Improving Transparency and Replication in Bayesian Statistics:

The WAMBS-Checklist

Sarah Depaoli[1*]

Rens van de Schoot[2,3*]

Cite as:

Depaoli, S., and van de Schoot, R. (in press). Improving transparency and replication in Bayesian Statistics: The WAMBS-Checklist. *Psychological Methods*.

[1] Psychological Sciences, University of California, Merced

[2] Utrecht University, Department of Methods and Statistics, The Netherlands

[3] North-West University, Optentia Research Program, Faculty of Humanities, South Africa

Correspondence should be addressed to Sarah Depaoli, Assistant Professor, School of Social Sciences, Humanities, and Arts, University of California, Merced, 5200 N. Lake Road, Merced, CA, 95343. Email: sdepaoli@ucmerced.edu.

Author Notes:

The second author was supported by a grant from the Netherlands organization for scientific research: NWO-VENI-451-11-008.

Abstract

Bayesian statistical methods are slowly creeping into all fields of science and are becoming ever more popular in applied research. Although it is very attractive to use Bayesian statistics, our personal experience has led us to believe that naively applying Bayesian methods can be dangerous for at least three main reasons: the potential influence of priors, misinterpretation of Bayesian features and results, and improper reporting of Bayesian results. To deal with these three points of potential danger, we have developed a succinct checklist: the WAMBS-checklist (When to worry and how to Avoid the Misuse of Bayesian Statistics). The purpose of the questionnaire is to describe 10 main points that should be thoroughly checked when applying Bayesian analysis. We provide an account of "when to worry" for each of these issues related to: (a) issues to check before estimating the model, (b) issues to check after estimating the model but before interpreting results, (c) understanding the influence of priors, and (d) actions to take after interpreting results. To accompany these key points of concern, we will present diagnostic tools that can be used in conjunction with the development and assessment of a Bayesian model. We also include examples of how to interpret results when "problems" in estimation arise, as well as syntax and instructions for implementation. Our aim is to stress the importance of openness and transparency of all aspects of Bayesian estimation, and it is our hope that the WAMBS questionnaire can aid in this process.

Key-words: Bayesian estimation; prior; sensitivity analysis; convergence; Bayesian checklist

Improving Transparency and Replication in Bayesian Statistics:

The WAMBS-Checklist

Bayesian statistical methods are slowly creeping into all fields of science and are becoming ever more popular in applied research. Figure 1 displays results from a literature search in Scopus using the term "Bayesian estimation" and, as can be seen, the number of empirical peer reviewed papers using Bayesian estimation is on the rise. This increase is likely due to recent computational advancements and the availability of Bayesian estimation methods in popular software and programming languages like WinBUGS and OpenBUGS (Lunn, Thomas, Best & Spiegelhalter, 2000), MlWiN (Browne, 2009), AMOS (Arbuckle, 2006), M*plus* (Muthén & Muthén, 1998-2015), BIEMS (Mulder, Hoijtink, and de Leeuw, 2012), JASP (Love et al., 2015), SAS (SAS Institute Inc., 2002-2013), and STATA (StataCorp., 2013). Further, there are various packages in the R programming environment (Albert, 2009) such as STAN (Stan Development Team, 2014) and JAGS (Plummer, 2003) that implement Bayesian methods.

When to use Bayesian Statistics

There are (at least) four main reasons why one might choose to use Bayesian statistics. First, some complex models simply cannot be estimated using conventional statistics (see e.g., Muthén & Asparouhov, 2012; Kruschke, 2010, 2011; Wetzels, Matzke, Lee, Rouder, Iverson & Wagenmakers, 2011). Further, some models (e.g., mixture or multilevel models) require Bayesian methods to improve convergence issues (Depaoli & Clifton, 2015; Skrondal & Rabe-Hesketh, 2012), aid in model identification (Kim, Suh, Kim, Albanese, & Langer, 2013), and produce more accurate parameter estimates (Depaoli, 2013; 2014). Second, many scholars prefer Bayesian statistics because they believe population parameters should be viewed as random (see e.g., Dienes, 2011; Van de Schoot, Hoijtink, Mulder, Van Aken, Orobio de Castro, Meeus, Romeijn, 2011). Third, with

Bayesian statistics one can incorporate (un)certainty about a parameter and update this knowledge through the prior distribution. Fourth, Bayesian statistics is not based on large samples (i.e., the central limit theorem) and hence may produce reasonable results even with small to moderate sample sizes, especially when strong and defensible prior knowledge is available (Depaoli & Scott, in press; Hox, van de Schoot, & Matthijsse, 2012; Moore et al., 2015; van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & van Loey, 2015; Zhang, Hamagami, Wang, & Nesselroade, 2007).

For a full introduction to Bayesian modeling, we refer the novice reader to, among many others: Bolstad (2007); Carlin and Louis (2009); Christensen, Johnson, Branscum, and Hanson (2010); Depaoli and Boyajian, (2014); Gelman and Hill (2007); Kaplan, 2014; Kruschke (2010); Jackman (2009); Lynch (2007); Ntzoufras (2009); or van de Schoot and Depaoli (2014). Likewise, a more technical introduction can be found in Gelman, Carlin, Stern and Rubin (2004), Lee (2007), or Press (2003).[1]

Making Decisions when Implementing Bayesian Methods

Although it is very attractive to use Bayesian statistics, estimating models within this framework involves making some nontrivial decisions throughout the estimation process. Likewise, these decisions can become increasingly more complex to judge based on the complexity of the model being estimated. Our personal experience has also led us to believe that naively applying Bayesian methods can be dangerous for three main reasons. First, Bayesian statistics makes use of (subjective) background knowledge formalized into a, so-called, prior distribution. The exact influence of the prior is often not well understood and priors might have a huge impact on the study results, which requires careful consideration we detailed in subsequent sections. Second, akin to many elements of frequentist statistics, some Bayesian features can be easily misinterpreted. As is

[1] For a comprehensive list of introductory, intermediate, and advanced readings on Bayesian statistics, see the following website: http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_introbayes_sect011.htm.

true with any statistical paradigm, misleading inferences can be drawn if results are not interpreted precisely. A danger here is that most statistical training, at least in the field of psychology, comes from a frequentist approach. In addition, without proper training, it may be that interpretations of Bayesian statistics can be confused with those in the frequentist framework. Third, reporting on Bayesian statistics follows its own rules since there are elements included in the Bayesian framework that are fundamentally different from frequentist settings. Given that Bayes is only slowly increasing its presence in the methodological and applied literature, there is not a strong precedence for how to report results; this became even more evident when we noted most of the papers we found from Scopus cited earlier failed to report each of the 10-points we have deemed important and detailed below. These points have been described in Bayesian textbooks and papers implementing Bayesian methodology. However, to our knowledge, there is no succinct summary of these important diagnostics that appears in a single source.

We conducted a systematic review on applied Bayesian papers published in psychology (see van de Schoot, Ryan, Winter, Zondervan-Zwijenburg, & Depaoli, under review). In this review, we discovered that the majority of papers we reviewed (99 empirical Bayesian papers were deemed eligible for the review) did not properly report important issues surrounding Bayesian estimation. For instance, 55.6% of the papers did not report the hyperparameters specified for the prior, 56.6% did not report checking for chain convergence, and 87.9% did not conduct a sensitivity analysis on the impact of priors. We address all of these issues here and discuss the importance of being completely transparent when reporting Bayesian results.

To deal with these points of potential danger, we have developed a succinct checklist: the WAMBS-checklist (When to worry and how to Avoid the Misuse of Bayesian Statistics), see Figure 2. The purpose of the Checklist is to describe 10 main points that should be thoroughly examined

when applying Bayesian analysis. We provide an account of "when to worry" for each of these issues related to: (a) issues to check before estimating the model, (b) issues to check after estimating the model but before interpreting results, (c) understanding the influence of priors, and (d) actions to take after interpreting results. To accompany these points of concern, we will present diagnostic tools that can be used in conjunction with the development and assessment of a Bayesian model.

Intended Audience, Scope, Outline for the Current Paper, and Limitations

Intended audience. The intended audience of this paper consists of applied researchers who are implementing Bayesian techniques, novice Bayesian users, or PhD students wanting to implement Bayes. As a motivating example, take a situation where a PhD student wants to use Bayesian methods to solve an applied problem. In this situation, the supervisor is unfamiliar with such techniques and cannot be of direct help for diagnosing and solving problems with estimation and priors. The current paper can be used as a guide and tutorial for checking and diagnosing "problems" with priors as to ensure that the PhD student is proceeding with data analysis appropriately, even if the supervisor is unable to help with this process. We do not intend this paper to replace formal, proper training in implementing and interpreting Bayesian statistics. Although we hope this paper can act as a guide to proper use of these methods, thorough training is essential when conducting Bayesian methods.

One important warning regarding the use of this checklist is that fully addressing the items is likely to be a nontrivial task, especially for the novice user of Bayesian methods. For example, some of the first stages of the checklist could take several months or more to adequately address. However, we believe that the time investment is necessary and that the quality and replicability of results are ensured once this Checklist has been properly implemented.

Scope of the paper. In order to keep the current paper as general as possible with respect to implementing Bayesian methods, there are several concepts that we will be focusing on and several that we will not specifically address outside of providing references of additional sources. Throughout this paper, due to space considerations, we will assume that the user of these methods is estimating one particular model with one set of priors. We are assuming researchers are using any general model that implements MCMC (with any sampling method—e.g., Gibbs or Metropolis-Hastings), where the number of iterations in the chain is known. We are also assuming the researcher is implementing user-known or user-specified priors that can be freely altered within the software; note that some Bayesian programs do not allow the user to directly altering priors (e.g., the JASP and BIEMS programs).

There are many specialized topics in traditional statistical modeling that are also important to address under the Bayesian framework such as Bayesian model fit, missing data, model specification, model identification, and parameterization. These specialized topics are beyond the scope of the current paper, but we refer the reader to more technical sources such as Gelman et al., (2004) and Lee (2007) for details on such topics. Finally, we recognize that a variety of open-source and commercial software programs can be used to implement Bayesian methods. To keep this discussion as encompassing as possible, we do not focus on any particular software; however, we do at times make note of specific features to be aware of in different programs. Supplementary material representing work from a variety of software programs is presented online to aid in implementing various topics discussed and can be found here: www.sarahdepaoli.com. This material includes a variety of resources to aid in implementing the recommendations presented here. We include the following types of information separated into 7 Folders of content online:

- Examples for using 9 different Bayesian software programs (e.g., AMOS, JAGS, JASP, BUGS, and STAN, to name a few),

- A detailed exercise document walking through the use of the WAMBS checklist,
- All code, output, and data for the examples provided here, and
- A step-by-step set of directions for implementing the PSRF convergence diagnostic (discussed in Point 8) for assessing sensitivity analysis results.

Outline of the current paper. The current paper includes 10 main points, comprising the WAMBS-checklist, to consider when implementing Bayesian statistics. These points are broken down into four main categories: (a) To be checked before estimating the model—Point 1; (b) To be checked after estimation but before inspecting model results—Points 2-6; (c) Understanding the exact influence of the priors—Points 7-9; and (d) Interpretation of model results—Point 10. Within each of these main categories, we provide background information necessary for understanding each Point listed, and we provide a simple example to show how to fill in the WAMBS-checklist. Next, we present the individual Points, where we include a description of the issue, a description of what output should be provided to the supervisor for checking this Point, details of when to worry about certain outcomes, and guidelines for when a Bayesian expert should be consulted.[2] The information required from each of these sections can be addressed through the WAMBS-checklist provided in Figure 2.

In many places throughout the paper, we include small examples illustrating the different issues the can arise when applying the Checklist. Many of the examples provided use one of two main datasets to highlight issues such as convergence, priors, and sensitivity analysis. The first dataset contains longitudinal information on burn victims (see Van Loey, Maas, Faber, & Taal, 2003) and was selected because previous work on these data (e.g., Depaoli et al., 2015) showed problems in obtaining convergence and stable estimates. The second dataset is from the large-scale Early Childhood Longitudinal Study-Kindergarten class (NCES, 2001) database, where we illustrate how

[2] We are using the broad term of 'expert' here to capture the fact that Bayesian experts come from a variety of sub-disciplines. For the sake of the current paper, 'expert' refers to anyone with ample expertise in Bayesian statistics to advise on the current topics being addressed. This group may include statisticians, psychometricians, quantitative psychologists, education statisticians, or those with Bayesian expertise from other disciplines.

priors can be derived and thoroughly examined. All examples are available in the online material and we also provide an example online where a single dataset is used to illustrate the entire Checklist from start to finish (see Folder 7 in online material).

Limitations of the current paper. The Checklist detailed next should be followed precisely in order to ensure proper implementation and reporting of Bayesian methods. From a practical point of view, using this Checklist will be time-consuming and require a great deal of critical thinking and decision making—both of which will improve the quality of the work being presented. However, the implementation of this Checklist will not ensure that every aspect of the modeling process has been properly conducted. This paper should be viewed as a tool that can be used to improve clarity and replication of results when implementing Bayesian methods. It is not our intention to focus the paper on early errors that could have been committed—for example, errors in the data collection or model-building phase. Of course, it is possible that the incorrect model or set of priors can be chosen before the Checklist is addressed. In addition, it is important to recognize that errors at earlier stages (e.g., selecting an inappropriate model to estimate) will impact subsequent phases of the model estimation process addressed in this Checklist. The current paper should be used in conjunction with other tools and knowledge that can help the user avoid errors or mistakes that are beyond the scope here. Although the focus of this paper is not about early phases of why the model looks the way it does, or if the prior information came from the "proper place", using this Checklist will aid in proper review and dissemination of work because the estimation process will be completely transparent.

Stage 1: To be Checked before Estimating the Model

Background Information on Priors

When specifying priors, it is important to recognize that prior distributions fall into three main classes related to the amount of (un)certainty they contribute to the model about a given parameter: (1) *non-informative priors*, (2) *weakly-informative priors* and (3), *informative priors*.[3] Before describing each of these categories, we note that levels of informativeness fall on a continuum and are defined subjectively in line with the metric and scale of the particular parameter under study. For example, a prior (e.g., Uniform[0,1]) may be quite informative for one parameter (e.g., an intercept for a growth model measured on a continuous metric) and quite non-informative for another (e.g., a parameter on the probability scale); the level of informativeness is dependent on the scale of the parameter. We therefore recommend to use graphs to visualize how well the prior maps onto the scale of the parameter; Folder 1 of the Supplementary Material presents examples for doing this.

First, non-informative priors represent a complete lack of knowledge about the value of the parameter being estimated. A non-informative prior is typically denoted by a distribution that places an equal probability for each possible value under that distribution. Typically, a non-informative prior would be represented by a distribution with a relatively flat density, where the different values the parameter can take on have approximately equal likelihood under the distribution. If, for example, a continuous intercept for a growth model was being estimated, a non-informative prior might be Normal(0,10^{10})—of course, depending on the scale of the parameter. This prior distribution is centered at zero and has a very wide variance of 10^{10}, which provides complete ambiguity about the parameter value.

The next level of informativeness represents a prior distribution that holds some useful information, but that does not really influence the final parameter estimate to a large degree. These

[3] The term "non-informative prior" refers to the case where researchers supply vague information about the population parameter value; the prior is typically defined with a very wide variance (Gill, 2008). Although "non-informative" is one term commonly used in the Bayesian literature to describe this type of prior (see e.g., Gelman et al., 2004), other phrases such as "diffuse" (see e.g., Gill, 2008), or "flat" (Jeffreys, 1961) are also used to describe this type of prior. We use "non-informative" and "diffuse" interchangeably in the current paper.

prior distributions are referred to as *weakly-informative* priors. A weakly-informative prior is perhaps more useful than a strictly non-informative prior since some information is conveyed within the distribution. Essentially, weakly-informative priors do not supply any strict information, but yet are still strong enough to avoid inappropriate inferences that can be produced from a non-informative prior (Gelman, Jakulin, Pittau, & Su, 2008). Taking the same example of a continuous intercept for a growth model, a weakly-informative prior might be Normal(50,15). Perhaps the researcher knows an approximate intercept and also knows negative starting points of the growth trajectory are unlikely. The mean for this prior is set at 50 and the variance is 15. This distribution still covers a relatively wide-range of values, but it is allowing for quite a bit of variation surrounding the center of the distribution. In the case of some model parameters, such as a growth model intercept, the researcher will have information about the possible range of values for the intercept from basic descriptive statistics of the growth data. In this case, the researcher could use that information to help construct the weakly-informative prior that covers the range of possible values for the parameter. It may even be helpful for the researcher to plot the potential prior on a graph, where the x-axis relates to the scale of the parameter. In this case, seeing how the prior maps onto the scale of the parameter could be very insightful when constructing the prior.

The other end of the spectrum includes prior distributions that contain strict numerical information that is crucial to the estimation of the model. These priors are often referred to as *informative* prior distributions. Specifically, the hyperparameters for these priors (e.g., the prior mean and prior variance) are specified to express particular information reflecting a greater degree of certainty about the model parameters being estimated. In the case of our growth model intercept, an informative prior could be Normal(50,2). In this case, the researcher implies through the prior that the intercept is very close to 50 since the variance is set to such a small value. The information embedded in the informative prior can come from a variety of places, which is referred to as prior

elicitation (O'Hagan et al., 2006; Van Wesel et al., 2011). Some elicitation strategies include the following techniques. First, the researcher can ask an expert, or a panel of experts, to provide an estimate for the hyperparameters based on knowledge of the field; see, for example: Bijak and Wisniowski (2010); Fransman, et al., (2011); Howard, Maxwell, and Fleming (2000); Martin et al., (2012); and Morris, Oakley, and Crowe (2014). Second, the researcher can use the results of a previous publication as prior specification (Kaplan, & Depaoli, 2013). Third, they can use the results from a meta-analysis to define hyperparameter values for the prior, where multiple studies are combined to form information about the parameter (Ibrahim, Chen, & Sinha, 2000; Rietbergen et al., 2011). Fourth, a pilot study can be used with the same population of interest and a sampling method can be implemented to obtain an estimate for the parameter that can then be used to define a prior for a subsequent data set (Gelman, Bois, & Jiang, 1996). Finally, data-based priors can be derived based on a variety of methods including maximum likelihood (see, Berger, 2006, Brown, 2008, Candel & Winkens, 2003, van der Linden, 2008) or sample statistics (see e.g., Darnieder, 2011; Raftery, 1996; Richardson & Green, 1997; Wasserman, 2000). Note that there are some arguments against using such "double-dipping" procedures where the sample data are used to derive priors and then used in estimation; we refer the reader to Darnieder (2011) for more details on this topic.

Much research has indicated that priors can have an impact on parameter estimates and therefore also on substantive findings; for details on the different ways in which priors can adversely impact findings, see: Depaoli (2013); Gelman and Shalizi (2013); Johnson (2013); Seaman, Seaman, and Stamey (2012); and van de Schoot and Depaoli (2014). Moreover, whether the priors typically used as non-informative (or informative) priors are actually *acting* as non-informative (versus informative) priors has not been fully examined in the methodological literature. There is reason to believe that such supposedly non-informative priors may in fact be acting in an expectedly informative way. For example, a prior specification that is truly non-informative may have an

adverse impact on final parameter estimates via the posterior, especially when sample sizes are small; see Lambert, Sutton, Burton, Abrams, and Jones (2005) for a meta-analytic example of this. Alternatively, a prior that is meant to be non-informative but is actually acting as informative can have unintended effects on the posterior (see, Gelman, 2006). For example, a Dirichlet prior of D(10,10) for a 2-class mixture model can distort the posterior and push the classes to be equal in size even if they are far from equal (Depaoli, 2013); note that this is the default "non-informative" prior in M*plus*. Likewise, diffuse or non-informative priors can have an adverse impact on parameters that are transformed. Specifically, Seaman, Seaman, and Stamey (2012) showed that diffuse (non-informative) priors used in logistic regression actually acted as informative priors once the logit transformation was computed. The result was that the diffuse priors had an unintended impact on the posterior of the transformed parameters, which were the parameters ultimately interpreted in the model. As a result of the impact that even "default" diffuse priors can have, it is important to indicate and justify when default univariate priors are implemented in the data analysis process.[4]

Given that priors in general may have a rather large impact on final estimates, especially when sample sizes are small, it is important to understand the priors used in the model under investigation. That is, if a researcher specifies prior distributions, the results are affected by the subjective choices a researcher makes. The question is how much the results are influenced and whether the influence is wanted or unwanted. If a researcher uses Bayesian estimation without exactly understanding the role of the prior distribution, then the results and conclusions (!) might be impacted in a manner that makes them invalid. Therefore, priors can be dangerous and researchers should always convince their supervisors, the editor, and the reviewers the impact that the prior is having on the final conclusions. When using our diagnostic tool, the exact effect of the priors on the results can be

[4] Additional information will be provided in a subsequent section about common, default multivariate priors for covariance matrices.

uncovered, as well as whether the influence of the prior is wanted or unwanted. In conclusion, the first point of our diagnostic tool seems rather intuitive and simple, but the importance of understanding your priors cannot be stressed enough given the potential impact that it may have on conclusions.

Point 1: Do you understand the priors?

Item description. In order to convey your understanding of your prior, you must address five different points. First, one needs to specify the distributional form of the priors (e.g., normal, inverse gamma, etc)[5]. For a list of possible types of priors, see Appendix A on page 573-577 in Gelman et al. (2004). Second, the researcher must decide whether they will use conventional or "default" priors, which we also refer to as non-informative in this paper.[6] This distinction in the type of prior is rooted in whether Bayesian estimation is used as a method that incorporates previous knowledge into the estimation process (via weakly or informative priors) or simply as just another estimator (via non-informative priors); for a discussion on this topic, see Press (2003). Third, if weakly-informative or informative priors are used, then the researcher must include information about where the background knowledge used to form the prior came from, see O'Hagan et al. (2006) for more details on prior elicitation. Fourth, the researcher must include visual plots depicting weakly-informative and informative priors. Plotting priors can help to visually detect levels of informativeness. Many programs such as M*plus*, Amos, the R programing environment, and many online web tools can be used to plot priors—including the code we provide in the online Supplementary Material in Folder

[5] First, the researcher must select the distributional family and then, within the family, the specific form of the distribution is selected.

[6] We also note that the researcher will need to decide whether conjugate priors (those of the same parametric form as the posterior) are used or not. Conjugate priors are convenient for interpretation since the posterior will follow a known distribution form (Gelman et at., 2004; Gill, 2008). Although computational advances no longer require the need for conjugate priors, some models (e.g., finite mixture models) require them to speed up mixing time and aid in proper convergence. This topic of conjugacy is largely beyond the scope of this paper, but we refer the reader to Gelman et al. (2004) and Gill (2008) for more details about conjugate priors.

2. Fifth, specific hyperparameter values must be determined and reported for all priors. This final request is perhaps the most difficult because it is tied to the issue of prior elicitation so it is important to be meticulous and thorough when determining hyperparameter values.

What to show to your supervisor. Table 1 provides an example of how to summarize these five points. Perhaps the most important portion of this table is where the information used to determine the hyperparameters came from. This information is especially important to cover with one's supervisor to ensure that the correct source of information was used to construct the prior. If the user is unsure of how to solicit or specify a certain prior, then the supervisor should be consulted during the prior-specification phase as well.

When to worry. Worry if you cannot fill in all of the information for Table 1.

When to ask an expert. If after going through these recommendations, reading the literature, talking to the supervisor, and following a Bayesian course the researcher still has questions about the elicitation process, then an expert can be consulted to help.

Stage 2: To be Checked after Estimation but Before Inspecting Model Results

Background Information on Convergence

All textbooks introducing Bayesian statistics caution users to always inspect the trace plots (Bolstad, 2007; Carlin & Louis, 2009; Christensen, Johnson, Branscum, Hanson, 2010; Gelman, Carlin, Stern & Rubin, 2004; Gelman & Hill, 2007; Jackman, 2009; Lynch, 2007; Ntzoufras, 2009). This section is about the importance of these plots and how to assess them. After specifying the prior distribution and entering the data into the software, the posterior distribution needs to be obtained. To approximate the posterior, often the Gibbs sampler is used; although, other samplers

can also be implemented here.[7] The idea behind the Gibbs sampler is that the conditional distribution of one set of parameters given other sets can be used to make random draws of parameter values (see for more information about the Gibbs sampler: Geman and Geman, 1984; Casella & George, 1992). This process results in an approximation of the joint distribution of all the parameters. The Gibbs sampler consists of t iterations ($t = 1, \ldots, T$) to obtain new values in each step drawing from a conditional posterior parameter distribution. Typically, a large number of iterations are performed to construct the posterior distribution. If we plot the estimates of all iterations after burn-in (the iterations discarded before convergence is obtained), then a histogram is obtained. It is typically desired to visually depict the samples pulled from the posterior, and the histogram or a Kernel density plot can be used to visually represent the samples.

Before inspecting the Kernel density plot there is one issue of high importance: namely, convergence of the trace plot. As previously mentioned, after running enough iterations, the Gibbs sampler converges to the posterior distribution of interest. Theoretical results imply that the Gibbs sampler always converges if run long enough. The solution to when convergence is not met, however, is simple providing proper specification of the model: use more iterations and only use that part of the chain which has reached convergence. However, the question is how many iterations to use and as such, how to determine convergence of our statistical chains?

The decision of whether a chain has converged can be based on statistical criteria, but should always be accompanied by a visual inspection of the trace-plot, as will become clear below. Although Sinharay (2004) and others (see e.g., Brooks and Roberts, 1998) discuss several diagnostic tools to determine convergence, there is no consensus which statistical criterion can be considered as the 'best' one. Much of this lack of consensus is due to the fact that the various convergence criteria

[7] We generalize to the Gibbs sampler here, but the same issues we discuss will arise with other samplers. In turn, the same solutions we suggest can also be implemented.

focus on different aspects of chain convergence; it is much more difficult to assess convergence in distribution than convergence to a particular number (akin to maximum likelihood via the EM algorithm). However, it is also important to note that convergence is still equally important and sometimes difficult to assess for maximum likelihood estimation.

To determine whether the algorithm has converged, one should check the stability of the generated parameter values. A visual check of the stability of the generated parameter values implies estimating multiple chains (when possible), where each chain starts at a disparate place in the parameter space. Then the researcher should visually observe from which iteration onwards the generated parameter values display a stable pattern in the mean and in the variance of the parameter across chains (i.e., the mean of the chain is stable and the variance, or fluctuations in the chain is stable). Note that this visual check should be carried out for each and every estimated parameter, even if the parameter is not of particular substantive interest.

It is important to note that there are several other commonly implemented convergence diagnostics in programs such as M*plus* and R; for example, the Geweke diagnostic (Geweke, 1992), the Heidelberger and Welch diagnostic (Heidelberger & Welch, 1983), and the Raftery and Lewis diagnostic for determining the length of the burn-in and post-burn-in portions of the chain (Raftery & Lewis, 1992). The interested reader is referred to Kaplan and Depaoli (2012), Sinharay (2004), or Kim and Bolt (2007) for an overview of additional convergence diagnostics commonly implemented in the Bayesian literature. At times, we will also suggest implementing some of these techniques to satisfy the Checklist.

Point 2: Does the trace-plot exhibit convergence?

Item description. For each parameter estimated in the model, extract the trace plot and put it in Table 2, column 2. For each trace plot, one must visually inspect chain convergence (i.e., the mean and the

variance of the chain show stability). If the visual inspection does not show chain convergence, then run more iterations and increase the burn-in phase. The number of iterations should be increased until all of the parameters in the model show visual convergence in the trace plots. If the number of iterations has been increased and convergence still has not been obtained, then perhaps there are still not enough iterations. We recommend having at least 10,000 iterations in the burn-in phase and 10,000 iterations in the post burin-in phase, but some complex models (e.g., multilevel or mixture models) may require up to 500,000 or one million iterations in the burn-in phase. These are very rough guidelines, but our point is that the researcher should be open to the idea that the chain length necessary to converge (i.e., one with a stable mean and stable variance) may be very long.[8] Most of the time, non-convergence can be remedied by increasing the length of the chain. However, if running the chain for a large number of iterations does not yield convergence, then consider changing starting values or altering the model. Mathematically, every model will converge to the target distribution, and if convergence is not obtained after going through these recommendations then there may be another issue causing problems (e.g., model mis-specification or a model that is not identified).

What to show to your supervisor. Show the supervisor Table 2, column 2 with converged trace plots for every parameter in the model. It may be that in order to fill in this table with converged trace plots for every parameter that you will have to rerun the model several times using different lengths of burn-in and post burn-in portions of the chain to obtain visual convergence for every parameter.

When to worry. Worry if at least one trace plot does not show convergence after implementing the suggestions listed above.

[8] If a very large number of iterations are required, and this number is unreasonable to compute, then the researcher may consider using a different estimation algorithm or Bayesian program to help speed computation time.

When to ask an expert. If substantially increasing the number of iterations (e.g., up to two million iterations) does not solve the issue, then an expert should be consulted.[9]

Point 3: Does convergence remain after doubling the number of iterations?

Item description. Once visual convergence appears to have been established through the trace plots, a second check of convergence is necessary using: (1) another visual check, (2) a convergence diagnostic, and (3) computation of relative deviation. This second check is specifically to avoid obtaining what we call *local convergence*. 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.

In order to check for local convergence, rerun the model with twice the number of iterations. As an example, see Table 3, where we show results for two different chains. Example data used to illustrate this point were based on a reanalysis of longitudinal data presented in Van Loey, Maas, Faber, and Taal (2003). Specifically, we estimated a 4-class Bayesian latent growth mixture model examining different trajectories representing posttraumatic stress disorder (PTSD) changes over the course of a year following a traumatic burn event. Initial model estimation using 6,000 total iterations in the chain, and the first half discarded as burn-in, indicated model convergence via convergence diagnostics and visual inspection.[10] However, upon extending the length of the chain, we found that local convergence had actually been obtained. Table 3 shows an example of the mean of the slope for the first latent class. Once extending the chain out substantially to 50,000 iterations

[9] The number of iterations is not of direct concern as long as convergence has been obtained. There are many features that can result in needing a larger number of iterations such as poor starting values, complex statistical models (e.g., multilevel or mixture), and the sampler implemented in the MCMC estimation algorithm.

[10] Note that this shorter chain does appear to display patterns associated with large autocorrelation. Large degrees of autocorrelation can typically be viewed in the chain as patterns showing systematic deviations from the mean (or other central tendency measure) of the chain. It is also likely that this portion of the chain does not reflect convergence given the results and the context of the model parameter. However, the purpose of this example was to illustrate what local convergence might look like. It can be seen through the longer chain displayed in Table 3 that convergence was obtained once lengthening the chain substantially.

(first half discarded as burn-in), we can see that the chain stabilized in a different area of the parameter space. In this case, the shorter and longer chains both exhibited convergence based on convergence diagnostics. However, despite the convergence diagnostics indicating the chain was stable, it is clear that local convergence was initially obtained under 6,000 iterations. Supplementary documents for this example can be found in the online material in Folder 4.

We recommend assessing for local convergence using some additional criteria. Specifically, convergence diagnostics can be used to help establish convergence, and next relative deviation can be computed to assess potential differences after extending the length of the chain. We also discuss cases where cumulative average plots can aid in diagnosing convergence problems and when multiple chains should be used in the estimation process.

One convergence diagnostic test that can be incorporated here is the Geweke diagnostic (Geweke, 1992); note that there are others that can also be used, but Geweke can be used to specifically compare the running mean of two chains to identify potential differences (Smith, 2005). After doubling the number of iterations, the Geweke convergence diagnostic can be implemented to see how stable the full length of the chain is. The Geweke diagnostic uses a z-test for the first and last portions of a chain. If the z-test yields a significant test statistic, then the two portions of the chain significantly differ and full chain convergence was not obtained. To test local convergence, the Geweke convergence diagnostic can be used on the first half and last half of the chain. If the z-statistic rejects, then the two portions of the chain are assumed to be significantly different from one another. In this case, one can conclude that local convergence was an issue and a longer burn-in phase is likely necessary. This process should be repeated until Geweke indicates that local convergence was not an issue via a non-significant z-statistic. Implementing the Geweke convergence diagnostic is rather straightforward using the BOA (Bayesian Output Analysis) package

in R (Smith, 2005). Specifically, one would save out the CODA files (from BUGS or akin) and then read these files into the BOA package in R. A guide for implementing diagnostics in BOA can be found in the online Supplementary Material in Folder 6.

Another approach that can be taken here, which addresses a similar issue as the Geweke diagnostic, is to compute a relative deviation score between the estimates. Both of these approaches (the Geweke and the relative deviation computation) address similar inquiries (i.e., whether the estimate is stable and convergence has been obtained), but they yield slightly different interpretations. In the case of the relative deviation, some information about the magnitude of the scale for the parameter is being retained whereas this is not the case with the Geweke diagnostic. However, both approaches are meant to assess whether convergence can be assumed with the number of iterations specified in the chain.

The relative deviation can be computed between the estimates obtained during the converged result obtained for the initial model (Analysis 1) and the model where the number of iterations was doubled (Analysis 2); this relative deviation should be computed for each parameter in the model. Computing relative deviation will provide information about the fluctuations in the estimates across both chains. The researcher can then substantively interpret any fluctuations observed in the chains. The formula for computing relative deviation for each model parameter is: Relative deviation (in percent) = [(initial converged analysis – analysis with double iterations)/initial converged analysis]*100. The researcher should then use substantive knowledge about the metric of the parameter of interest, as well as substantive interpretations of the amount of fluctuation exhibited between chains, to determine when levels of relative deviation are negligible or problematic. For example, with a regression coefficient of 0.001, a 10% relative deviation level might not be substantively relevant. However, with an intercept growth parameter of 50, a 10% relative deviation

level might be quite meaningful. The specific level of relative deviation should be interpreted in the substantive context of the model. Some examples of interpretations are:

if relative deviation is < |1|%, then do not worry;

if relative deviation > |1|%, then rerun with 4-times and compare (called Analysis 3).

Compare results from Analysis 2 and Analysis 3 by computing relative deviation.

By providing these examples of interpretation, we are not trying to present a new "rule" for interpreting relative deviation. Rather, we use this as a guideline for researchers to interpret findings. Another option that researchers can use here is to look at a cumulative average plot for the mean of the posterior. This type of plot would be able to detect if the mean of the posterior was not consistent and stable throughout the post burn-in iterations.

Finally, it is also the case that a single Markov chain may not be able to expose all issues with convergence. Specifically, in a context where a distribution has multiple modes, a single chain may not be able to adequately display this information. As a result, we would recommend researchers to implement multiple chains (e.g., at least 2) for a given model parameter to explore the possibility of multiple modes existing in the posterior. The main point is to assess whether fluctuations in the chains impact the results. Thus, critical thinking about what dictates a substantive fluctuation is necessary.

What to show to your supervisor. A model where all chains have passed the visual check, the Geweke convergence diagnostic test, relative deviation levels that are considered to be substantively negligible or a stable cumulative average plot, and the first portion of Table 4 related to Point 3, which captures relative deviation and Geweke inforamtion for each model parameter. Note that

results for relative deviation and the Geweke diagnostic will likely coincide given that the Geweke diagnostic uses the mean of different fractions of the chain to assess convergence.

When to worry. If after doubling and perhaps rerunning the model with 4-times the number of iterations, results are still not comparable (e.g., if relative deviation results indicate substantial substantive fluctuations, or if the Geweke convergence diagnostic test statistic is still significant), then worry.

When to ask an expert. If problems exist after changing starting values, double-checking model specification and checking the literature to see if default priors implemented should have been altered, then consult an expert. Note that at this point, the subjective priors should not be changed, assuming that Point 1 was implemented properly, but the expert can help identify other potential issues that may be creating a problem.

Point 4: Does the histogram have enough information?

Item description. The amount of information, or smoothness, of the histogram should be checked visually for each model parameter. The purpose of this Point, as well as Point 5, is to ensure that the samples pulled from the posterior are ample enough and adequate representations of the posterior distribution. Notice that the plots for our simple example show histograms with no gaps or other abnormalities, see Table 2 column 3. This level of information is desired for histograms. Alternatively, we see a variety of plots in Figure 3, which represent histograms from the estimated chain. Looking at the histograms, rather than the smoothed densities, is important in order to assess whether there were "enough" iterations in the chain to approximate the posterior. We need to ensure that there was enough samples drawn in the chain in order to properly reconstruct the posterior. In Figure 3, we can see that plots (a) and (b) clearly do not display a smoothed and precise histogram. In both of these cases, more samples should be drawn to ensure that there is "enough"

information to fully capture various features of the posterior (e.g., central tendency and variation of the posterior). In contrast, plot (c) is showing more information, and finally plot (d) illustrates a histogram with enough information to approximate the mean and variance of the posterior. We can confidently draw substantive conclusions about the shape of plot (d).

The practical issue here is how to make the decision that "enough" information is incorporated into the posterior. Much of this decision is subjective and directly tied to the point at which the researcher feels the posterior is substantively interpretable (e.g., the mean and variance can be derived and interpreted with confidence by the researcher). With computationally complex models, this decision is likely going to be a trade-off between computational time and the amount of information gathered for the parameter estimates. If computation time is incredibly long, then we recommend running the chain until the level of information included in the histogram makes substantive sense and can be interpreted; of course, this is a subjective judgment call. We also note that the shape and smoothness of a histogram is linked to the number of bins used to create the plot. If too few bins are used, then some information may be lost and it would be more difficult to establish whether or not there is "enough" information in the chain. An objective check that can be done in this situation is to re-estimate the model with different starting values and compute the size of the effect between estimates to ensure results are stable. If the difference in estimates is substantively irrelevant between the two sets of starting values, then results are likely stable. The main point here is that enough samples have been compiled to form the posterior such that substantive conclusions can be appropriately drawn.

What to show to your supervisor. Histograms with a high level of information and column 3 of Table 2.

When to worry. Worry if a smooth histogram is not obtained for each parameter.

When to ask an expert. There is no direct need to consult an expert, but the number of iterations should be increased until smoothness in the histograms is obtained.

Point 5: Do the chains exhibit a strong degree of autocorrelation?

Item description. The very nature of a Markov chain is that the iterations in the chain are dependent on one another, and this dependency is captured by the amount of autocorrelation present in a chain. To remove (or decrease) the amount of autocorrelation in the chain, some researchers will use a process called *thinning*, where every t-th sample ($t > 1$) is selected to form the post burn-in samples in order to lessen the dependency in the posterior. It is important to stress that thinning is not a necessary component for obtaining convergence since convergence can still be obtained with dependent samples, providing a long enough chain is specified. In fact, thinning is typically not viewed as optimal because of the impact it can have on sample variance estimates for parameters (Geyer, 1991; Link & Eaton, 2011). Specifically, when a chain is thinned, sample variance estimates from the iterations must be down-weighted to account for larger lags (or higher thinning intervals) in order to produce a decent variance estimate.

Even though we do not recommend thinning in general, high degrees of autocorrelation can be indicative of other problems with the model estimation process that should be addressed. For example, 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. If convergence is also not obtained with an extreme number of iterations, then these issues can be indicative of a model specification problem. In these cases, the validity of the model results can be questionable. As a result, the cause of autocorrelation should always be investigated in order to determine if other features (e.g., the sampling algorithm or structure of the model) need modification to obtain valid results. Researchers should always examine autocorrelation plots for the model parameters. If the chains have high levels

of dependency, but convergence was obtained and the model was estimated properly otherwise, then autocorrelation can be ignored. However, if the patterns of autocorrelation suggest other estimation problems, or problems with the specification of the model, then model modification may be necessary.

What to show your supervisor. Autocorrelation plots for all parameters as seen in Table 2, column 4.

When to worry. It depends. If there is natural dependency among samples in the chain that is left unaccounted for in the model, then a longer chain is generally needed before convergence is achieved. Whenever possible, the source(s) of natural dependency should be incorporated into the model. Convergence will be obtained with a long enough chain, and the amount of autocorrelation present is not a problem for interpretation of results as long as convergence was obtained. However, if the dependency among samples in a chain seems overly excessive, or shows strange patterns when comparing across similar types of parameters in the model, then the sampling algorithm or the specification of the model may need to be modified.

When to ask an expert. If the autocorrelation plots suggest there may be a problem with the sampling algorithm (e.g., some parameters are showing rather excessive autocorrelation, thus requiring much longer mixing times), then an expert can be consulted to help determine whether an alternative sampling algorithm might be necessary.

Point 6: Does the posterior distribution make substantive sense?

Item description. Substantive abnormalities in the posterior distribution should be examined (e.g., through Kernel density plots). The main things that should be checked in a posterior distribution are that it: is smooth, makes substantive sense, does not have a posterior standard deviation that is greater than the scale of the original parameter, does not have a range of the posterior credible

interval greater than the underlying scale of the original parameter, and does not show great fluctuations in the variance of the posterior.

What to show to your supervisor. Posterior distributions that are smooth and make substantive sense, and column 5 of Table 2 should also be filled out.

When to worry. Worry if the posterior does not make sense substantively.

When to ask an expert. If the results show convergence in Points 2-6 but the posterior does not make sense substantively, talk to your supervisor and go into the literature to find out if there is an alternative substantive justification for these findings. If these recommendations fail, then see an expert.

Stage 3: Understanding the Exact Influence of the Priors

For cases where only non-informative or default-setting priors are used, Points 7-9 can be skipped. However, if (weakly) informative priors were implemented for any model parameters, then points 7-9 should be addressed.

Warning: It is imperative that decisions made during Points 7-9 are presented in a completely transparent manner. If Points 7-9 indicate that results from the sensitivity analysis are problematic (e.g., some parameters are extremely sensitive to the prior specification), then any changes made to the model or priors should be presented in a completely clear way and the Checklist should be started over with the model that incorporates any changes made.

Although this next section recommends playing around with some of the prior settings, it is important to note that this is an exercise to improve the understanding of the prior specified in Point 1 and <u>not</u> a method for changing the original prior and continuing forward. There are several

dangers of adjusting priors at this stage of the process. This issue is related to questionable research practices. For example, if a researcher changes the prior after seeing results of Points 7-9, then this can be considered as manipulating the results. Further, priors can be altered to influence results in whatever way the researcher wants. Finally, if the original priors are updated after seeing the results of Points 7-9 and the new prior is implemented with the same data, then it is the considered double-use of the data. All three of these examples are highly discouraged and may even be considered violating the moral integrity of science. Specific to Bayesian work, openness and transparency in the selection of priors is imperative for this reason.

If Points 7-9 indicate instability of results through a sensitivity analysis (e.g., a parameter is particularly sensitive to prior settings), then it is possible that the model was mis-specified or the parameters are not fully identified by the data or model. In this case, researchers should consider making the necessary changes to the model to combat any identification or mis-specificaiton issues. However, once any changes have been made, the process of implementing the Checklist should start over from the beginning. The following points are designed to help researchers better understand Bayesian results and understand the impact of the priors selected in Point 1 above.

Point 7: Do different specifications of the multivariate variance priors influence the results?

The information provided for Point 7 is decidedly more technical than the other points presented here. Handling a multivariate variance prior has technical complexities that are often not elaborated upon in applied Bayesian papers, but some severe issues can arise if this prior is not specified properly. Although it is true that univariate priors used for variances (or standard deviations/precisions) have some similar complexities (see e.g., van de Schoot et al., 2015), we highlight some of the issues specific to the multivariate treatment of these priors. In our experience, the multivariate priors used in this situation can be quite difficult to navigate and require detailed

consideration during implementation. It is our aim in this section to describe some of these complexities in order to introduce researchers to this multivariate prior, describe some of the problems that can arise, and provide guidance for handling this type of prior.

Background information. Just as with Point 1 and the univariate model priors, it is also important to understand your multivariate prior for a covariance matrix. A multivariate prior such as this would be placed on the matrix of variances and covariances. In cases where data are distributed multivariate normal (MVN), the data distribution can be written as MVN($[\mu_{Y1}, \mu_{Y2}]$, Σ), where Σ (the covariance matrix) is commonly specified as following an inverse Wishart distribution (IW). The IW distribution is perhaps the most common prior specification for covariance matrices. There are two hyperparameters for the IW distribution such that $\Sigma \sim IW(\Omega, d)$, where Ω is a positive definite scale matrix and d is an integer representing the degrees of freedom for the IW distribution. The integer d can vary depending on the informativeness of the prior distribution.

Overall, the Wishart family of multivariate priors is important to handle properly. There have been many comments published on the optimal specification of the (inverse) Wishart prior. Specifically, by O'Malley and Zaslavsky (2005) and subsequently Gelman and Hill (2007) have recommended using a scaled inverse-Wishart prior, where the covariance matrix is broken up into a diagonal matrix of scale parameters and an un-scaled covariance matrix which is then given the prior (for additional details see Gelman, 2006, September 1). The exact specification of the Wishart prior has also been found to have a large impact when variances (diagonal elements) in the covariance matrix are small (Schuurman, Grasman, & Hamaker, in press). It is especially important to assess hyperparameters for the multivariate prior through a sensitivity analysis in order to examine the potential impact of the prior. For example, as a preliminary analysis Depaoli (2012) conducted a sensitivity analysis for the IW prior in the context of a mixture confirmatory factor analysis model

and found that even slightly modifying the hyperparameters of the IW changed final estimates substantially.

Given the complexity of the (inverse) Wishart prior, our advice is twofold. First, the default settings for the prior can be specified to ensure that the prior is properly specified according to a multivariate probability distribution. Second, if the default settings are changed for the (inverse) Wishart prior, then we recommend consulting an expert to ensure that the prior is positive definite and that it will perform properly during estimation. Note that the practitioner may still choose to consult with an expert, even if the default settings are used. The complexity of this prior often warrants careful consideration.

Item description. The action to take in order to examine the impact of this multivariate prior is to always assess an alternative setting for the prior and compare structural and measurement model results to the original default setting results obtained in previous stages. There are three specifications of the inverse Wishart (IW) that are discussed as non-informative in Asparouhov and Muthén (2010, page 35). The first specification is IW(0, -p-1) for covariance matrices, where p is the dimension of the covariance matrix, and mimics a uniform prior bounded at (-∞,∞). The second specification is IW(0,0). The last specification discussed is IW(**I**, p+1), where this prior mimics the case where off-diagonal elements (covariances) of the covariance matrix would have uniform priors bounded at [-1,1] and diagonal elements (variances or residual variances) distributed as IG(1,.5), where IG represents the inverse gamma distribution. Once a second specification of the IW is implemented, then the effect of the prior should be computed between the two IW specifications.[11] In order to compute this effect, one can use the following formula for the parameter estimates:

[11] Here we use the terminology of "effect of the prior" rather than the term "bias", which is commonly used in comparable frequentist settings. Within the Bayesian framework, "bias" does not take on the same meaning since bias that is directly related to priors will diminish and disappear under large sample sizes given that the impact of the prior will also diminish (Gelman et al., 1996). As a result, we refer to differences between prior settings as "effects".

Effect of the prior (in terms of % difference) = [(initial prior specification – subsequent prior specification)/initial prior specification]*100. We present an example in Table 4.[12]

What to show to your supervisor. The size of the effect for all model parameters between the first and second specifications of the IW prior should be provided see Table 4, column 2 (section ii).

When to worry. When the size of the effect between the two IW specifications are substantively meaningful (e.g., $> |1|$% for any measurement or structural model parameter), then worry about the impact of the IW prior.

When to ask an expert. If the technical details related to this section are confusing, or anytime the size of the effects are substantively meaningful (e.g., $>|1|$%).

Point 8: Is there a notable effect of the prior when compared with non-informative priors?

Item description. In order to understand the impact that the subjective (i.e., weakly-informative or informative) prior is having on model results, we recommend that comparisons be made between subjective and non-informative priors. Specifically, this Point involves estimating the model with all non-informative priors and comparing results via the size of the effect of the subjective prior as defined in Point 1. This Point is simply to understand the subjectivity of the prior and can aid in the discussion of the prior impact. We note here that priors will impact different parameters in different ways. For example, mean structure parameters are less sensitive to priors, but parameters in more complex models may be more sensitive (e.g., latent variable models, mixture class models, or multilevel models).

[12] We have also included a column in Table 4 corresponding to the PSRF convergence diagnostic. We expand on using this statistic in the next section.

If the size of the effect is relatively small (e.g., less than |1|%) *and* the substantive conclusion remains the same, then the subjectivity embedded into the prior has no impact.[13] The researcher should continue to use informative priors but recognize their limited impact in the discussion. If the size of the effect is moderate (e.g., |1-10|%) *and* the substantive results are the same, then the subjectivity of the prior may have had a moderate impact on final results. However, if the size of the effect is moderate (e.g., |1-10|%) and the substantive results differ, then the subjectivity of the prior had a large impact. If the size of the effect is large (e.g., greater than |10|%) or substantive interpretations are different, then the subjectivity of the prior had a large impact on results. Although we have supplied some example cutoffs here, we recognize that relative deviation levels, or sizes of effects, are largely interpreted in the substantive context of the metric of the parameters being examined. Therefore, we are not attempting to create new rules of thumb that span across all research scenarios. Rather, this section should be meant as a guide for interpreting one's own results.

Another method that can be used to examine the similarity of results obtained across models with two sets of priors specified is to use the Gelman and Rubin (1992a;b) convergence diagnostic. This diagnostic produces something called a potential scale reduction factor (PSRF). If this factor is near 1.0, within some preset bound, then two chains are said to have converged. Typically this method is used to examine convergence between two chains in the same model. However, we propose a novel use of the PSRF where two equal-length chains from separate analyses (e.g., one model with two sets of priors) are compared. If the PSRF is quite large, then this is another

[13] The size of the effect for a parameter using a percent relative deviation computation can be computed using the following formula: [(model with initial subjective parameter) – (model with non-informative prior)/(model with initial subjective parameter)]*100. The estimates for each model would be embedded into this equation and the size of the effect is produced. The relative size of this effect should be interpreted in the context of what is substantively meaningful (i.e., 5% might represent a large effect in one context and a very small effect in another context). In the case of our reference to specific cut-off values (e.g., 1%), our intention was to present an *example* of a negligible difference between effects. However, results should always be interpreted in the context of the particular model and parameters being investigation.

indication that the results obtained from the different prior settings may be substantively different. This finding could point toward the prior having a meaningful impact on results, which should be thoroughly described in the discussion section of a paper. For an example of implementing the PSRF in this novel manner, please refer to online Supplementary Material in Folder 6. Next, we provide an example of interpreting such findings.

As an example, we pulled reading achievement data from the Early Childhood Longitudinal Study-Kindergarten (ECLSK) database (NCES, 2001) for children throughout kindergarten and first grade (2 timepoints of data collected in each grade). We have estimated a latent growth curve model with the 3,856 children akin to the model presented in Kaplan (2002). To illustrate this point, we used the intercept estimate from Kaplan's model (Kaplan, 2002, page 204) as the mean hyperparameter for the prior we specified for the intercept of the growth model. Further, we arbitrarily specified a variance hyperparameter of 1, thus giving us a subjective prior for the latent intercept mean of N(31.37, 1). Results from an analysis with 10,000 burn-in and 10,000 post burn-in iterations using the OpenBUGS software are presented in Table 5. We then estimated a second model using a diffuse prior for the intercept mean ($N(0,10^6)$ to see how much of an impact our subjective prior has on results. We can see in Table 5 that the percent of relative deviation is quite low (under 1%); the corresponding PSRF value is 1.477, which may be considered high given that PSRF values are often required to be smaller than 1.1 or 1.2 when assessing convergence. However, this use of the PSRF is novel and we might expect more drastic changes in the model since more than just starting values have been modified across chains.[14] Although the PSRF is a bit high, we can interpret results as being relatively stable when the subjective prior is compared to a diffuse prior. In

[14] The PSRF is typically used to assess chains when the only difference is in the starting values specified for the chain. This use that we propose is completely novel (to our knowledge), and further research should be conducted on the use of the PSRF in this context in order to provide additional guidelines for cut-off values of the diagnostic when comparing across types of priors, etc.

this case, we can conclude that our theory (incorporated into the subjective prior) had little impact on the results. Supplementary documents for this example can be found in the online material in Folder 5.

For Bayesian updating, it is essential that researchers report the priors used, even if there is no substantive impact on results. In the case where there is a large difference between results, do not despair. These findings are interesting and fun. The entire focus of the discussion section can turn toward the discrepancy between results obtained using informative versus non-informative priors. This discussion illustrates the mismatch between theory and data, and it is up to the researcher to come up with an explanation.

What to show to your supervisor. The size of the effect between informative (or weakly-informative) priors and non-informative priors, see Table 4 section iii.

When to worry. Do not worry at all. Either results match and the subjectivity of the prior does not have an impact, or results differ and that becomes an interesting talking point for the discussion.

When to ask an expert. Never. All findings are fun and if discrepancies do not make sense at first, then turn to the literature for explanations.

Point 9: Are the results stable from a sensitivity analysis?

Item description. If informative or weakly-informative priors are used, then we suggest running a sensitivity analysis of these priors. When subjective priors are in place, then there might be a discrepancy between results using different subjective prior settings. A sensitivity analysis for priors would entail adjusting the entire prior distribution (i.e., using a completely different prior distribution than before) or adjusting hyperparameters upward and downward and re-estimating the model with these varied priors. Several different hyperparameter specifications can be made in a

sensitivity analysis, and results obtained will point toward the impact of small fluctuations in hyperparameter values.

Take the same ECLSK reading achievement example where we estimate a simple latent growth curve model. If we assume a normal distribution for the intercept prior, then we will need to specify values for the mean and variance hyperparameters of the subjective prior. If we have the intercept mean hyperparameter specified at 31.37 (based on a previous analysis presented in Kaplan, 2002), then we can start the sensitivity analysis at this point by varying this value upward and downward to see how much of an impact the mean hyperparameter of this prior has on the final estimate for the intercept. Specifically, we can examine a series of priors with mean hyperparameters specified in 5-point increments from this initial value of 31.37. Specifically, we can test mean hyperparameters of : 21.37, 26.37, 31.37 (the value in our original prior), 36.37, and 41.37. In this case, there would be five different models estimated and we can compare through the Gelman and Rubin convergence diagnostic whether the chains resulting from the respective priors are comparable. We can also compute the size of the effect to assess how different the results are when the prior mean is specified at these levels.

The purpose of this sensitivity analysis is to assess how much of an impact the location of the mean hyperparameter for the prior has on the posterior. For the ECLSK reading data, we have reported results from the sensitivity analysis in Table 5. Note the column labeled "PSRF" indicates how comparable the new priors specified through the sensitivity analysis are to the original prior with a mean hyperparameter of 31.37. We can see that PSRF values all indicate non-convergence with values beyond $1.0 \pm .05$; notice greater evidence of non-convergence as the mean hyperparameter becomes more extreme (i.e., 21.37 and 41.37 have comparatively larger PSRF values). However, percent of relative deviation is quite low (under 1% for all comparisons) so we

can interpret results as being relatively stable with the use of different mean hyperparameters. We would then continue our sensitivity analysis investigation for the intercept to examine the variance hyperparameter (see bottom of Table 5). In this case, we tested the prior of N(41.37,0.1) to see the impact of varying both hyperparameters. This is perhaps an extreme example of a sensitivity analysis cell, but it illustrates the type of results one might get if the prior has a substantial impact on estimates. Specifically, we see in Table 5 that the PSRF value is quite high (over 20), indicating non-convergence between the chains from the two priors. We can also see a very large effect between the estimates (over 16%), which further indicates this prior specification had a substantial impact on the results. In this case, the researcher would describe the substantive differences in the discussion section.

As another illustration, Figure 4 exhibits how changes in substantive results can be tracked throughout a sensitivity analysis. In this case, we have one parameter with a normally distributed prior, where we varied the mean and variance hyperparameters. Each of the lines in the plot represent a different mean hyperparameter. The *y*-axis represents the posterior parameter value obtained for the estimate, and the *x*-axis represents the variance hyperparameter for the prior. In this example, significance for the parameter of interest is identified with a solid line, and non-significance of this parameter is denoted with a dashed line. These results show that with a large variance hyperparameter of 1000, the value of the mean hyperparameter makes no difference on the final parameter estimate. As the variance hyperparameter decreases, the mean hyperparameter has more influence on the final estimate obtained. A plot like this can be very helpful in examining how significance patterns and substantive interpretations change as the prior is modified. Supplementary documents for this example using ECLSK data can be found in the online material in Folder 5. For more examples and details on sensitivity analysis of priors see: Hamra, MacLehose, and Cole (2013);

Heydari, Miranda-Moreno, Lord, and Fu (2014); Lopes and Tobias (2011); Millar (2004); and van de Schoot et al., (2015).

Upon receiving results from the sensitivity analysis, assess the impact that fluctuations in the hyperparameter values have on the substantive conclusions. Results may be stable across the sensitivity analysis, or they may be highly instable based on substantive conclusions. Whatever the finding, this information is important to report in the results and discussion sections of a paper. We should also reiterate here that original priors should not be modified, despite the results obtained.

What to show to your supervisor. A report showing the (in)stability of the results for the entire sensitivity analysis; a table akin to Table 5 or a plot akin to Figure 4 will be particularly useful.

When to worry. Do not worry necessarily, but note if there is a great deal of instability in substantive conclusions, even with small fluctuations in the hyperparameter values. Even if there are only minor substantive differences, this is an important factor to discuss in the paper. If there are some parameters that are particularly sensitive to changes in the prior through the sensitivity analysis, then this could be an indication that the model was mis-specified or there are identification issues regarding certain parameters in the model. If this is the case, then the model should be re-specified and the Checklist should be repeated with the new model starting with Point 1. In this case, it is important to be 100% transparent about any changes that were made in the model or priors as a result of the original sensitivity analysis findings. However, if after carefully checking the priors and the model the results are still sensitive to different prior specifications, then this may just be the result of the study. In this case, the sensitivity of results to the prior settings should be thoroughly described in the discussion section.

When to ask an expert. If there is a great deal of instability in substantive conclusions, even with small fluctuations in the hyperparameter values. This high instability could be a symptom of a larger problem with the model or the priors.

Stage 4: After the Interpretation of the Model Results

Point 10: Is the Bayesian way of interpreting and reporting model results used?

Item description. After Points 1-9 have been carefully considered and addressed, model interpretation and the reporting of results become the next concern. First, we will consider the issue of properly interpreting Bayesian results, which is then followed by a discussion on reporting Bayesian results.

There are some important distinctions between the interpretation of frequentist and Bayesian statistics. One of the most notable distinctions is that the Bayesian framework no longer deals in terms of point estimates compared to frequentist approaches. Results obtained under the Bayesian framework reflect the posterior distribution obtained, where each parameter is estimated with a density capturing uncertainty in the true value. It is common for researchers to summarize the posterior density with the mean, median, or mode of the distribution. This summary should be interpreted as the central tendency measure for the posterior distribution, rather than as a point estimate. In order to capture the spread (and potentially the shape) of the posterior, Bayesian credible intervals are constructed. The Bayesian credible interval is akin to the frequentist confidence interval, but the interpretations rely on different probability theories thus making interpretations different across the two frameworks. For example, a 95% frequentist confidence interval of [0.05, 1.12] for a regression coefficient would indicate that over long-run frequencies, 95% of the confidence intervals constructed in this manner (e.g., with the same sample size, etc.) would contain the true population value. In contrast, the 95% Bayesian credible interval of [0.05, 1.12] would be

interpreted such that there is a .95 probability of the population regression coefficient falling between 0.05 and 1.12, indicating that this regression coefficient likely represents a positive effect.

As it has hopefully been made clear at this point, Bayesian analyses have many distinct features that are not a part of traditional frequentist methods. As a result, there are different considerations when writing up results for a Bayesian model to ensure that all information has been properly conveyed and that results can be replicated or priors can be extracted and then updated in future Bayesian models. There are several key components that must be included in a write-up of Bayesian results and these include information from all of the previous Points 1-9 listed above, see for an example Matzke, Dolan, Logan, Brown, and Wagenmakers (2013).

What to include in the write-up. When writing an empirical Bayesian manuscript, one could use the following list to ensure completeness (see also, van de Schoot and Depaoli, 2014). If one of the following points is not adequately addressed in the text, then we feel the paper should not be published in its current form. Particular to Bayesian work, openness and transparency is imperative. Specifically, it is imperative that in the analytic strategy section a paragraph should be devoted to providing information about how priors were obtained and why each prior was specified in that way. Also the hyperparameters should be reported in a table or in an online supplementary file. Next, information about estimation and convergence must be detailed. The methods section should reveal the program used for estimation, the sampler (e.g., Gibbs) implemented, the number of chains, the number of burn-in iterations, seed and starting values for the chains, the number of post burn-in iterations, and how convergence was checked or monitored (e.g., visual inspection and convergence criteria such as the Gelman and Rubin diagnostic). All of the points addressed earlier for identifying and checking convergence should also be detailed so that the reader understands the extent to which chain convergence was checked. One could refer to our WAMBS-checklist to ensure convergence

has been established. Next, the impact of the priors should be carefully described. If (weakly) informative priors were used, then the substantive differences to non-informative priors must be compared to understand the impact of the prior. Likewise, results of the sensitivity analysis must also be described in the text as a means to further describe the impact of the prior on the final model results. Again, the WAMBS-checklist could be used for this investigation.

What to show to your supervisor. The supervisor should be given a full write-up of the results, as well as any relevant information appearing in the discussion section. We have provided a hypothetical example of a Bayesian results section in Appendix A.

When to worry. Worry if you cannot understand or convey the differences between conventional and Bayesian model results, or if you are not able to create results and discussion sections that reflect all of the information constructed in Points 1-9 described above.

When to ask an expert. Never. There is likely no need to consult an expert for interpretation. Instead, the researcher can consult the many reference books and papers we have listed in the introduction to aid in Bayesian statistics interpretation. However, if you are unable to write up a section of the Bayesian findings after consulting examples and other readings on interpretation and reporting Bayesian results, then consult an expert.

Conclusion

It is our aim to highlight some of the most important nuances of implementing Bayesian methods and to provide the succinct, but comprehensive, *When to worry and how to Avoid the Misuse of Bayesian Statistics* (WAMBS) checklist to aid in avoiding the misuse of Bayes. If the 10 points in this checklist are carefully considered and addressed, then many of the common problems or mistakes that arise in Bayesian estimation can be avoided or corrected.

There are some limitations and warnings surrounding the use of the WAMBS checklist that should be highlighted. First, we note that this checklist may be rather tedious and time-consuming to implement in cases where models have many parameters being estimated. For example, item response theory models can become cumbersome with sometimes thousands of person parameters being estimated under large sample size cases. We acknowledge from experience that implementing such a checklist under cases where there are many model parameters is difficult. However, we also feel a thorough check is imperative, regardless of the number of model parameters being estimated. If, for example, a model with 5000 model parameters is estimated but a few parameters do not reach stable convergence, then the model results would not be appropriate to report. It is important to check all model parameters, however tedious that may be. To assist in implementing such massive parameter checks, we recommend the use of software (e.g., the MplusAutomation package implemented in R) that can aid in handling a large number of model parameters. We feel the added complexity of using a checklist is a heavy price to pay when implementing Bayesian statistics because it does require a good deal of work. However, we also feel using the Checklist is a necessary price to pay to ensure that results are trustworthy, the model estimation process is transparent and can be replicated, and that estimation and reporting of results exhibit best practice.

Second, we did not directly deal with issues tied to the assessment of model fit/selection within the Bayesian framework. Properly assessing model fit and model selection are important issues to handle alongside the implementation of this checklist. Data-driven model selection techniques (e.g., comparing deviance information criteria across competing models) are typically considered after the estimation of model parameters. Therefore, it is likely that this Checklist would be implemented simultaneously with the model selection process. Although we do not directly address Bayesian model fit and selection here (these topics would likely warrant their own checklists, in fact), we recognize that these issues would likely be handled in parallel to the issues addressed

through the WAMBS-checklist. At a very minimum level, researchers should consult an expert about issues surrounding model fit and selection.

Finally, there may be times when specific points are not easily satisfied when implementing the WAMBS-checklist. For example, it is possible that convergence may not be obtained even after doubling the number of iterations. In cases where specific points in the checklist are not satisfied, we have provided some additional guidelines to act as a starting point for troubleshooting and continuing to strive to satisfy each point. These guidelines are presented in Table 6 and should act as "next steps" in thinking if some of the Points are not fulfilled using the Checklist guidelines presented here.

To conclude, this Checklist should act as a guide for implementation and for writing up findings. We stress that openness and transparency are vital for implementing any statistical tool, but this is especially the case for Bayesian tools. One of the main goals for the WAMBS-checklist is to aid in improving replicability of results in Bayesian statistics. There are so many points within the process of implementing Bayesian methods where things can go awry (e.g., misinterpretation, problems with non-convergence, unintended impact of priors). It is our aim to promote clarity during implementation and dissemination of Bayesian modeling, and we hope that the WAMBS-checklist assists with this goal. Finally, we hope that researchers working within the Bayesian framework realize the advantages of estimation and interpretation, and we hope they have fun interpreting any (mis)match between data and theory!

References

Albert, J. (2009). *Bayesian computation with R*. New York: Springer.

Arbuckle, J. L. (2006). *Amos (Version 7.0) [Computer Program]*. Chicago: SPSS.

Asparouhov, T. & Muthén, B. (2010). *Bayesian analysis using Mplus: Technical implementation*. Technical Report. Version 3.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 3*, 385–402.

Bijak, J., & Wisniowski, A. (2010). Bayesian forecasting of immigration to selected European countries by using expert knowledge. *Journal of the Royal Statistical Society, 173*, 775–796.

Bolstad, W. M. (2007). *Introduction to Bayesian statistics*. John Wiley & Sons.

Brooks, S. P., & Roberts, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, *8*(4), 319-335.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics, 2*, 113–152.

Browne, W.J. (2009) *MCMC Estimation in MLwiN v2.1*. Centre for Multilevel Modelling, University of Bristol.

Candel, J. J. M., & Winkens, B. (2003). Performance of empirical Bayes estimators of level-2 random parameters in multilevel analysis: A Monte Carlo study for longitudinal designs. *Journal of Educational and Behavioral Statistics, 28*, 169–194.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis*. Boca Raton, FL: Chapman and Hall/CRC Press.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*, 167-174.

Christensen, R., Johnson, W.O., Branscum, A.J., & Hanson, T.E. (2010). *Bayesian ideas and data analysis: An introduction for scientists and statisticians*. Boca Raton, FL: CRC Press.

Darnieder, W. F. (2011). *Bayesian methods for data-dependent priors*. (Dissertation, Ohio State University).

Depaoli, S. (2012). Measurement and structural model class separation in mixture-CFA: ML/EM versus MCMC. *Structural Equation Modeling, 19*, 178-203.

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods, 18*, 186-219.

Depaoli, S. (2014). The Impact of Inaccurate "Informative" Priors for Growth Parameters in Bayesian Growth Mixture Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 239-252.

Depaoli, S., and Boyajian, J. (2014). Linear and nonlinear growth models: Describing a Bayesian perspective. *Journal of Consulting and Clinical Psychology, 82,* 784-802.

Depaoli, S., and Clifton, J. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling, 22,* 327-351.

Depaoli, S., and Scott, S. (2015). Frequentist and Bayesian estimation of CFA measurement models with mixed item response types: A Monte Carlo investigation. *Structural Equation Modeling.* Online first publication.

Depaoli, S., van de Schoot, R., van Loey, N., and Sijbrandij, M. (2015). Using Bayesian statistics for modeling PTSD through latent growth mixture modeling: Implementation and discussion. *European Journal of Psychotraumatology, 6*, 27516.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274-290.

Fransman, W., Van Tongeren, M., Cherrie. J.W., Tischer M., Schneider, T., et al. (2011). Advanced Reach Tool (ART): Development of the Mechanistic Model. *Ann Occup Hyg, 55*, 957-979.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*, 515–533.

Gelman, A. (2006, September 1). Modeling the group-level covariance matrix for varying-intercept, varying-slope multilevel models: Updated paper by O'Malley and Zaslavsky. Retrieved August 19, 2015, from http://andrewgelman.com/2006/09/01/modeling_the_gr/

Gelman, A., Bois, F., & Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association, 91*, 1400-1412.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London, UK: Chapman & Hall.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 4*, 1360–1383.

Gelman, A., & Rubin, D. B. (1992a). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-511.

Gelman, A., & Rubin, D. B. (1992b). A single series from the Gibbs sampler provides a false sense of security. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), Bayesian statistics 4 (pp. 625-631). Oxford: Oxford University Press.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 8-38.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), 721-741.

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

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood (Tech. Rep.). Computing Science and Statistics: Proceedings 23rd Symposium on the Interface.

Gill, J. (2008). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: CRC Press.

Hamra, G.B., MacLehose, R.F., Cole, S.R. (2013). Sensitivity analyses for sparse-data problems - Using weakly informative Bayesian priors. *Epidemiology 24*, 233-239.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*(6), 1109-1144.

Heydari, S., Miranda-Moreno, L. F., Lord, D., & Fu, L. (2014). Bayesian methodology to estimate and update safety performance functions under limited data conditions: A sensitivity analysis. *Accident Analysis and Prevention, 64*, 41-51

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, *5*(3), 315.

Hox, J., van de Schoot. R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6*, 87-93.

Ibrahim, J. G., Chen, M. H., & Sinha, D. (2005). *Bayesian survival analysis*. John Wiley & Sons, Ltd.

Jackman, S. (2009). *Bayesian analysis for the social sciences* (Vol. 846). John Wiley & Sons.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford Classic Texts in the Physical Sciences. Oxford Univ. Press, Oxford.

Johnson, V. E. (2013). Uniformly most powerful Bayesian tests. *The Annals of Statistics, 41,* 1716-1741.

Kaplan, D. (2014). *Bayesian statistics for the social sciences.* New York: Guilford.

Kaplan, D. (2002). Methodological advances in the analysis of individual growth with relevance to education policy. *Peabody Journal of Education*, *77*(4), 189-215.

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. Hoyle (ed.), *Handbook of structural equation modeling* (pp 650-673). New York: Guilford.

Kaplan, D. & Depaoli, S. (2013). Bayesian statistical methods. In T. D. Little (ed.), *Oxford Handbook of Quantitative Methods*. (pp 407- 437). Oxford: Oxford University Press.

Kim, J. S., & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*(4), 38-51.

Kim, SY., Suh, Y., Kim, JS., Albanese, M., & Langer M.M. (2013). Single and Multiple Ability Estimation in the SEM Framework: A Non-Informative Bayesian Estimation Approach. *Multivariate and Behavioral Research, 48,* 563-591.

Kruschke, J.K. (2010). *Doing Bayesian analysis*. Burlingon, MA: Academic Press.

Kruschke, J.K. (2011). Introduction to Special Section on Bayesian Data Analysis. *Perspectives on Psychological Science, 6*, 272-273

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in medicine*, *24*(15), 2401-2428.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. West Sussex, UK: John Wiley and Sons.

Link, W. A., and Eaton, M. J. (2011). On thinning of chains in MCMC. *Methods in Ecology and Evolution, 3,* 112-115.

Lopes, H. F., & Tobias, J. L. (2011). Confronting prior convictions: On issues of prior sensitivity and likelihood robustness in Bayesian analysis. *The Annual Review of Economics, 3*, 107-131

Love, J., Selker, R., Marsman, M., Jamil, T., Verhagen, A. J., Ly, A., Gronau, Q. F., Smira, M., Epskamp, S., Matzke, D., Wild, A., Rouder, J. N., Morey, R. D. & Wagenmakers, E.-J. (2015). JASP (Version 0.6.6)[Computer software].

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325-337.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists* (Vol. 2). New York: Springer.

Martin, T. G., Burgman, M. A., Fidler, F., Kuhnert, P. M., Low-Choy, S., Mcbride, M. And Mengersen, K. (2012), Eliciting Expert Knowledge in Conservation Science. *Conservation Biology, 26*, 29–38.

Matzke, D., Dolan, C. V., Logan, G. D., Brown, S. D., & Wagenmakers, E. J. (2013). Bayesian parametric estimation of stop-signal reaction time distributions. *Journal of Experimental Psychology: General*, *142*(4), 1047.

Millar, R. B. (2004). Sensitivity of Bayes estimators to hyper-parameters with an application to maximum yield from Fisheries. *Biometrics, 60*(2), 536-542

Moore, T. M., Reise, S. P., Depaoli, S., & Haviland, M. G. (2015). Iteration of partially specified target matrices: Applications in exploratory and Bayesian confirmatory factor analysis. *Multivariate Behavioral Research*, *50*, 149-161.

Morris, D.E., Oakley, J.E., & Crowe, J.A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software, 52*, 1–4

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran 90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software, 46(2), 1-39*.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods*, *17*, 313-335.

Muthén, L., & Muthén, B. (1998-2015). *Mplus user's guide* (Seventh ed.). Los Angeles, CA: Muthén & Muthén.

NCES. (2001). Early childhood longitudinal study: Kindergarten class of 1998-99: Base year public-use data files user's manual (Tech. Rep. No. NCES 2001-029). U.S. Government Printing Office.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Wiley Series in Computational Statistics, Hoboken, NJ.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. West Sussex: Wiley

O'Malley, A. J., & Zaslavsky, A. M. (2005). Cluster-level covariance analysis for survey data with structured nonresponse. Technical report. Department of Health Care Policy, Harvard Medical School, Cambridge, MA.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. http: //www-fis.iarc.fr/~martyn/software/jags/.

Press, S. J. (2003). *Subjective and objective Bayesian statistics: Principles, models, and applications* (2nd ed.). New York, NY: Wiley

Raftery, A. E., & Lewis, S. (1992). How many iterations in the Gibbs sampler. *Bayesian statistics*, *4*(2), 763-773.

Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 163–187). New York: Chapman & Hall.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, 59*, 731–792.

Rietbergen, C., Klugkist, I., Janssen, K. J., Moons, K. G., & Hoijtink, H. J. (2011). Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemporary Clinical Trials*, *32*(6), 848-855.

SAS Institute Inc., *SAS 9 Help and Documentation*, Cary, NC: SAS Institute Inc., 2002-2013.

Schuurman, N. K., Grasman, R. P.P.P, & Hamaker, E. L. (in press). A comparison of Wishart prior specifications for variance-covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*.

Seaman III, J. W., Seaman Jr, J. W., & Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician*, *66*(2), 77-84.

Sinharay, S. (2004). Experiences with Markov Chain Monte Carlo Convergence Assessment in Two Psychometric Examples. *Journal of Educational and Behavioral Statistics, 29*, 461-488.

Skrondal, A., & Rabe-Hesketh, S. (2012). *Multilevel and longitudinal modeling using Stata*. STATA press.

Smith, B. J. (2005, March 23). Bayesian Output Analysis program (BOA), version 1.1.5. http://www.public-health.uiowa.edu/boa.

Stan Development Team (2014). *STAN Modeling Language Users Guide and Reference Manual, Version 2.2*.

StataCorp. 2013. *Stata 13 Base Reference Manual*. College Station, TX: Stata Press.

van de Schoot, R., Broere, J., Perryck, K., Zondervan-Zwijnenburg, M., & van Loey, N. (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*. doi: http://dx.doi.org/10.3402/ejpt.v6.25216

van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *The European Health Psychologist, 16*, 73-82.

van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M. A., Orobio de Castro, B., Meeus, W., & Romeijn, J. W. (2011). Evaluating expectations about negative emotional states of aggressive boys using Bayesian model selection. *Developmental Psychology*, *47*(1), 203.

van de Schoot, R., Ryan, O., Winter, S., Zondervan-Zwijenburg, M. A. J., & Depaoli, S. (under review). A systematic review of empirical Bayesian applications in psychology.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavorial Statistics, 33*, 5–20.

Van Loey, N. E. E., Maas, C. J. M., Faber, A. W., & Taal, L. A. (2003). Predictors of chronic post-traumatic stress symptoms following burn injury: Results of a longitudinal study. *Journal of Traumatic Stress, 16*(4), 361-369.

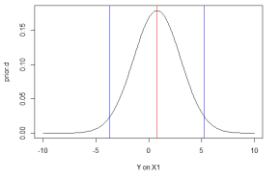
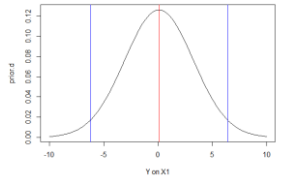
Van Wesel, F., Hoijtink, H., & Klugkist, I. (2011). Choosing priors for constrained analysis of variance: methods based on training data. *Scandinavian Journal of Statistics*, *38*(4), 666-690.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62*, 159–180.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology an empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291-298.

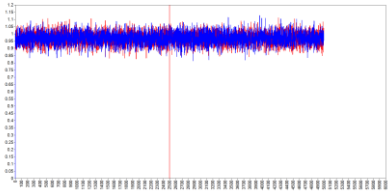
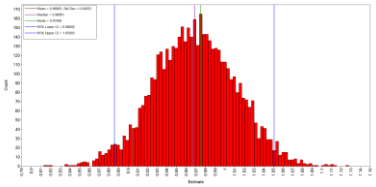
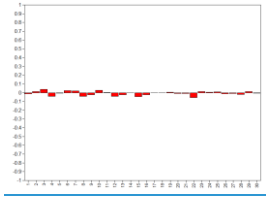
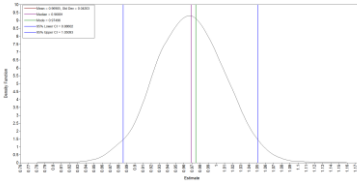
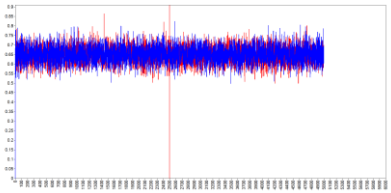
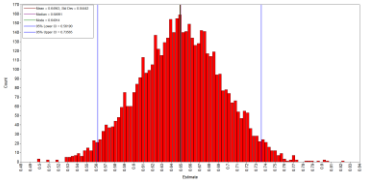
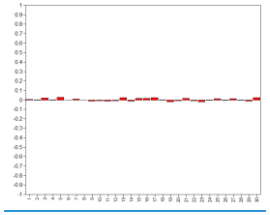
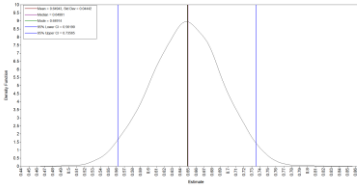
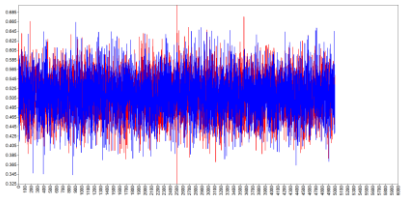
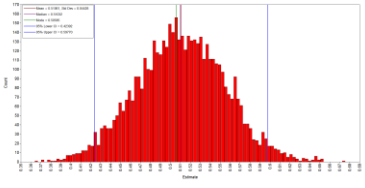
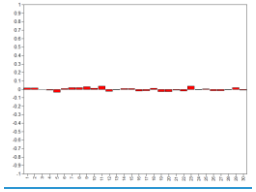
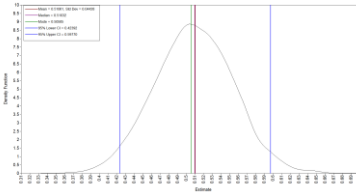
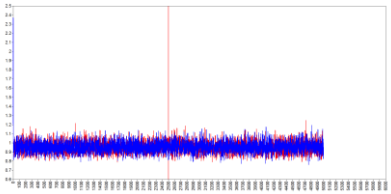
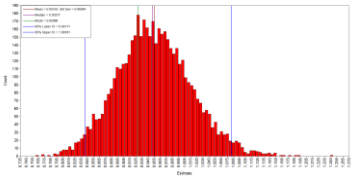
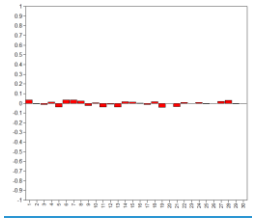
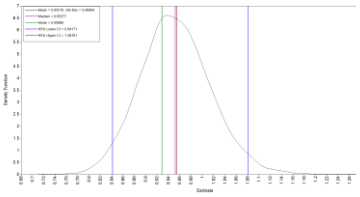
Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., & Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development, 31*(4), 374-383.

Table 1. Table to show your supervisor for Point 1: Do you understand the priors? Consider a basic regression analysis with 1 dependent variable (Y) and two predictors (X_1 and X_2).

Parameters	**Distributional form of the priors** (e.g., normal, inverse gamma, etc)	**Type of prior** (non-, weakly, highly informative)	**Source of background information**	**Picture of Plot**	**Hyperparameters**
Y on X_1	Normal	Highly Informative	Table x on page xx of the meta-analysis of Author et al. (2000)		N(.8,5);
Y on X_2	Normal	Highly Informative	Obtained from expert knowledge, see Appendix X for more information.		N(.1,10);
Y: Mean	Normal	Non-Informative (software default)	n/a	n/a	N(0,10^{10});
Y: Residual variance	Inverse Gamma	Non-Informative (software default)	n/a	n/a	IG(-1,0);

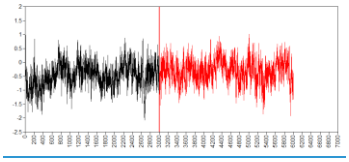
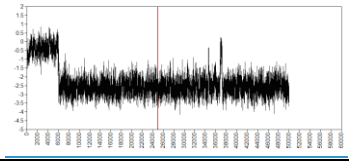
Note. The example is purely hypothetical and serves only to illustrate how to fill in the table. Supplementary documents for this example can be found in the online material in Folder 3.

Table 2. Table to show your supervisor for Points 2, 4-6. Consider a basic regression analysis with 1 dependent variable (Y) and two predictors (X_1 and X_2).

Parameters	**Trace plot (Point 2)**	**Histogram (Point 4)**	**Autocorrelation (Point 5)**	**Kernel density plot (Point 6)**
Y on X_1				
Y on X_2				
Y: Mean				
Y: Residual variance				

*Note***.** Supplementary documents for this example can be found in the online material in Folder 3.

Table 3. An Example of Local Convergence using PTSD Latent Growth Trajectories: Slope Mean Parameter for Latent Class 1

Length of Chain	Parameter Estimate (SD)	Trace Plot	Geweke *z*-statistic (Significant or not):[a]
Shorter chain: 6,000 iterations	-0.309(0.417)		Non-significant
Longer chain: 50,000 iterations	-2.574(0.535)		Non-significant

[a] The Geweke convergence diagnostic compares the first and last halves of the post burn-in portion of the chain. If the *z*-statistic is significant for the Geweke diagnostic, then there is evidence of local convergence. In this case, the burn-in would need to be increased substantially and the local convergence test should be conducted again by doubling the number of iterations to ensure a static statistic is obtained. However, as we see in this Table, it is also possible to obtain results of a non-significant Geweke statistic when local convergence was exhibited. In this case, running the chain out much longer was necessary to identify local convergence problems and obtain a static statistic. The computation of relative deviation becomes particularly important to capture differences in the chains under this circumstance. Supplementary documents for this example can be found in the online material in Folder 4.

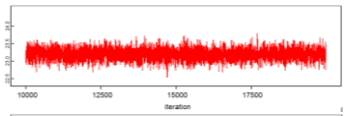
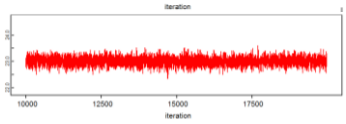
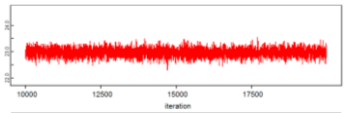
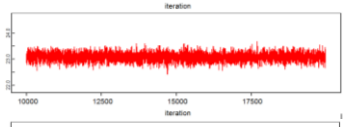
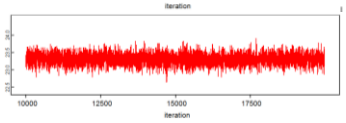
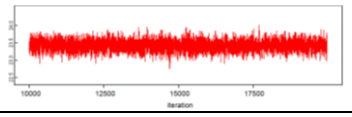
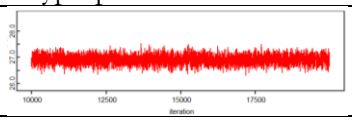
Table 4. Computing Relative Deviation or the Effect of Priors for Model Parameters: Points 3 (section i), 7 (section ii), and 8 (section iii).

Parameters	**Relative Deviation or Size of Effect**	**Convergence Diagnostic**
(i)	**Deviation for Point 3**[a] [(initial converged analysis – analysis with double iterations)/initial converged analysis]*100	Geweke z-statistic (Significant or not):
Y on X_1	[(0.969-0.970)/ 0.969]*100= -0.10	Non-significant
Y on X_2	[(0.650-0.650)/ 0.650]*100= 0.00	Non-significant
Y: Mean	[(0.510-0.511)/ 0.510]*100= -0.19	Non-significant
Y: Residual variance	[(0.953-0.951)/ 0.953]*100= 0.21	Non-significant
(ii)	**Size of the effect for Point 7** [(initial priors – default/non-informative priors)/ initial priors]*100	PSRF (convergence or not)
Y on X_1	[(0.969-0.969)/ 0.969]*100= 0.00	Convergence
Y on X_2	[(0.650-0.650)/ 0.650]*100= 0.00	Convergence
Y: Mean	[(0.510-0.510)/ 0.510]*100= 0.00	Convergence
Y: Residual variance	[(0.953-0.949)/ 0.953]*100= 0.42	Convergence
(iii)	**Size of the effect for Point 8** [(initial priors – default/non-informative priors)/ initial priors]*100	PSRF (convergence or not)
Y on X_1	[(0.969-0.969)/ 0.969]*100= 0.00	Convergence
Y on X_2	[(0.650-0.650)/ 0.650]*100= 0.00	Convergence
Y: Mean	[(0.510-0.510)/ 0.510]*100= 0.00	Convergence
Y: Residual variance	[(0.953-0.953)/ 0.953]*100= 0.00	Convergence

Note. The Geweke convergence diagnostic compares the first and last halves of the post burn-in portion of the chain. If the z-statistic is significant for the Geweke diagnostic, then there is evidence of local convergence. In this case, the burn-in would need to be increased substantially and the local convergence test should be conducted again by doubling the number of iterations to ensure a static statistic is obtained. PSRF = potential scale reduction factor computed from the Gelman and Rubin convergence diagnostic for two chains. Typically values beyond 1.0 ± .05 point toward non-convergence; in this case, priors leading toward different estimates. Supplementary documents for this example can be found in the online material in Folder 3.

[a] initially with 5,000 iterations, alternative model with 10,000 iterations

Table 5. An Example of a Sensitivity Analysis to Examine the Impact of Priors: Points 8 and 9

Chain Comparison	Intercept Estimate (SD)	Trace Plot	PSRF	Size of Effect (Relative Deviation)[a]
Point 8: Compare Subjective Prior to Diffuse Prior				
Subjective Prior: N(31.37,1)	23.19(0.149)			
Compared to: N(0, 10^6)	23.01(0.152)		1.477	0.776%
Point 9: Sensitivity Analysis for Subjective Prior—Altering the Mean Hyperparameter (alter hyperparameters upward and downward)				
Compared to: N(21.37, 1)	22.97(0.149)		1.645	0.948%
Compared to: N(26.37, 1)	23.08(0.149)		1.194	0.474%
Compared to: N(36.37, 1)	23.31(0.150)		1.194	-0.517%
Compared to: N(41.37, 1)	23.42(0.150)		1.646	-0.992%
Point 9: Sensitivity Analysis for Subjective Prior—Altering the Mean and Variance Hyperparameters				
Compared to: N(41.37,.1)	26.91(0.166)		20.442	-16.041%

Note. PSRF = potential scale reduction factor computed from the Gelman and Rubin convergence diagnostic for two chains. Typically values beyond 1.0 ± .05 point toward non-convergence; in this case, priors leading toward different estimates. Note that estimates may be different from Kaplan (2002) due to no covariates being present in the current example. Supplementary documents for this example can be found in the online material in Folder 5.

[a] Percent of relative deviation can be computed as: [((estimate using subjective prior)-(estimate using new prior))/(estimate using subjective prior)]*100. Interpreting percent of relative deviation results is largely subjective and dependent on the metric of the parameters. However, relative deviation under 1% would likely be considered negligible.

Table 6. Actions to Take when WABMS Checklist Points are not Fulfilled

Points	Actions when points are not fulfilled
1	Keep reading and talking to experts until you can explain your priors. Dive into the literature. Explore your network of experts and organize an expert meeting to help inform your selection of priors.
2	Keep increasing the number of iterations until this point is satisfied, despite the length of time it is taking to estimate. Be prepared to wait a long time for models to converge, especially if the model is complex. Be sure your model is properly specified as that can also create problems in convergence.
3	If this point is not reached, then keep doubling the number of iterations. Be sure your model is properly specified as that can also create problems in convergence.
4	Keep sampling until the histogram has a smooth shape, whatever that shape may be. Note that the notion of a "smooth" shape is a bit subjective and often linked to substantive interpretations of parameters. Ultimately, you need to make sure you can trust your results and this can be left to some subjective assessments along the way.
5	If you found higher degrees of autocorrelation according to the plots, then run with double (or more) the number of iterations. If the entire parameter space has been covered and your feel the target distribution (e.g., for the statistic) has been converged upon, then there is nothing to do/worry about.
6	If you see something you do not expect, then look for a model specification error. If an error is suspected or found, then estimate the new model. If there was no error, then consult a statistician to determine whether there is a more technical problem (e.g., with the parameterization of the model or the priors, etc.).
7-9	Do nothing but interpret the findings. Spend extra time justifying your priors and being transparent about your choice. Make a theory-based argument about your model and priors. Then be really transparent about your data-prior conflict.
10	Not applicable.

Figure Captions

Figure 1. Number of papers published with 'Bayesian estimation' in the title or abstract (Source: Scopus).

Figure 2. WAMBS-Checklist**.**

Figure 3. Illustrating the level of information in a histogram, which represents the estimate for the posterior. (A) and (B) illustrate cases where more samples are needed to accurately portray the posterior; (C) and (D) illustrate histograms with adequate information for capturing the nature of the posterior, with (D) representing the most information of the four plots.

Figure 4. Illustrating how substantive results can change and be tracked through a sensitivity analysis of priors.

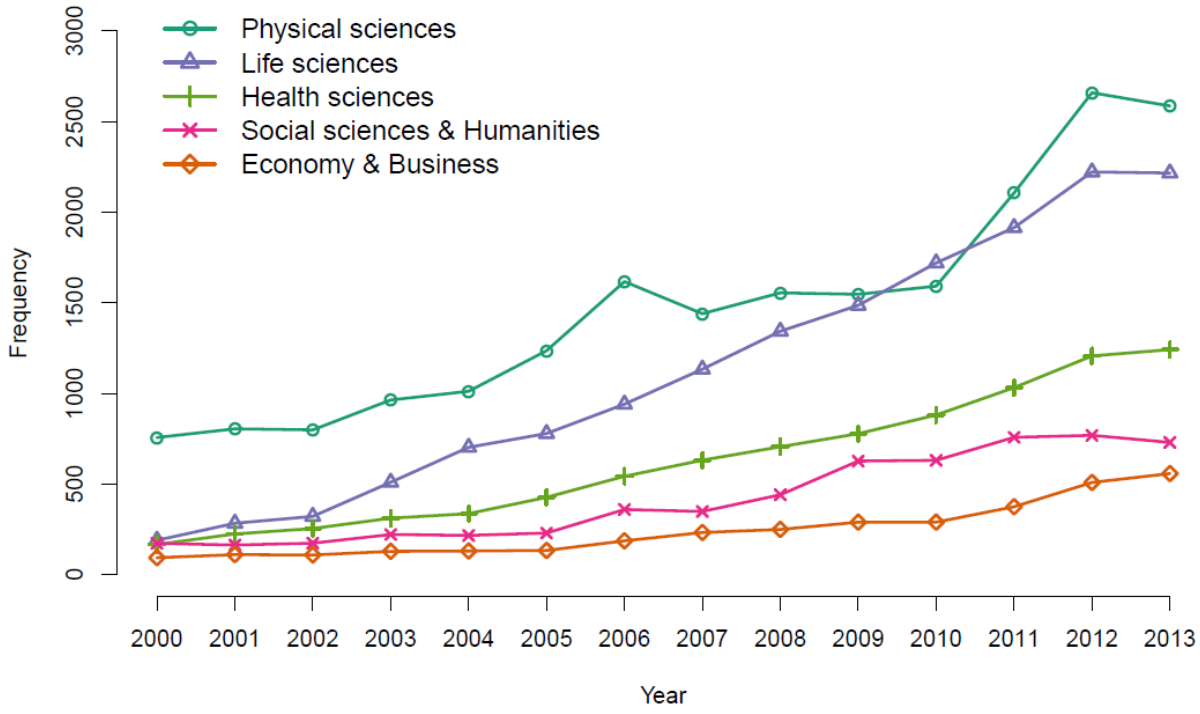


Figure 1. Number of papers published with 'Bayesian estimation' in the title or abstract (Source: Scopus).

THE WAMBS-CHECKLIST

When to worry, and how to Avoid the Misuse of Bayesian Statistics
DEPAOLI & VAN DE SCHOOT (2016)

	Did you show your supervisor…?	Should you worry?	Should you consult an expert?
TO BE CHECKED BEFORE ESTIMATING THE MODEL			
Point 1: Do you understand the priors?	Table 1	YES / NO	YES / NO
TO BE CHECKED AFTER ESTIMATION BUT BEFORE INSPECTING MODEL RESULTS			
Point 2: Does the trace-plot exhibit convergence?	Table 2, column 2	YES / NO	YES / NO
Point 3: Does convergence remain after doubling the number of iterations?	Table 4, columns 2, 3 (i) and akin to Table 3	YES / NO	YES / NO
Point 4: Does the histogram have enough information?	Table 2, column 3	YES / NO	n/a
Point 5: Do the chains exhibit a strong degree of autocorrelation?	Table 2, column 4	YES / NO	YES / NO
Point 6: Does the posterior distribution make substantive sense?	Table 2, column 5	YES / NO	YES / NO
UNDERSTANDING THE EXACT INFLUENCE OF THE PRIORS			
Point 7: Do different specifications of the multivariate variance priors influence the results?	Table 4, columns 2, 3 (ii)	YES / NO	YES / NO
Point 8: Is there a notable effect of the prior when compared with non-informative priors?	Table 4, columns 2, 3 (iii)	NEVER	n/a
Point 9: Are the results stable from a sensitivity analysis?	Sensitivity analysis akin to Table 5 or Figure 4	NEVER	YES / NO
AFTER INTERPRETATION OF MODEL RESULTS			
Point 10: Is the Bayesian way of interpreting and reporting model results used? *(a) Also report on: missing data, model fit and comparison, non-response, generalizability, ability to replicate, etc.*	Text – see Appendix	YES / NO	YES / NO

Figure 2. WAMBS-Checklist.

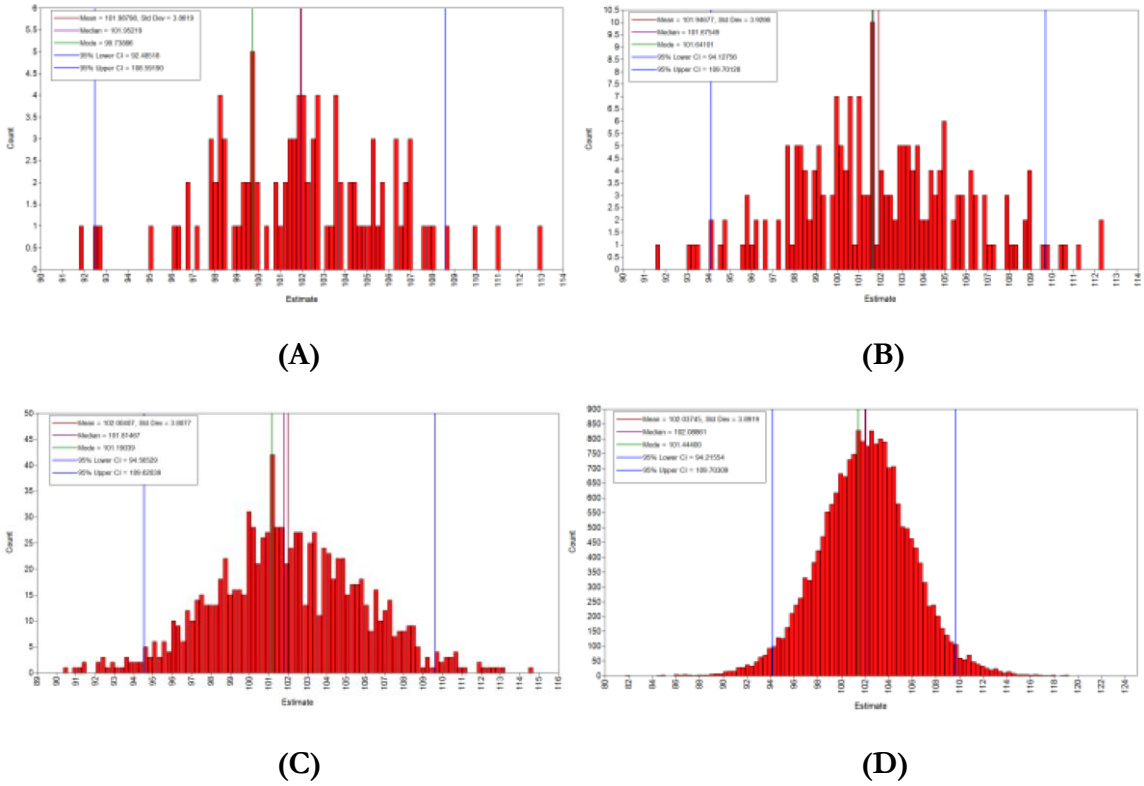


Figure 3. Illustrating the level of information in a histogram, which represents the estimate for the posterior. (A) and (B) illustrate cases where more samples are needed to accurately portray the posterior; (C) and (D) illustrate histograms with adequate information for capturing the nature of the posterior, with (D) representing the most information of the four plots.

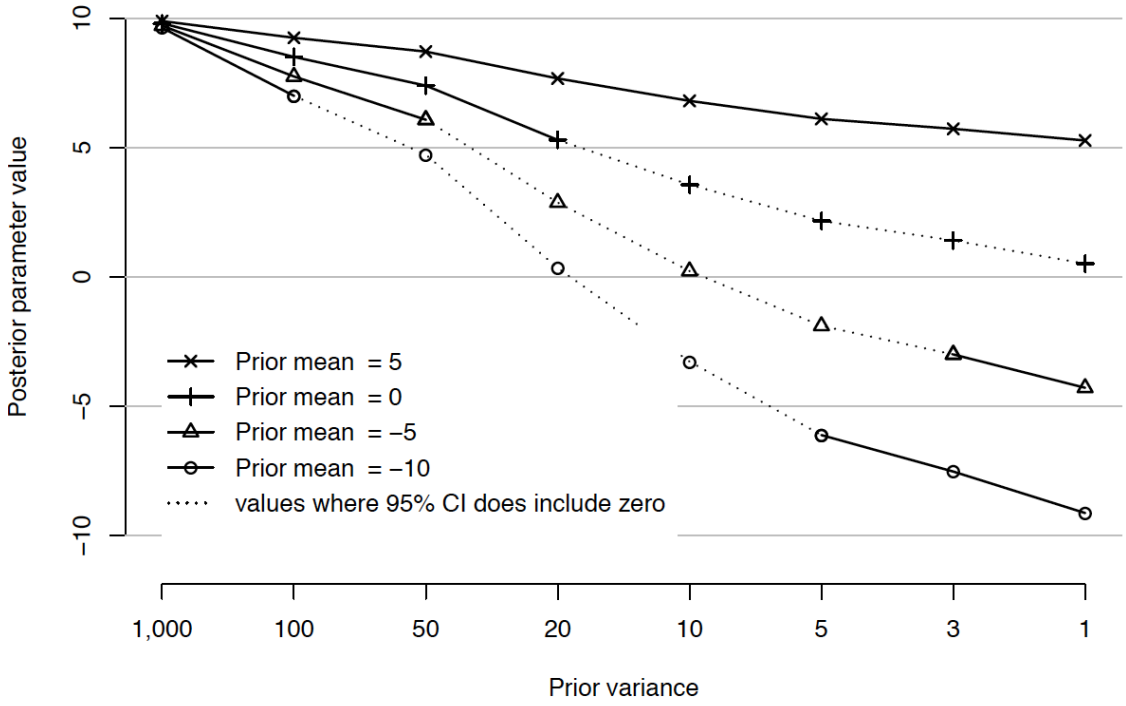


Figure 4. Illustrating how substantive results can change and be tracked through a sensitivity analysis of priors.

Appendix A

The following provides an example of how to write up Bayesian results to adhere to Point 10. Take a simple example of a regression model with predictors X_1 and X_2, and outcome Y. The following is a portion of a contrived Results section for the model results.

For in the analytical strategy:

The regression model was estimated using Bayesian estimation in the M*plus* version 7.3 software program (Muthén & Muthén, 1998-2015) using a seed value of 200 and starting values based on the ML-estimates. Three Markov chains were implemented for each parameter and distinct starting values were provided for each of the chains. To assess chain convergence, the Gelman and Rubin convergence diagnostic was implemented as described in the M*plus* manual with a stricter convergence criterion than the default setting: 0.01 instead of 0.05. An initial burn-in phase of 10,000 iterations was specified, with a fixed number of post burn-in iterations of 10,000. The Gelman and Rubin diagnostic indicated that convergence was obtained with these fixed iterations for each of these three chains. Next, the trace plots for each model parameter were visually inspected. For each of the model parameters, all three chains appeared to converge in that they were visually stacked with a constant mean and variance in the post burn-in portion of the chain. To ensure that convergence was obtained and that local convergence was not an issue, we estimated the model again but with the number of burn-in and post burn-in iterations doubled (40,000 iterations total). Again, the Gelman and Rubin (1992a; 1992b) convergence diagnostic indicated convergence was obtained and the visual inspection of plots was consistent with that finding. Percent of relative deviation can be used to examine how similar (or different) parameter estimates are across multiple analyses. Upon computing the percent of relative deviation for model parameters obtained in these two analyses, we found that results were almost identical with relative deviation levels less than |1|%. The computation for percent of relative deviation for a given model parameter is as follows: [(estimate from initial model) – (estimate from expanded model)/(estimate from initial model)]*100.

We implemented an informative prior for the regression of Y on X_1, and relied on the default prior settings of the software for the other parameters (see M*plus* manual). The background information for specifying the hyperparameters, $\sim N(.5, 0.1)$ stems from the meta-analysis conducted by Author et al. (200x), see Table x on page x. Note that all of the points of the WAMBS-checklist (Depaoli & van de Schoot, 2016) were addressed and the results from this checklist can be requested from the first author (or downloaded as supplementary material).

At the end of the results section:

Because it is important to understand the impact of this theoretically-driven prior, we estimated the model using default non-informative priors in M*plus* as a method for detecting how much influence our informative prior had on the posterior results. Findings from the default prior settings were substantively different in that the default settings indicated the regression parameter estimate was negative and the informative prior settings found that it was positive. Next, we conducted a sensitivity analysis for the informative prior to see what kind of impact the prior might be having. As

indicated in the Methods section, the informative prior based on theory for the regression parameter was distributed normal with hyperparameters of $N(.5, 0.1)$. In this sensitivity analysis, we varied the mean hyperparameter upward and downward by .2 and examined the additional priors: $N(.3, 0.1)$ and $N(.7, 0.1)$. Upon estimating models implementing these two priors, we computed the effect of the priors with the results from the original $N(.5, 0.1)$ prior. The effect of the prior captures the differences between prior settings as "effects" and can be computed using the following equation: Effect of the prior = [(initial prior specification – subsequent prior specification)/initial prior specification]*100. Using this assessment, we found that statistical and substantive findings were comparable for all models in the sensitivity analysis. The Discussion section will detail explanations for the differences in results and the impact these findings may have on the theory under investigation.