1

INTRODUCTION TO BAYESIAN STATISTICS

Milica Miočević

DEPARTMENT OF PSYCHOLOGY, MCGILL UNIVERSITY, MONTREAL, CANADA

Roy Levy

T. DENNY SANFORD SCHOOL OF SOCIAL AND FAMILY DYNAMICS, ARIZONA STATE UNIVERSITY, ARIZONA, UNITED STATES OF AMERICA

Rens van de Schoot

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

Introduction

Bayesian statistics are becoming more popular in many fields of science. See, for example, the systematic reviews published in various fields from educational science (König & Van de Schoot, 2017), epidemiology (Rietbergen, Debray, Klugkist, Janssen, & Moons, 2017), health technology (Spiegelhalter, Myles, Jones, & Abrams, 2000), medicine (Ashby, 2006), and psychology (Van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017) to psychotraumatology (Van de Schoot, Schalken, & Olff, 2017). Bayesian methods appeal to researchers who only have access to a relatively small number of participants because Bayesian statistics are not based on large samples (i.e., the central limit theorem) and hence may produce reasonable results even with small to moderate sample sizes. This is especially the case when background knowledge is available. In general, the more information a researcher can specify before seeing the data, the smaller the sample size required to obtain the same certainty compared to an analysis without specifying any prior knowledge.

In this chapter, we describe Bayes' theorem, which is the foundation of Bayesian statistics. We proceed to discuss Bayesian estimation and Bayes Factors (BFs). The chapter concludes with a brief summary of take-home messages that will allow readers who are new to Bayesian statistics to follow subsequent chapters in this book that make use of Bayesian methods. The applications of Bayesian statistics described in this volume cover the following topics: the role of exchangeability between prior and data (Chapter 2, Miočević et al.), applying the WAMBS checklist (Chapter 3, Van de

Schoot et al.) using informative priors when fitting complex statistical models to small samples (Chapter 4, Veen & Egberts), regression analysis with small sample sizes relative to the number of predictors (Chapter 5, Van Erp), data analysis with few observations from a single participant (Chapter 8, Lek & Arts), updating results participant by participant (Chapter 9, Klaassen), clinical trials with small sample sizes and informative priors based on findings from other trials (Chapter 10, Kavelaars), tests for evaluating whether a finding was replicated (Chapter 12, Zondervan-Zwijnenburg & Rijshouwer), and a comparison between frequentist two-step modeling and Bayesian methods with informative priors (Chapter 17, Smid & Rosseel). Due to space considerations, this chapter does not offer an exhaustive discussion of Bayesian statistics and the differences between Bayesian and classical (frequentist) statistics; for approachable texts on Bayesian statistics in the social sciences, we refer readers to books by Kaplan (2014) and Kruschke (2014), and the chapter by Gigerenzer (1993).

Bayes' theorem

Bayesian statistics are a branch of statistics that implements Bayes' theorem to update prior beliefs with new data:

$$p(\boldsymbol{\theta}|data) = \frac{p(data|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(data)} \propto p(data|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1.1}$$

where $\boldsymbol{\theta}$ denotes a set of parameters (e.g., regression coefficients), $p(\boldsymbol{\theta}|data)$ is the posterior distribution of the parameters, which was obtained by updating the prior distribution of the parameters, $p(\boldsymbol{\theta})$, with the observed data represented by the likelihood function, $p(data|\boldsymbol{\theta})$. The term $p(data)$ is the marginal probability of the data that can be considered a normalizing constant that ensures that the posterior distribution integrates to 1. As the right-hand side of Equation 1.1 shows, excluding this term yields a result that is proportional to the posterior distribution.

In the Bayesian framework, the updated (posterior) beliefs about the parameters in a statistical model are used for inference. The posterior distribution can be summarized to report the probability that a parameter lies within a given range. Bayes' theorem stems from the laws of conditional probabilities, which are not controversial. The controversial elements surrounding Bayesian statistics are *whether* to engage in Bayesian analysis and accept the requirement of specifying a prior distribution, and once the researcher chooses to use Bayesian inference, *how* to specify the prior distribution, $p(\boldsymbol{\theta})$. Applied researchers are often advised to base their prior distributions on previous findings, meta-analyses, and/or expert opinion; for considerations related to the choice of source of prior information, see Chapter 2. The exact influence of the prior is often not well understood, and priors will have a larger impact on the results when sample size is small (see Chapter 3). Bayesian analyses of small data sets using priors chosen by the researcher can sometimes lead to worse estimates than those obtained using uninformative priors or classical methods (Smid, McNeish, Miočević, & Van de Schoot, 2019). Thus, priors should be chosen carefully.

To illustrate a Bayesian statistical analysis, consider a normally distributed variable y (for example, IQ, used to illustrate Bayesian inference in the shiny application example from www.rensvandeschoot.com/fbi/; see also the Center for Open Science (OSF): https://osf.io/vg6bw/) with unknown mean μ and a known variance σ^2. In the frequentist framework, one would collect a sample of data (IQ scores), $y_1, \ldots y_n$, compute the sample mean $\bar{y}$, and use it as the estimate of the population mean of IQ. The standard error is a measure of the uncertainty surrounding the estimate.

In the Bayesian framework, the researcher would start the analysis by specifying a prior distribution for μ (population mean of IQ). When specifying a prior distribution, researchers have to select a distributional form (e.g., normal distribution, *t*-distribution, beta distribution), and specify the parameters of the prior distribution, known as hyperparameters. A common choice of prior distribution for the population mean μ is the normal distribution, which is described by the prior mean (μ_0) and prior variance (σ_0^2) or prior standard deviation (σ_0) or prior precision (τ_0^2) hyperparameters. The mean hyperparameter (μ_0) may be seen as encoding the researcher's best guess about the population mean being estimated, and the variance hyperparameter (σ_0^2) encodes the informativeness (or uncertainty) of the prior distribution. The smaller the variance hyperparameter, the more informative the prior distribution, and the more weight it carries in the analysis. Visually, this analysis is presented in Figure 1.1, where we observe three different situations: panel A depicts an analysis with a sample size of 20 participants from a population where the mean is 100, and the standard deviation is 15; panel B represents the analysis with a sample of 50 participants from that same population; and panel C represents the analysis with a sample of 200 participants from the same population. The prior distribution is the same in all three analyses. Notice how the density of the posterior distribution "moves" closer to the likelihood function as sample size increases from 20–200.

This example has an analytical solution; that is, under the specifications just described, the posterior $p(\mu|y)$ has a known form. It can be shown (Gelman et al., 2013) that the posterior $p(\mu|y)$ is a normal distribution with posterior mean:

$$\mu_p = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \tag{1.2}$$

and posterior variance

$$\sigma_p^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \tag{1.3}$$

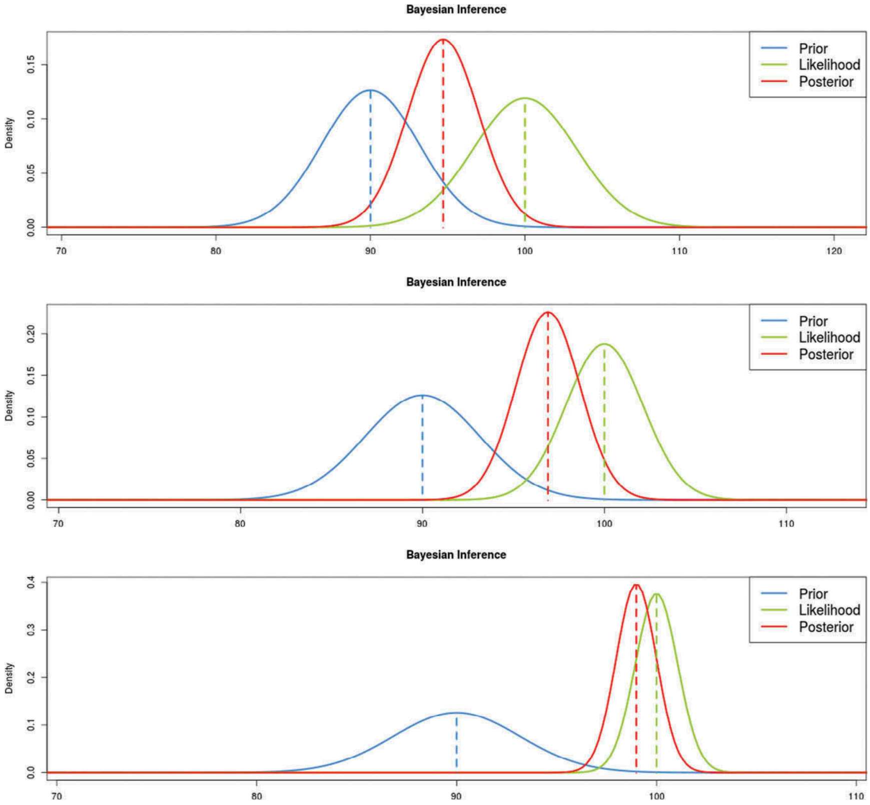

FIGURE 1.1 Plots of the Bayesian computation of a mean parameter with a known variance obtained using the shiny application available at www.rensvandeschoot.com/tutorials/fbi/ (see also the OSF: https://osf.io/vg6bw/)

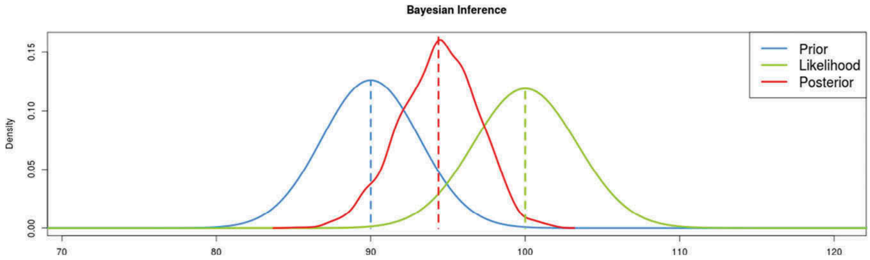

FIGURE 1.2 Plots of the Bayesian computation of a mean parameter with an unknown variance obtained using the shiny application available at www.rensvandeschoot.com/tutorials/fbi/ (see also the OSF: https://osf.io/vg6bw/)

where μ_0 denotes the mean of the normal prior distribution, $\bar{y}$ denotes the observed mean in the sample, n is the sample size, σ_0^2 is the variance hyperparameter in the prior, and σ^2 is the variance in the observed sample. Both the prior and posterior are normal distributions; this is the case because the normal prior distribution is a conjugate prior for the mean parameter. All conjugate prior distributions, when multiplied by the likelihood function, yield posterior distributions from the same distributional family. We can use Equations 1.2 and 1.3 to obtain the analytical solution for the mean and variance of the posterior for the mean of IQ. If we select the prior mean of IQ to be $\mu_0 = 90$ and the prior variance equal to $\sigma_0^2 = 10$, and we observe a sample of 20 participants for which the sample mean of IQ is $\bar{y} = 100$ and the sample variance is $\sigma^2 = 225$, we end up with a posterior distribution centered around $\mu_p = \frac{\frac{1}{10}90 + \frac{20}{225}100}{\frac{1}{10} + \frac{20}{225}} = 94.71$ with a posterior variance equal to $\sigma_p^2 = \left(\frac{1}{10} + \frac{20}{225}\right)^{-1} = 5.29$. Notice how the posterior mean, μ_p, is a "compromise" between the prior mean μ_0, and the mean of the variable in the observed data set, $\bar{y}$. Notice also how decreasing the prior variance (σ_0^2) gives the prior mean more weight, and how increasing the sample size n gives the observed data more weight in determining the posterior mean.

Bayesian estimation

In the example where it is of interest to estimate the mean of a population with a known variance, it is possible to obtain the posterior distribution analytically. However, most statistical models in the social sciences are more complex, and the posterior distribution cannot be obtained analytically. In these situations, results are obtained by progressively approximating the posterior distribution using Markov Chain Monte Carlo (MCMC; Brooks, Gelman, Jones, & Meng, 2011). MCMC is an iterative procedure, like maximum likelihood (ML). However, unlike ML, which seeks to maximize the likelihood function, MCMC seems to approximate the entire posterior distribution. Figure 1.2 illustrates an approximation of the posterior for the same analysis as in panel A of Figure 1.1 obtained using MCMC instead of using the analytical solution; note that the distribution is no longer smooth because it is an approximation of the posterior. In the following paragraphs, we briefly survey some of the practical aspects involved in utilizing MCMC for Bayesian analyses.

In a Bayesian analysis, MCMC proceeds by simulating values from distributions such that, in the limit, the values may be seen as draws from the posterior distribution (for visual representations of multiple chains, see Figure 3.4 in Chapter 3). A properly constructed chain will eventually converge to the point where the subsequent simulated values may be seen as samples from the posterior; however, there is no guarantee as to *when* that will happen. Though there is no way of definitively knowing that a chain has converged to the posterior distribution, there are several techniques one can use to find evidence of convergence (Cowles & Carlin, 1996).

In the social sciences literature, the most commonly encountered convergence diagnostics are those offered by the majority of software packages, which include the Potential Scale Reduction factor (Gelman & Rubin, 1992), Geweke's diagnostic (1992), and trace plots of draws plotted against the iteration number for each parameter (Brooks, 1998; see Chapter 3 for information about how to obtain and interpret trace plots). Several convergence diagnostics rely on running multiple chains from dispersed starting values for different chains in order to assist with the monitoring of convergence (Gelman & Shirley, 2011). The generated values from the chain prior to convergence are referred to as burn-in iterations and are discarded; values from the chain after convergence are taken to be draws from the posterior and can be summarized to represent the posterior. In theory, the more draws are taken from the posterior, the better it is approximated.

A complicating factor for MCMC is the within-chain correlation of the draws (see Figure 3.8 in Chapter 3); for a more detailed discussion on autocorrelation and possible solutions see Chapter 3. It is often recommended to use thinning[1] to reduce the autocorrelation between the retained draws (Gelman & Shirley, 2011). However, some researchers argue that thinning can be problematic for obtaining precise summaries of the posterior (Link & Eaton, 2012) and that it is better to run longer chains than to thin. *Stopping time* refers to ending the sampling and depends on time constraints, how long the chain(s) ran before convergence, the researcher's confidence that convergence was reached, and the autocorrelation between draws (see Chapter 3). The number of draws to retain after convergence (i.e., post burn-in) should be determined in part by the precision with which the researcher wants to estimate the posterior, or its features. Estimating broad summaries, such as the posterior mean, tends to require fewer draws than features out in the tails, such as extreme percentiles (Kruschke, 2014).

To summarize the posterior, all non-discarded draws (i.e., all draws after burn-in) from all chains should be mixed together (Gelman & Shirley, 2011). Features of these draws (e.g., mean, standard deviation, intervals) are seen as estimates of the corresponding features of the posterior distribution. Common point summaries of the posterior are the mean, median, and mode. Common interval summaries are $(1 - \alpha)\%$ equal-tail credibility intervals, which are constructed from the $(\alpha/2)^{th}$ and $(1 - \alpha/2)^{th}$ percentiles of the posterior distribution, and highest posterior density credibility intervals which have the property that no values outside the interval are more probable than any values inside the interval.

Bayes Factors

Null hypothesis significance testing (NHST) has been the dominant approach to statistical inference in the social sciences since the 1940s (Gigerenzer, 1993). NHST belongs to the family of frequentist statistics, which define probability as the frequency of an event. Two quantities that stem from this school of statistics and rely on the above definition of probability are *p*-values and confidence intervals. The *p*-value quantifies the probability of finding the observed or a more extreme result given that the null hypothesis is true, and the $(1 - \alpha)\%$ confidence intervals tell us that upon repeated sampling, $(1 - \alpha)\%$ of the confidence intervals will contain the true value of the parameter (Jackman, 2009). The reliance on NHST and *p*-values has been criticized for decades (Bakan, 1966; Ioannidis, 2005; Rozeboom, 1960). Some researchers advocate for the replacement of *p*-values with alternatives such as effect size measures and confidence intervals (Cumming, 2014). Others have argued for abandoning the frequentist paradigm altogether because the *p*-value does not quantify the probability of the hypothesis given the data (Wagenmakers, Wetzels, Borsboom, & Van der Maas, 2011), nor does it provide any measure of whether the finding is replicable (Cohen, 1994), and confidence intervals do not have the properties they are ascribed to have and are easily misunderstood (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016).

In the Bayesian framework, it is possible to calculate the probability of a hypothesis given the data, and to compute the posterior odds in favor of one hypothesis (or model) relative to another hypothesis (or model; Kass & Raftery, 1995). The ratio of posterior probabilities is equal to the ratio of prior probabilities multiplied by the ratio of marginal likelihoods under each hypothesis:

$$\frac{p(H_2|data)}{p(H_1|data)} = \frac{p(H_2)}{p(H_1)} \times \frac{\int_{\theta(2)} p(data|\theta_{(2)})p(\theta_{(2)}|data)d\theta_{(2)}}{\int_{\theta(1)} p(data|\theta_{(1)})p(\theta_{(1)}|data)d\theta_{(1)}} \tag{1.4}$$

The last term on the right-hand side, the ratio of marginal likelihoods, is also called the Bayes Factor (Kass & Raftery, 1995; Raftery, 1993). BFs are a way of comparing two competing hypotheses (H_1 and H_2) and are calculated by dividing the integrated likelihoods of the two models (Jeffreys, 1998). Chapters 9 and 12 make use of *BF*; the readers will notice that there are notational differences between chapters, and this is the case in the literature as well. However, the meaning and interpretations of *BF* are the same as described in this chapter, unless the authors indicate otherwise. If the prior probabilities of the two models are both set to 0.5, then the posterior odds equal the BF. If the prior probabilities are not .5, then the BF is not equal to the posterior odds. However, the BF still captures the weight of evidence from the data in favor of one hypothesis. A BF of 1 indicates that the data do not support one hypothesis more than the other, a BF below 1 indicates that the data provide support for H_1 over H_2, and a BF above 1 indicates that the data support H_2 over H_1. The computation of the BF does not require nesting of the models being compared. Unlike classical hypothesis tests, BFs can support a null hypothesis. In the words of Dienes (2014, p. 1), BFs "allow accepting and rejecting the null hypothesis to be put on an equal footing", but as indicated by Konijn, Van de Schoot, Winter, & Ferguson (2015), we should avoid BF-hacking (cf., "God would love a Bayes Factor of 3.01 nearly as much as a BF of 2.99"). Especially when BF values are small, replication studies and Bayesian updating are still necessary to draw conclusions (see Chapter 12 for more on this topic).

Conclusion

In this brief introductory chapter, we sought to inform readers about the fundamental concepts in Bayesian statistics. The most important take-home messages to remember are that in Bayesian statistics, the analysis starts with an explicit formulation of prior beliefs that are updated with the observed data to obtain a posterior distribution. The posterior distribution is then used to make inferences about probable values of a given parameter (or set of parameters). Furthermore, BFs allow for comparison of non-nested models, and it is possible to compute the amount of support for the null hypothesis, which cannot be done in the frequentist framework. Subsequent chapters in this volume make use of Bayesian methods for obtaining posteriors of parameters of interest, as well as BFs.

Note

[1] Thinning is the practice of retaining only every k^{th} draw, where the thinning parameter k is chosen so that the retained draws are approximately independent. However, thinning represents a loss of information and is not necessary, and "as long as a sequence has converged and the number of iterations retained is substantial, it makes no practical difference if we keep all or every 25th or every 50th iteration" (Scheines, Hoijtink, & Boomsma, 1999, p. 42).

References

Ashby, D. (2006). Bayesian statistics in medicine: A 25 year review. *Statistics in Medicine*, *25* (21), 3589–3631. doi:doi.org/10.1002/sim.2672.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. doi:10.1037/h0020412.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society. Series D (the Statistician)*, *47*(1), 69–100. doi:10.1111/1467-9884.00117.

Brooks, S. P., Gelman, A., Jones, G. L., & Meng, X.-L. (Eds.). (2011). *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman & Hall/CRC Press.

Cohen, J. (1994). The earth is round (p <. 05). *American Psychologist*, *49*(12), 997–1003. doi:10.1037/0003-066X.49.12.997.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*(434), 883–904. doi:10.2307/2291683.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. doi:10.1177/0956797613504966.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton: FL: CRC Press.

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

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

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

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren and C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124.

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

Jeffreys, H. (1998). *The theory of probability*. Oxford: Oxford University Press.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi:10.1080/01621459.1995.10476572.

König, C., & Van de Schoot, R. (2017). Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review*, 1–24. doi:10.1080/00131911.2017.1350636.

Konijn, E. A., Van de Schoot, R., Winter, S. D., & Ferguson, C. J. (2015). Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. *Communication Methods and Measures*, *9*(4), 280–302.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and stan* (2nd ed.). Boston, MA: Academic Press.

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, *3*(1), 112–115.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. doi:10.3758/s13423-015-0947-8.

Raftery, A. E. (1993). Bayesian model selection in structural equation models. *Sage Focus Editions*, *154*, 163.

Rietbergen, C., Debray, T. P. A., Klugkist, I., Janssen, K. J. M., & Moons, K. G. M. (2017). Reporting of Bayesian analysis in epidemiologic research should become more transparent. *Journal of Clinical Epidemiology*, *86*, 51–58.e52. doi:10.1016/j.jclinepi.2017.04.008.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. doi:10.1037/h0042040.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*(1), 37–52. doi:10.1007/BF02294318.

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

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., & Abrams, K. R. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment*, *4*(38), 1–130.

Van de Schoot, R., Schalken, N., & Olff, M. (2017). Systematic search of Bayesian statistics in the field of psychotraumatology. *European Journal of Psychotraumatology*, *8*(sup1). doi:10.1080/20008198.2017.1375339.

Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. doi:10.1037/met0000100.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, *100*(3), 426–432. doi:10.1037/a0022790.