.

Although a sensitivity analysis needs you to play around with some of the prior settings, it is important to note that this can only be an exercise to improve the understanding of the priors. It is not a method for changing the original prior. That is, if a researcher actually changes the prior after seeing the results of Steps 7–9, then this is considered as manipulating the results, related to questionable research practices or even fraud.

To understand how the prior for the residual variance impacts the posterior, we compared the current results with a model that uses different hyperparameters for the Inverse Gamma prior for the residual variance. So far we used $\sigma_\varepsilon^2 \sim IG(.5, .5)$, but we can also use $\sigma_\varepsilon^2 \sim IG(.01, .01)$ and see if doing so makes a difference (many other variations are possible). To quantify the impact of the prior, we again calculated the relative bias (computed the same way as in Step 3); see the second column of Table 3.1. The results are robust, because there is only a minimum amount of relative bias for the residual variance.

Is there a notable effect of the prior when compared to non-informative priors?

In order to understand the impact of our informative priors on the posterior results, we also compare our subjective priors with non-informative priors:

- $\beta_{intercept} \sim \mathcal{N}(0, 10^6)$
- $\beta_{age} \sim \mathcal{N}(0, 1000)$
- $\beta_{age^2} \sim \mathcal{N}(0, 1000)$
- $\sigma_\varepsilon^2 \sim IG(1, .5)$

We computed the relative bias, and as can be seen in the third column of Table 3.1, there is some bias between the two models. To understand the impact of our informative priors, we plotted the priors and posteriors for both models and for all parameters in Figure 3.9. In the last column the two posteriors are plotted in the same graph, and, as can be seen, the informative priors do impact the posterior results when compared to the non-informative priors.

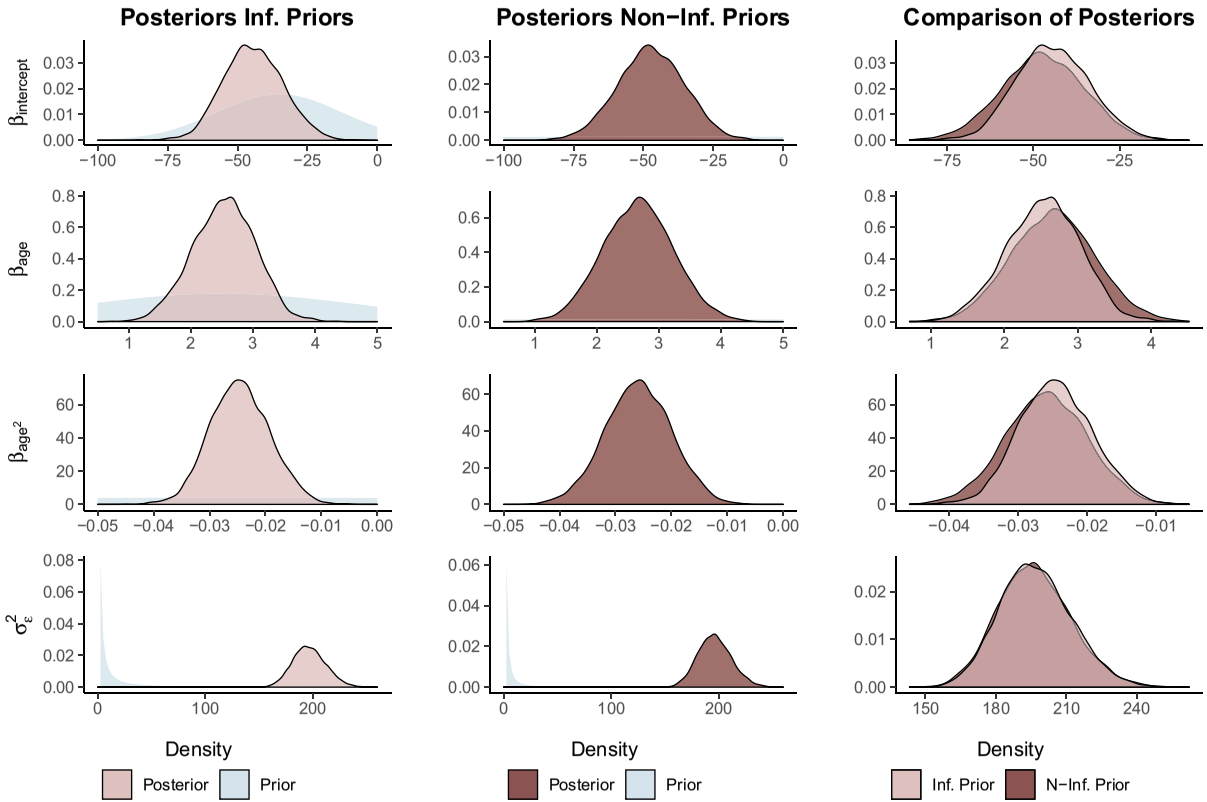

FIGURE 3.9 Priors and posteriors for the models with and without informative priors

Are the results stable with a sensitivity analysis?

In addition to the previous steps, we not only checked the sensitivity of the results to different prior specifications (our informative priors, a factor 10 times more informative, and non-informative), but we also checked the stability of the posterior estimates across participants by sequential updating. That is, with Bayesian statistics the priors can be updated with a sample size of $n = 1, \ldots, n = N$. Thus, in each step the original prior is updated with more and more data so that the posterior becomes more dominated by the data and less by the prior. The stability of the results is indicative of how small the sample size could have been with different prior settings.

As we assume our data to be exchangeable (see Chapter 2, Miočević, Levy, and Savord), it should not matter in which order the data points were observed if our posterior distributions are to pass this second sensitivity analysis. Therefore, we created five permutations of the data in which only the order of the participants was randomly changed. For each of these five data sets, we ran the model with the three different prior specifications, resulting in 15 different models. After the first update with the first participant, the posteriors were updated again with adding the second participant to the data, and so on.

As can be seen in Figure 3.10, updating the model 333 times results in similar posterior results for all the five different data sets, which makes sense since the data is exchangeable and the order in which the data is analyzed should not matter. But when inspecting, for example, the results for the intercept with the priors as specified in Step 1, it can be seen that only after roughly 100 participants are the results stable. Stated differently, if we had included only 50 PhD recipients in our data, the uncertainty in the posterior would have been much larger, even allowing zero plausibility (grey line; the blue line resembles the prior mean). This effect is much lager for the non-informative priors and a much larger data set is needed to obtain stable results. It is not surprising, however, that with precise priors (small prior variance) our data does not change the estimates much: after a few samples our posterior estimates from the permuted data sets are highly similar.

In conclusion, the data, with $n = 333$, could have been a bit smaller with our informative priors, but not much. Only with highly informative priors, the sample size could have been smaller.

TABLE 3.2 Results for the model using our informative priors

	Mean	*SD*	*2.5%*	*50%*	*97.5%*
$\beta_{intercept}$	-44.425	10.579	-64.325	-44.668	-23.387
β_{age}	2.532	.503	1.522	2.544	3.477
β_{age^2}	-.025	.005	-.034	-.025	-.014
σ^2_ε	196.923	15.266	168.758	196.255	228.166

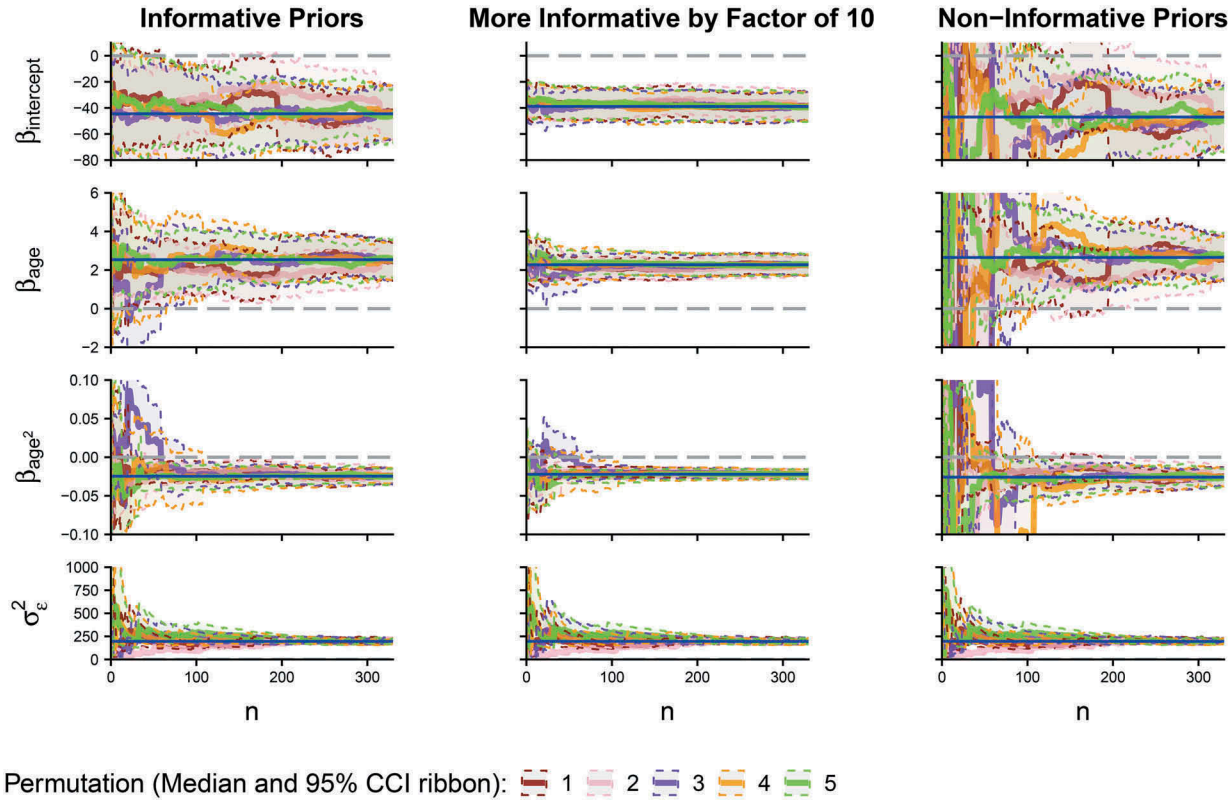

FIGURE 3.10 Results sensitivity analysis

Is the Bayesian way of interpreting and reporting model results used?

The posterior parameter estimates can be summarized using, for example, the median of the posterior distributions, and can be found in Table 3.2. Based on these point summaries, it appears the delay peaks at around the age of 50 $(2.532/-(2*.025))$. Considering that 0 is not included in the 95% interval of the linear effect and the quadratic effect, we can conclude that there is a small positive linear effect and a small negative quadratic effect[4].

It is also informative to inspect the posterior predictive results which are similar to the prior predictive results, except that because we now inspect the posteriors we can use our updated beliefs from after observing the data. If we inspect the posterior predictive plots in Figure 3.2 (bottom panel), we can see that we are not able to perfectly predict the delay in PhD completion using the candidate's age, which also becomes evident by the R^2 of 6%. Moreover, it is not surprising to see that predictions based on our model tend to be more off the mark for cases with longer and shorter delays than with normal delays, whilst our uncertain estimates capture all standard cases. Furthermore, if we compare our prior and posterior predictive distributions (see Figure 3.3), we are less uncertain and more consistent in what we expect *after* observing the data. So, accurate predictions of delay for individual cases may not be possible, but we can predict general trends at group level.

Conclusion

The chapter shows how to properly implement a Bayesian analysis following the steps of the WAMBS checklist. Following this checklist is important for the transparency of research, which is important no matter which estimation paradigm is being implemented. However, it is *even more* important within the Bayesian framework, because there are so many places where bad research practices can be "hidden" within this estimation perspective, especially concerning the prior specification and its impact on the posterior results. Clear reporting and sufficient amounts of detail for reproducing results are important first steps in ensuring that Bayesian results can be trusted and properly interpreted. We therefore recommend including results of the WAMBS checklist as an appendix or as a supplementary file to any Bayesian paper; for an example, see Zweers (2018).

In the end, properly conducting and reporting results is important, but the key is understanding the impact of the prior, especially when sample size is small, since this will ultimately be the element that potentially shifts theories and practices within a field.

Acknowledgement

This work was supported by Grant NWO-VIDI-452-14-006 from the Netherlands organization for scientific research.

Notes

1 Note that the only priors you can tweak in the app are the priors for the intercept and regression coefficients. In Step 7 of the checklist we will return to the specification of the prior for the residuals, σ_{ε}^2.
2 Historically, Inverse Gamma priors have been used for (residual) variance parameters, due to their conjugate properties that allow for using the Gibbs sampling algorithm. Some alternatives are being discussed in, for example, McNeish (2016). The hyper-parameters are, in this example, mildly informative based on the discussion in Van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, and Van Loey (2015).
3 The relative bias should be interpreted with caution and only in combination with substantive knowledge about the metric of the parameter of interest. For example, with a regression coefficient of .001, a 5% relative deviation level might not be substantively relevant. However, with an intercept parameter of 50, a 1% relative deviation level might already be quite meaningful.
4 Note that testing such results by means of the Bayes Factor is being discussed in Chapters 9 (Klaassen) and 12 (Zondervan-Zwijnenburg & Rijshouwer).

References

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv Preprint arXiv:1701.02434*.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). doi:10.18637/jss.v076.i01.

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240–261. doi:10.1037/met0000065.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. doi:10.1214/ss/1177011136.

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. P. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 116–162). Boca Raton, FL: Chapman & Hall/CRC Press.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, A. F. M. Smith, A. P. Dawid, and J. O. Berger (Eds.), *Bayesian Statistics 4* (pp. 169–193). Oxford: Oxford University Press.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

McElreath, R. (2017). Markov chains: Why walk when you can flow? Retrieved from http://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/.

McNeish, D. (2016). 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*(1), 27–56. doi:10.3102/1076998615621299.

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. doi:10.18637/jss.v085.i04.

Merkle, E. C., Rosseel, Y., Garnier-Villarreal, M., Jorgensen, T. D., Hoofs, H., & Van de Schoot, R. (2019). blavaan: Bayesian latent variable analysis, Version 0.3-4. Retrieved from https://CRAN.R-project.org/package=blavaan.

Plummer, M., Stukalov, A., & Denwood, M. (2018). rjags: Bayesian graphical models using MCMC, Version 4-8. Retrieved from https://cran.r-project.org/web/packages/rjags/index.html.

Smid, S. C., McNeish, D., Miočević, M., & Van de Schoot, R. (2019). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*. doi:10.1080/10705511.2019.1577140.

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*(1), 25216. doi:10.3402/ejpt.v6.25216.

Van de Schoot, R., Sijbrandij, M., Depaoli, S., Winter, S. D., Olff, M., & Van Loey, N. E. E. (2018). Bayesian PTSD-trajectory analysis with informed priors based on a systematic literature search and expert elicitation. *Multivariate Behavioral Research*, *53* (2), 267–291. doi:10.1080/00273171.2017.1412293.

Van de Schoot, R., Yerkes, M. A., Mouw, J. M., & Sonneveld, H. (2013). What took them so long? Explaining PhD delays among doctoral candidates. *PLoS One*, *8*(7), e68839. doi:10.1371/journal.pone.0068839.

Van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363–388. doi:10.1037/met0000162.

Zondervan-Zwijnenburg, M. A. J., 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*(4), 305–320. doi:10.1080/15427609.2017.1370966.

Zweers, I. (2018). Chapter 5. Similar development in separate educational contexts? Development of social relationships and self-esteem in students with social-emotional and behavioral difficulties in inclusive classrooms and exclusive schools for special education – Supplementary materials. Retrieved from osf.io/yf3mu.