Psychon Bull Rev (2018) 25:155–177
DOI 10.3758/s13423-017-1272-1

BRIEF REPORT

Bayesian data analysis for newcomers

John K. Kruschke[1] · **Torrin M. Liddell**[1]

Published online: 12 April 2017

Abstract This article explains the foundational concepts of Bayesian data analysis using virtually no mathematical notation. Bayesian ideas already match your intuitions from everyday reasoning and from traditional data analysis. Simple examples of Bayesian data analysis are presented that illustrate how the information delivered by a Bayesian analysis can be directly interpreted. Bayesian approaches to null-value assessment are discussed. The article clarifies misconceptions about Bayesian methods that newcomers might have acquired elsewhere. We discuss prior distributions and explain how they are not a liability but an important asset. We discuss the relation of Bayesian data analysis to Bayesian models of mind, and we briefly discuss what methodological problems Bayesian data analysis is not meant to solve. After you have read this article, you should have a clear sense of how Bayesian data analysis works and the sort of information it delivers, and why that information is so intuitive and useful for drawing conclusions from data.

Keywords Bayesian model · Bayesian analysis · Null hypothesis significance test · Bayes factor · Highest density interval · Region of practical equivalence · p value · Confidence interval · Replication crisis

✉ John K. Kruschke
johnkruschke@gmail.com;
http://www.indiana.edu/∼kruschke/

[1] Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington, IN 47405, USA

This article explains the basic ideas of Bayesian data analysis. The article uses virtually no mathematical notation. The emphases are on establishing foundational concepts and on disabusing misconceptions. This article does not rehearse the many reasons to be wary of p values and confidence intervals (see, for example, Kruschke, 2013; Kruschke & Liddell, 2017; Wagenmakers, 2007). We assume that you already are curious about Bayesian methods and you want to learn about them but you have little previous exposure to them. For readers with some previous experience, we hope that the framework described here provides some useful unification and perspective.

The first section of the article explains the foundational ideas of Bayesian methods, and shows how those ideas already match your intuitions from everyday reasoning and research. The next sections show some simple examples of Bayesian data analysis, for you to see how the information delivered by a Bayesian analysis can be directly interpreted. We discuss Bayesian parameter estimation, Bayesian model comparison, and Bayesian approaches to assessing null values. The final sections focus on disabusing possible misconceptions that newcomers might have. In particular, we discuss when prior distributions are critical in an analysis and when they are not, we discuss the relation of Bayesian data analysis to Bayesian models of mind, and we briefly discuss what methodological problems Bayesian data analysis is not meant to solve. After you have read this article, you should have a clear sense of how Bayesian data analysis works and the sort of information it delivers, and why that information is so intuitive and useful for drawing conclusions from data. We hope the article provides a clear conceptual framework that makes subsequent learning much easier.

The main idea: Bayesian analysis is reallocation of credibility across possibilities

The main idea of Bayesian analysis is simple and intuitive. There are some data to be explained, and we have a set of candidate explanations. Before knowing the new data, the candidate explanations have some prior credibilities of being the best explanation. Then, when given the new data, we shift credibility toward the candidate explanations that better account for the data, and we shift credibility away from the candidate explanations that do not account well for the data. A mathematically compelling way to reallocate credibility is called Bayes' rule. The rest is just details.

You already are Bayesian in everyday reasoning

Bayesian reallocation of credibility across possibilities is so intuitive that you already do it in everyday life. Here are two examples.

Sherlock Holmes' reasoning is Bayesian

The fictional detective Sherlock Holmes was famous for remarking to his sidekick, Dr. Watson: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Doyle, 1890, Ch. 6). That simple reasoning is Bayesian. Mr. Holmes begins a case with a set of candidate culprits for the crime. Some suspects seem more suspicious than others, and what makes detective novels fun is that the true culprit is usually someone who has very small prior probability of being guilty. Holmes then does his detective work and collects new information, what in scientific investigations would be called the data. From the data, suspicion is reallocated across the suspects. If the data eliminate some suspects, the remaining suspects must be more suspicious, even if their prior probability was small.

Bayesian analysis does exactly the same reallocation, but using precise mathematics. Figure 1 illustrates the reallocation graphically. Suppose there are four possible suspects for a crime, named Amy, Bob, Cai, and Don. Based on knowledge at the beginning of the investigation, Amy is thought to be most suspicious and Don least suspicious, as plotted in the prior distribution of panel A. During the course of the investigation, new data give airtight alibis to Amy, Bob, and Cai. Therefore all suspicion is re-allocated to Don, as plotted in the posterior distribution of panel B. The sort of graphical display in Fig. 1 will be used later in this article to illustrate realistic applications in data analysis. The horizontal axis denotes the range of possibilities, and the vertical axis denotes the credibility, or probability, of each possibility.

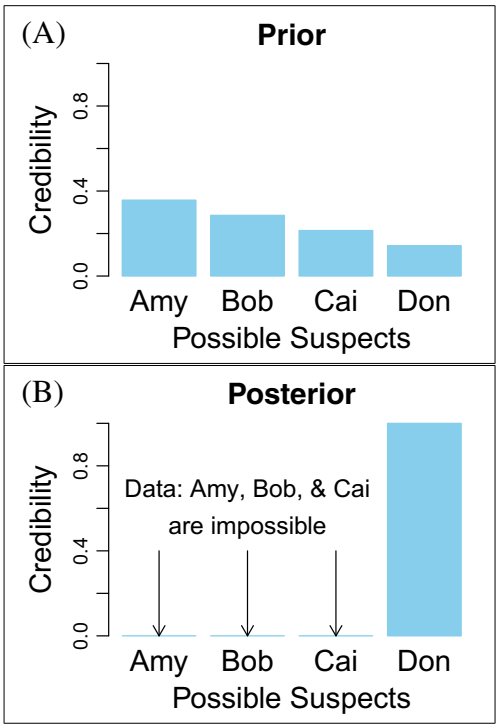

Fig. 1 The Bayesian reasoning of Sherlock Holmes. **A** Prior distribution of credibility across suspects, of the claim that the suspect committed the crime. **B** After data indicate that suspects Amy, Bob, and Cai could not have committed the crime, the posterior distribution loads all credibility onto the claim that Don committed the crime

In Holmes' reasoning, a tacit premise is that the actual culprit is included in the set of suspects. A more accurate phrasing for Holmesian reasoning is this: When you have eliminated the impossible, whatever remains, however improbable, must be the least bad option from among the possibilities you are considering. In general, Bayesian reasoning provides the relative credibilities within the set of considered possibilities.

Exoneration is Bayesian

Every pre-school child knows the logic of exoneration. Suppose there's a window broken during a time when some children were playing nearby. The angry adult knows that Tommy and Joey were among the children at the scene. Innocent little Tommy is greatly relieved when Joey confesses to accidentally breaking the window, because Tommy knows he is exonerated by Joey's confession. The logic of exoneration is Bayesian: Reasoning starts with a set of candidate causes of the event, then collects new data such as a confession, and then reallocates credibility accordingly. If the data fully implicate one suspect, the remaining (unaffiliated) suspect must be less suspicious. Bayesian analysis does the same reallocation, but with exact mathematics.

The possibilities are parameter values

What do the examples of the previous section have to do with data analysis? The connection is this: In data analysis, the candidate explanations are values of parameters in mathematical descriptions. For example, suppose we randomly poll ten people about an upcoming referendum, and we find that seven intend to vote yes and the remaining three intend to vote no. Given those data, what should we believe about the proportion of people voting yes in the general population? How credible is a proportion of 0.60, or 0.70, or 0.80, and so on? Intuitively, we should allocate more credibility to proportions near 0.70 than to proportions far from 0.70 because the data showed 7/10 yes votes. But with only ten people polled, we should not allocate credibility too narrowly only to proportions extremely close to 0.70, because proportions such as 0.60 or 0.80 could easily have generated 7/10 in the data.

In this case, the underlying proportion is a parameter in a simple coin-flipping model of the data. We conceive of each randomly polled response as a flip of a coin that has some underlying probability of coming up yes. We start with some prior allocation of credibilities across the continuum of possible parameter values. The prior allocation could be quite vague and spread evenly across the range of candidate values from 0 to 1, or the prior could give some candidate proportions higher credibility than others if previous knowledge recommends it. Then we collect data and re-allocate credibility to parameter values that are consistent with the data. Bayesian analysis provides the mathematical form of the reallocation and more details will be provided later in the article. In this section the emphasis is on the notion of a parameter value as a candidate explanation of the data.

Parameter values are meaningful in the context of their model

We care about parameter values because they are meaningful. To illustrate, suppose we collect the weights of all the children in third grade of a particular school. We describe the set of data as being randomly generated from a normal distribution with mean μ and standard deviation σ. If we tell you that certain values of μ and σ are good descriptions of the data, then you have a pretty good idea of what the data look like. The values of the parameters are meaningful in the context of the model. As another example, in linear regression, the regression coefficient indicates how much the predicted value changes when the predictor increases by one unit. As yet another example, in analyses of multiple-group designs, contrast parameters indicate differences of group means. In all these situations, data are described by mathematical models, and the parameters are meaningful. Bayesian analysis tells us the relative credibilities of the parameter values. That's why the information provided by a Bayesian analysis is so useful and intuitive.

Parameters can be discrete or continuous

In many mathematical descriptions of data, the parameters are continuous, such as means, standard deviations, and regression coefficients. The posterior distribution on continuous parameters is a continuous distribution, rising and falling smoothly across the range of the joint parameter space. As nearly all statistical models involve continuous parameters, it is continuous parameter distributions that dominate Bayesian data analysis.

But descriptive parameters can also be discrete, not continuous. Perhaps the most common example is disease diagnosis from an imperfect diagnostic test. The patient takes a diagnostic test that provides a datum, namely, a result of "positive" or "negative." The test is imperfect, and has a non-zero false-alarm rate and an imperfect correct-detection rate. The parameter value to be inferred is the underlying state of the patient, which has two discrete values: "has disease" or "does not have disease." There is a prior distribution on the discrete parameter values, which typically has low probability on "has disease" and high probability on "does not have disease." The probability of having or not having the disease is re-allocated in light of the test result.

There are many other examples of Bayesian inference over discrete parameters. We already considered examples in the context of Sherlock Holmes and exoneration. In those cases, the parameter (i.e., the variable over which credibility was re-allocated) was the set of candidate culprits. The data collected by sleuthing reallocated probability more heavily onto some suspects than others. As another example of a discrete parameter, consider a historical document of disputed authorship for which there are several possible authors. Based on characteristics from examples of the authors' writings with established provenance, we can infer the relative posterior probabilities that each candidate author wrote the document in question (e.g., Mosteller & Wallace, 1984; Peng, Schuurmans, & Wang, 2004). Analogous techniques can be applied to inferring the probable identity of a speaker from his or her voice, or inferring the probability that an incoming email is spam based on its content.

Another important case of a discrete parameter, that we will revisit later, is model comparison. For a given set of data, there can be multiple models. Each model involves its own parameters *and* prior distribution over its parameters. The models are labeled by a discrete indexical parameter ("1" for the first model, "2" for the second model, and so on). When new data are considered, credibility shifts over the parameter distributions within each model, and

credibility simultaneously shifts over the discrete indexical parameter. The re-allocated posterior probabilities of the model indices are the relative degrees to which we believe each model, given the data. In particular, when one model represents a null hypothesis and a second model represents an alternative hypothesis, this discrete model comparison is one way of implementing hypothesis testing in a Bayesian framework (e.g., Edwards, Lindman, & Savage, 1963; Kruschke, 2011, 2015; Wagenmakers, 2007).

Because the mathematics of Bayesian inference is relatively easy to describe for discrete parameters and for simple diagnostic tests, introductory expositions tend to use examples with discrete parameters. Consequently, beginners can get a mistaken impression that Bayesian inference always involves discrete hypotheses. For instance, the classic article of Edwards et al. (1963) emphasized Bayesian discrete hypothesis testing, even though the article started with a description of Bayesian inference for continuous parameters. The more general application is for continuous parameters; discrete parameters are just a special case.

Bayesian analysis provides the relative credibilities of parameter values

The goal of Bayesian data analysis is to provide an explicit distribution of credibilities across the range of candidate parameter values. This distribution, derived after new data are taken into account, is called the posterior distribution across the parameter values. The posterior distribution can be directly examined to see which parameter values are most credible, and what range of parameter values covers the most credible values.

The posterior distribution can be directly interpreted. We can "read off" the most credible parameter values and the range of reasonable parameter values. Unlike in frequentist statistical analysis, there is no need to generate sampling distributions from null hypotheses and to figure out the probability that fictitious data would be more extreme than the observed data. In other words, there is no need for p values and p value based confidence intervals. Instead, measures of uncertainty are based directly on posterior credible intervals.

You already are Bayesian in data analysis

Bayesian reasoning is so intuitive that it's hard to resist spontaneously giving Bayesian interpretations to results of traditional frequentist data analysis. Consider, for example, the t test for a single group of 50 people to whom we administered a "smart drug" and then an intelligence-quotient (IQ) examination. We would like to know if the mean IQ score of the group differs from the general population average of 100. Suppose the t test yields $t(49) = 2.36$, $p = 0.023$, with 95% confidence interval on μ extending from 100.74 to 109.26. What does this mean?

Your intuitive interpretation of p values is Bayesian

Consider the result that $p = 0.023$. This means that the probability that μ equals the "null" value of 100 is only 0.023, right? This is a natural, intuitive interpretation and is the one that many or most people give (e.g., Dienes, 2011; Gigerenzer, 2004; Haller & Krauss, 2002, and references cited therein). Unfortunately, it is not the correct interpretation of the p value. It seems that people are interpreting a p value as if it were the result of a Bayesian analysis. Bayesian parameter estimation can provide the probability that μ is within a narrow interval around the null value. A Bayesian hypothesis test provides the probability that μ equals the null value relative to an alternative hypothesis that μ could span a wide range of values. In other words, we naturally interpret a frequentist p value as if it were some form of Bayesian posterior probability.

But a frequentist p value is not a Bayesian posterior probability. The p value is the probability that the observed data summary (such as its t value), or something more extreme than observed would be obtained if the null hypothesis were true and the data were sampled according to the same stopping and testing intentions as the observed data. In other words, the p value is the probability that fictional, counterfactual data from the null hypothesis would be more extreme than the observed data, when those data were sampled and tested as intended by the current researchers. Different stopping and testing intentions therefore yield different p values (e.g., Kruschke, 2013, 2015, Ch. 11; Kruschke & Liddell, 2017). In summary, the correct interpretation of the p value is very counterintuitive, and the intuitive interpretation of the p value is as Bayesian posterior probability. Unfortunately, the p value is not a Bayesian posterior probability.

Your intuitive interpretation of confidence intervals is Bayesian

Consider the 95% confidence interval from 100.74 to 109.26. That means there is a 95% probability that the mean μ falls between 100.74 and 109.26, right? This is a natural, intuitive interpretation, and is the one that most people give (e.g., Morey, Hoekstra, & Rouder, 2015). It is also the correct interpretation of a Bayesian credible interval. That is, if we were to report from a Bayesian analysis that the posterior distribution on μ has its 95% most credible values between 100.74 and 109.26, then we would correctly say that we believe the mean μ has 95% probability of falling between

100.74 and 109.26. In other words, we naturally interpret a frequentist confidence interval as if it were a Bayesian credible interval.

But a frequentist confidence interval is not a Bayesian credible interval. A 95% confidence interval is the range of parameter values we would not reject at $p < .05$ (Cox, 2006, p. 40). In other words, the confidence interval is tied to the same fictional, counterfactual data sampling as the p value. Different stopping and testing intentions yield different confidence intervals (e.g., Kruschke, 2013, 2015, Ch. 11; Kruschke & Liddell, 2017). In summary, the correct interpretation of the confidence interval is very counterintuitive, and the intuitive interpretation of the confidence interval is as a range on a Bayesian posterior probability distribution.

Build your intuition with simple examples

The previous section attempted to convey two key ideas: Bayesian inference is reallocation of credibility across possibilities, and, the possibilities are values of parameters in a mathematical description of data. Thus, we start with a prior distribution over parameter values, then consider new data, and arrive at a posterior distribution over parameter values. The posterior distribution places higher credibility on parameter values that are more consistent with the data. Figure 1 summarized this sort of re-allocation graphically, as a shift in the heights of bars over the space of discrete possibilities. In this section we consider a few more simple examples of this process, applied to more typical cases with continuous ranges of possibilities.

The probability of dichotomous data

Suppose we want to know how well a drug cures a disease. Suppose we have a specific dosage and a specific operationalization for measuring whether a patient is cured or not. Of 50 patients treated, 36 are cured. From those data, what are credible values of the underlying probability of cure? Intuitively, the most credible probability is about $36/50 = 0.72$, but we know that the data are a random sample and hence there is a range of uncertainty.

We start with a model of the data that has a meaningful parameter. Our model is simple: The probability of cure is the value of parameter θ, which has possible values on the continuous interval from 0 to 1. Before considering the new data, we establish a prior distribution of credibility of the parameter values. For purposes of illustration, we will suppose that we have little prior knowledge of the cure rates for this disease and drug, and we will suppose that all values in the range from 0 to 1 are essentially equally possible, as shown in panel A of Fig. 2.

Bayesian analysis provides a quantitatively precise redistribution of credibility over the parameter values, given the data. Panel B of Fig. 2 shows the posterior distribution. Notice that the posterior distribution is peaked near the proportion of cures in the data, and the credibility drops off for values of θ above or below the data proportion.

The range of values of θ that includes the 95% most credible values is marked in the posterior distribution as the 95% HDI. HDI stands for *highest density interval*, which refers to the technical terminology of "probability density" instead of the colloquial but accurate term "credibility." Crucially, every parameter value within the HDI has higher credibility (i.e., higher probability density) than any parameter outside the HDI. We refer to the "most credible" parameter values as the parameter values with highest probability density. The 95% HDI contains a total probability of 95%. We use a 95% probability mass for the HDI, as opposed to 90% or 99% or whatever, merely because of familiarity by analogy to 95% confidence intervals in frequentist statistics. We say that the 95% HDI contains the 95% most credible values of the parameter.

Importantly, the posterior distribution reveals the credibility of every value of the meaningful parameter. From the posterior distribution, we merely "read off" whatever we may want to know about the parameter estimate, such as the most credible value (i.e., the mode of the distribution), the value of median or mean credibility, and the exact range of

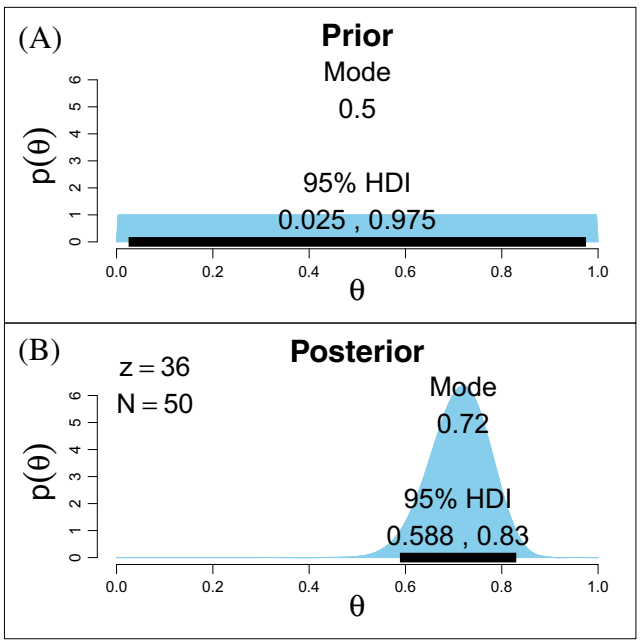

Fig. 2 Estimating the probability θ that a patient is cured when given a particular drug. **A** Prior distribution is broad over possible values of the parameter θ. **B** For $N = 50$ patients with $z = 36$ cures, the posterior distribution over the parameter θ is much narrower. HDI limits and modes are displayed to first three digits only. Compare with Fig. 1

uncertainty as indicated by the HDI. Later in this article it will be explained how to assess null values, such as whether the cure rate of the drug is meaningfully greater than a baseline value of, say, $\theta = 0.5$. For now, notice that it is easy to see where the null value falls in the posterior distribution.

The mean and standard deviation of normally distributed metric data

Consider now a case in which the data are metric, not dichotomous as in the previous example. Suppose we are considering the influence of a "smart drug" on IQ scores. We administer the smart drug to $N = 63$ people chosen at random, and the sample mean is 110 with a standard deviation of 20, and a histogram of the data appears to be roughly normal (not shown here). Suppose we wish to describe the set of data as a normal distribution that has mean μ and standard deviation σ. What are credible values of μ and σ for the normal model, and what is our uncertainty in the credible parameter values?

We start with a prior distribution on the infinite two-dimensional parameter space that represents all possible combinations of values for the mean (μ) and standard-deviation (σ) parameters of the normal model. For our present purposes, we suppose that we have very limited prior knowledge about how the data might come out, except for rough knowledge about the magnitude of the measurements. In this case, the data are IQ scores, which are normed for the general population to have a mean of 100 and a standard deviation of 15. But we don't know much about scores from a smart-drug group, so we may suppose that the mean will be somewhere in the interval between 0 and 250, and the standard deviation will be somewhere in the interval between 0 and 40. (Later in the article we discuss the choice of prior distribution.) The broad prior distribution on this interval is shown in panel A of Fig. 3.

The posterior distribution is shown in panel B of Fig. 3. Notice that the most credible values of the parameters are very close to the sample mean and standard deviation, and the posterior distribution explicitly reveals the uncertainty in the estimate. In particular, notice that the 95% HDI spans a small range relative to the prior distribution. Anything we want to know about the parameters we simply "read off" the posterior distribution.

The posterior distribution provides a *joint* distribution over both parameters simultaneously. There is no need to

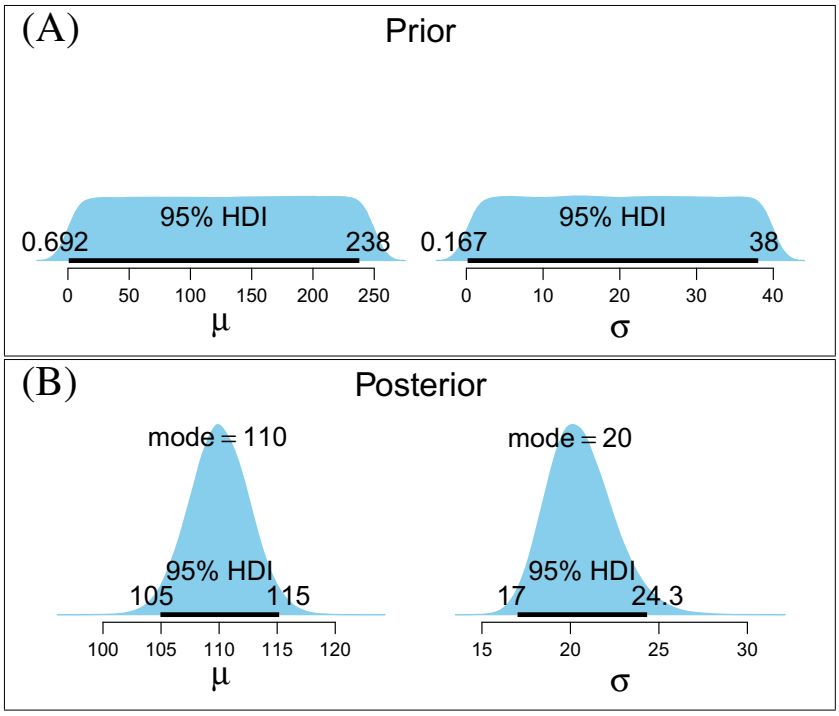

Fig. 3 **A** Prior distribution is broad over possible values of the parameters, μ and σ. The parameters are in a two-dimensional joint space, although only the marginal distributions are shown here. The actually-used distribution was uniform but the distribution shown here has rounded shoulders because of smoothing for display purposes only. **B** After taking into account a given set of data, the posterior distribution over the parameters is much narrower than in the prior. Notice that the range of the abscissa is much smaller in the graphs of the posterior (**B**) than in the graphs of the prior (**A**). The *vertical axis* (unlabelled) is the relative probability density. HDI limits and modes are displayed to first three digits only

generate different sampling distributions of different test statistics to evaluate different parameters, such as a sampling distribution of t to test the mean and a separate sampling distribution of χ^2 to test the standard deviation.

Bayesian analysis applies to *any* parameterized model of data

After seeing a few simple examples such as those illustrated in the previous sections, newcomers often wonder whether Bayesian analysis applies to some other favorite situation or analysis. Can Bayesian methods be used for factor analysis? Yes (e.g., Arminger & Muthén, 1998; Ghosh & Dunson, 2009; Merkle, 2016; Mezzeti, 2012; Song & Lee 2001, 2012). Can Bayesian methods be applied to item response theory? Yes (e.g., http://tinyurl.com/BayesianIRT,[1] Albert, 1992; Azevedo, Andrade, & Fox, 2012; Santos, Azevedo, & Bolfarine, 2013). Can Bayesian methods be used for cluster analysis? Yes (e.g., Guglielmi, Ieva, Paganoni, Ruggeri, & Soriano, 2014; Lau & Green, 2007; Richardson & Green, 1997). Can Bayesian methods be used for time-series analysis? Yes (e.g., Geweke & Whiteman, 2006; McCulloch & Tsay, 1994; Pole, West, & Harrison, 1994). Can Bayesian methods be used for hierarchical conditional-logistic models of nominal data? Yes: (e.g., Liddell & Kruschke, 2014).

In principle, *any* parameterized model of data can have its parameters estimated via Bayesian methods. The beauty of the Bayesian approach is that the posterior distribution can be directly examined for determining what parameter values are most credible and how uncertain the estimate is. We just "read off" the intuitively meaningful answer directly from the posterior distribution. There is no need for additional assumptions and complications in deriving sampling distributions for p values and p-value based confidence intervals.

Bayesian model comparison

In some situations, the researcher may have two or more candidate models of the data, and be interested in evaluating the relative veracities of the models. For example, a scatterplot of $\langle x_i, y_i \rangle$ data might be modeled as a linear trend of x with normally distributed y values, or as an exponential trend with Weibull distributed y values. The models could be "nested" such that one model merely restricts some of the parameter values of the other model, or the models could be non-nested and involve completely different parameters and mathematical structure.

We suppose that model 1 involves parameter θ while model 2 involves parameter ϕ. The key idea is that the model index M is itself another parameter. The model-index parameter M has discrete values, namely $M = 1$ or $M = 2$ (when there are two models, but there could be more). The parameters within the models, θ and ϕ, will in general be continuous but could be discrete. Thus the overall parameter space is a multi-dimensional space involving $M \times \theta \times \phi$. Bayesian inference is re-allocation of credibility across the overall parameter space. The probability distribution shows the relative credibilities of the model index values, and the relative credibilities of parameter values within each model. To judge the relative veracities of the models, we simply "read off" the posterior probabilities of the model indices.

Figure 4 illustrates the parameter space of model comparison and how Bayesian inference re-allocates credibility across that space. The figure caption provides detailed explanation, and we encourage the reader to examine the figure now.

An inherent quality of Bayesian model comparison is that the prior distributions on the parameters "inside" the models (θ and ϕ in our example) can strongly affect the posterior distribution of the model indices. This sensitivity of the model-index posterior to the within-model priors is caused by the fact that the model-index probability reflects the models' abilities across their entire parameter spaces, not

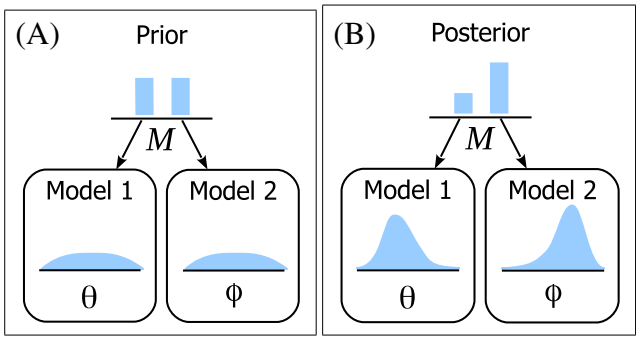

Fig. 4 Bayesian model comparison. In both panels **A** and **B**, model 1 involves a parameter denoted θ, whereas model 2 involves a parameter denoted ϕ. In general, θ and ϕ may be multidimensional and may be continuous or discrete valued. At the *top* of both panels, the discrete model-index parameter is denoted M. The *two bars* indicate the probabilities that $M = 1$ and $M = 2$. The full space of possibilities is the multi-dimensional parameter space involving $M \times \theta \times \phi$. In panel **A**, the prior distribution is shown iconically as broad distributions on parameters θ and ϕ within the two models, and equal probabilities on the discrete values $M = 1$ and $M = 2$. In panel **B**, the posterior distribution shows that credibility has been re-allocated across all the parameters, with $M = 2$ having higher probability than $M = 1$. Bayesian inference has re-allocated credibility across the entire multi-dimensional parameter space $M \times \theta \times \phi$ simultaneously

[1] The full URL is http://doingbayesiandataanalysis.blogspot.com/2015/12/bayesian-item-response-theory-in-jags.html and a PDF version is available at https://osf.io/79ugq/.

exclusively at their best-fitting parameter values. In other words, the prior distributions on the model parameters are an integral part of the meanings of the models. Model 1 asserts a description of the data as a mathematical form with parameter values θ approximately in a particular zone as specified by its prior distribution. Model 2 asserts a description of the data as a different mathematical form with its parameter values ϕ approximately in another particular zone as specified by its prior distribution. If the prior distribution of a model gives high credibility to parameter values that happen to fit the data well, then the posterior probability of that model index will tend to be high. But if the prior distribution of a model dilutes its prior distribution over a wide range of parameter values that do not fit the data well, then the posterior probability of that model index will tend to be low.

Bayesian model comparison automatically takes into account model complexity. This is important because a complex model will always be able to fit data better than a restricted model (nested in the complex model), even when the simpler restricted model actually generated the data. For example, suppose we have a scatter plot of $\langle x_i , y_i \rangle$ data that was actually generated by a linear trend. A quartic polynomial will always be able to fit the data better than a linear trend because the quartic will over-fit random noise in the data. We would like the model comparison method to be able to declare the simpler model as better than the complex model in this sort of situation. Bayesian methods inherently do this. The reason is that complex models involve more parameters than restricted models, and higher-dimensional parameter spaces require the prior distribution to be diluted over a larger space. The diluted prior distribution means that the prior probability is relatively small for any particular parameter value that happens to fit the data. Therefore the posterior probability of a complex model tends to be downweighted by its diluted prior probability distribution. A complex model can win, however, if the data are much better fit by parameter values in the complex model that are not accessible by the restricted model.

In summary, Bayesian model comparison is an excellent method for assessing models because it is both intuitively informative (just read off the posterior probabilities) and automatically takes into account model complexity. There is no need to compute p values by generating sampling distributions of imaginary data from restricted models. Bayesian model comparison does, however, require the analyst to think carefully about the prior distributions on the parameters within the models, and to think carefully about the prior distribution on the model index. One way to keep the prior distributions within the two models on an equal playing field is by informing both models with the same representative previous data. That is, both models are started with a diffuse proto-prior, and both models are updated with the same previous data. The resulting posterior distributions from the previous data act as the prior distributions for the model comparison (Kruschke, 2015, Section 10.6.1, p. 294). For further reading about Bayesian model comparison and setting useful priors within models, see examples in Kary et al. (2016) and in Vanpaemel and Lee (2012). The setting of prior probabilities on the model index is also important but less often considered. Setting the prior probabilities of two models at 50/50 is not an expression of uncertainty but is instead an expression of strong prior knowledge that the models have equal prior probability. Bayesian methods also allow expressing uncertainty in the prior probabilities of the model indices (see, e.g., http://tinyurl.com/ModelProbUncertainty[2]).

Two approaches to assessing null values

In many fields of science, research focuses on magnitudes of phenomena. For example, psychometricians might be interested in where people or questionnaire items fall on scales of abilities, attitudes, or traits. But in other domains, questions might focus on presence versus absence of an effect, without much concern for magnitudes. Is the effect of treatment different from the control group or not? Researchers would like to know whether the estimated underlying effect is credibly different from the null value of zero or chance. In this section, we will consider two Bayesian approaches to assessing null values. We will see that the two approaches correspond to different levels in the model-comparison diagram in Fig. 4.

Intervals on parameter estimates

Framing a theory as merely "any non-null effect" can lead to Meehl's paradox: As sample size increases and therefore as estimation precision increases, it gets easier to *confirm* the theory than to *disconfirm* the theory (Meehl, 1967, 1997). Science should work the other way around, and posit theories that are challenged more severely by more precise data. Meehl's paradox arises whenever there is a non-zero effect, regardless of how small, because with enough data the non-zero effect will be detected and the null value will be rejected, thereby confirming the anything-but-null theory.

A theory should instead be framed such that increased precision of data yields a greater challenge to the theory. A solution was described by Serlin and Lapsley (1985, 1993): Theories should predict a magnitude of effect, with

[2]The full URL is http://doingbayesiandataanalysis.blogspot.com/2015/12/lessons-from-bayesian-disease-diagnosis_27.html and a PDF version is available at https://osf.io/r9zfy/.

a region of practical equivalence (ROPE) around the predicted magnitude. (Serlin and Lapsley called the region around the predicted value a "good-enough belt.") As more data are collected, the estimate becomes more precise with a smaller range of uncertainty. When that range falls outside the ROPE the theory is disconfirmed, and when the range of uncertainty falls entirely within the ROPE then the theory is confirmed for practical purposes. In particular, a null value surrounded by a ROPE can be accepted, not only be rejected. Among frequentists, this approach is used in the method called *equivalence testing* (e.g., Lakens, 2017; Rogers, Howard, & Vessey, 1993; Westlake, 1976, 1981). A related framework is also used in clinical non-inferiority testing (e.g., Lesaffre, 2008; Wiens, 2002). The approach forms an intuitive basis for our first way to assess null values with a Bayesian posterior distribution. For a comparison of decisions made by equivalence testing and HDI with ROPE, see https://tinyurl.com/TOSTvsHDIandROPE.[3]

We set a ROPE that defines values that are equivalent to the null for practical purposes. For example, consider the effect size δ in a normal distribution, which is defined as the mean μ minus the null value M_0 all over the standard deviation σ, that is, $\delta = (\mu - M_0)/\sigma$. The effect size is the parameter that corresponds to Cohen's d (Cohen, 1988). We might declare a ROPE around zero effect size from -0.1 to $+0.1$ because Cohen (1988) established a convention that a "small" effect size is 0.2, and we will set the ROPE limits at half of that value. Thus, this choice of ROPE says that any value of effect size that is less than half of small is practically equivalent to zero.

The ROPE is merely a decision threshold, and its limits are chosen always in the context of current theory and measurement precision. For example, Cohen's (1988) declaration that a "small" effect size is 0.2 was made in the context of what he experienced from typical research in social sciences: "The terms 'small,' 'medium,' and 'large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation." (Cohen, 1988, p. 25) Other fields of study might routinely measure phenomena with smaller effect sizes, or eschew such small effects and focus on phenomena with larger effect sizes. Serlin and Lapsley (1993, p. 211) said, "The width of the [ROPE] depends on the state of the art of the theory and of the best measuring device available. It depends on the state of the art of the theory ... [because] a historical look at one's research program or an examination of a competing research program will help determine how accurately one's theory should predict in order that it be competitive with other theories." By reporting the posterior distribution, readers from different fields and later times can use their own ROPEs to make decisions.

With a ROPE to define a region of values that are practically equivalent to the null value, a decision rule can be stated as follows:

> If the ROPE completely excludes the 95% HDI, then the ROPE'd value is rejected (because none of the 95% most credible values is practically equivalent to the ROPE'd value).
>
> If the ROPE completely includes the 95% HDI, then the ROPE'd value is accepted for practical purposes (because all of the 95% most credible values are practically equivalent to the ROPE'd value).
>
> If the ROPE and 95% HDI only partially overlap, then remain undecided about the ROPE'd value (because some of the 95% most credible values are practically equivalent to the ROPE'd value but some are not).

The decision accepts or rejects only the ROPE'd value, such as the null value, not the entire ROPE'd interval.

Figure 5 illustrates examples of applying this decision rule. Suppose we are interested in the effect of a "smart drug" on IQ scores. We know the general background IQ has a mean of 100 (and a standard deviation of 15). For purposes of illustration, we use a ROPE on effect size from -0.1 to $+0.1$, halving Cohen's (1988) convention for a small effect. The effect size is relative to the null-effect value of $M_0 = 100$. Suppose we collect data from $N = 63$ randomly selected people, and the sample has a mean of 110 and a standard deviation of 20. The resulting posterior distribution of the effect size is shown in panel A of Fig. 5, where it can be seen that the ROPE completely excludes the HDI. In other words, from these data we would reject the candidate value $\mu = 100$ as untenable for describing the sample data.

The posterior distribution on effect size in Fig. 5 is merely a different perspective on exactly the same posterior distribution previously shown for these data in Fig. 3. Recall that the posterior distribution is a *joint* distribution on the multidimensional parameter space. Every point in the space is a combination of μ and σ, which represents a corresponding effect size, $(\mu - 100)/\sigma$. Every point in parameter space also has a corresponding posterior credibility. The posterior distribution of effect sizes is shown in Fig. 5.

From the posterior distribution we can directly say, "the 95% most credible values of effect size fall between 0.23 and 0.75." This statement is analogous to a frequentist confidence interval, but unlike a confidence interval the HDI actually refers to probabilities of parameter values, not to fictitious samples of data from alternative hypotheses. From the posterior distribution we can also make a statement analogous to a frequentist p value: "There is only 0.2% (0.002)

[3]The complete URL is http://doingbayesiandataanalysis.blogspot.com/2017/02/equivalence-testing-two-one-sided-test.html and a PDF version is available at https://osf.io/q686c/.

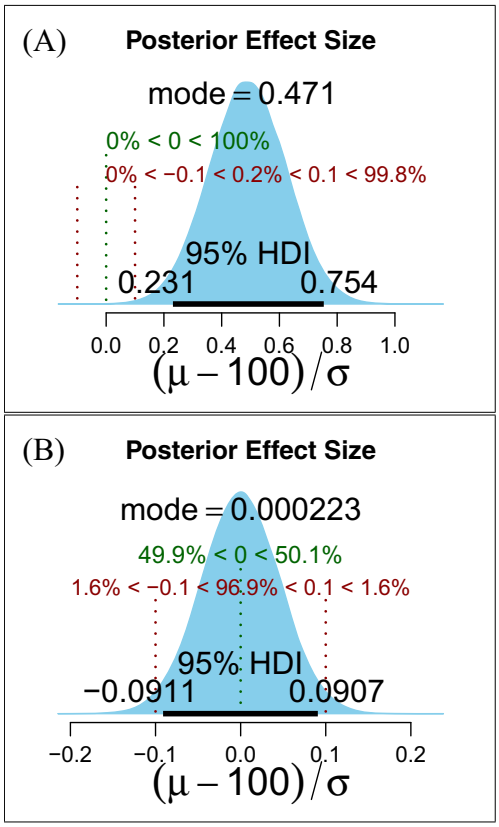

Fig. 5 Posterior distributions on effect size for IQ, $(\mu - 100)/\sigma$, marked with 95% HDI and ROPE. The null value of 0 is marked by a *vertical dotted line*, annotated with the percentage of the posterior distribution that falls below it and above it. The ROPE limits are also marked with *vertical dotted lines* and the percentage of the posterior distribution that falls below, within, and above the ROPE. **A** Posterior distribution of effect size when $N = 63$ with sample mean of 110 and sample standard deviation of 20. This distribution is just a different perspective on the same posterior distribution shown in Fig. 3. Notice that the 95% HDI falls entirely outside the ROPE, and there is only 0.2% probability that the effect size is practically equivalent to zero. **B** Posterior distribution of effect size when $N = 463$ with sample mean of 100 and sample standard deviation of 15. Notice that the 95% HDI falls entirely within the ROPE, and there is 96.9% probability that the effect size is practically equivalent to zero

probability that the effect size is practically equivalent to zero." This statement refers to the probability of an interval around the null value, not to a point null value. Unlike a p value, the statement is about the probability of parameter values, not about the probability of fictitious samples of data from a null hypothesis.

Suppose instead that our data consist of $N = 463$ randomly selected people, and the sample mean is 100 and the sample standard deviation is 15. The resulting posterior distribution on effect size is shown in panel B of Fig. 5. Notice in this case that the 95% HDI falls entirely inside the ROPE, and we would decide to accept the null value for practical purposes. Notice that this conclusion is based on having sufficient precision in the estimate of the effect size so we can safely say that more extreme values for the effect size are not very credible. From the posterior distribution we can make statements about probabilities such as, "the 95% most credible effect sizes fall between −0.091 and 0.091," and, "there is 96.9% probability that the effect size is practically equivalent to zero." These probability statements are carefully phrased so that they do not sound like hypothesis tests. The terminology of "Bayesian hypothesis test" is reserved for a different framing, described in the next section.

Comparing spike prior to alternative prior

A different way of assessing a null value is by expressing the null hypothesis as a particular prior distribution over the parameters and comparing it to an alternative prior distribution (e.g., Edwards et al., 1963; Jeffreys, 1961). For example, the "null" hypothesis that a coin is fair could be expressed by a prior distribution on the coin's bias parameter, θ, that is shaped like an infinitely dense spike at $\theta = 0.5$ with zero height at all other values of θ. The spike prior is compared against an alternative prior distribution that allows a wide range of possible values of θ, such as the one in panel A of Fig. 2. Bayesian inference assesses the relative credibility of the two prior distributions as accounts of the data.

This framing for Bayesian hypothesis testing is really a special case of Bayesian model comparison that was illustrated in Fig. 4. The two models in this case are the spike-prior null hypothesis and the broad-prior alternative hypothesis. Both models involve the same parameters but different prior distributions. Bayesian inference re-allocates credibility across all the parameters simultaneously, including the model-index parameter and the parameters within the models. The posterior probabilities on the model indices indicate the relative credibilities of the null and alternative hypotheses. This framework is illustrated in Fig. 6, which has an extensive explanatory caption that the reader is encouraged to examine now.

In principle, the decision rule about the models would focus on the posterior probabilities of the model indices. For example, we might decide to accept a model if its posterior probability is at least ten times greater than the next most probable model. In practice, the decision rule is often instead based on how much the two model probabilities have shifted, not on where the model probabilities ended up. The degree of shift is called the *Bayes factor*, which is technically defined as the ratio of probabilities of the data as predicted by each model. Put another way, the Bayes factor is a multiplier that gets us from the prior odds ratio to the posterior odds ratio. For example, if the prior odds of the two models are 50/50 and the posterior odds are 91/9, then

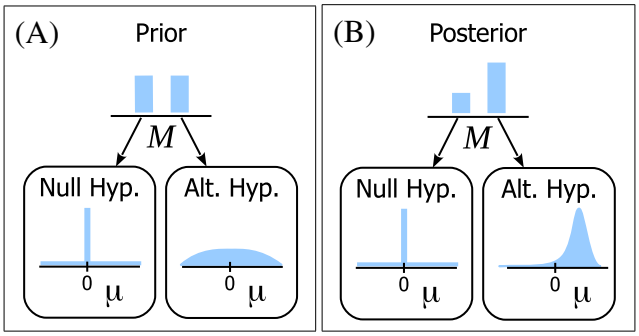

Fig. 6 Bayesian null hypothesis testing. This is a special case of model comparison that was shown in Fig. 4 (please see Fig. 4 for a description of the diagram's components). Both models involve the same parameters but differ in their assumptions about the prior distributions on those parameters. Here, the model comparison focuses on the particular parameter μ, but the models typically involve many other parameters simultaneously. In panel **A**, the prior distribution shows that the null hypothesis ($M = 1$) assumes a "spike" prior distribution on μ such that only the null value has non-zero probability, whereas the alternative hypothesis ($M = 2$) assumes a broad prior distribution on μ. The null value is denoted here generically by the *tic mark* at $\mu = 0$. In panel **B**, the posterior distribution shows that credibility has been re-allocated across the possible parameter values. In this case, the model-index parameter M shows that the alternative hypothesis ($M = 2$) has higher posterior probability, and within the alternative hypothesis the distribution over μ shows that the most credible values of μ are away from the null value. Within both models the other parameters (not shown) have also had their distributions re-allocated, differently in each model

the multiplier that gets from the prior odds to the posterior odds is very nearly 10.1. A Bayes factor of 10.1 also gets from prior odds of 9/91 to posterior odds of 50/50. When the Bayes factor exceeds a critical threshold, say 10, we decide to accept the winning model and reject the losing model. A Bayes factor between 3 and 10 indicates "moderate" or "substantial" evidence for the winning model (Jeffreys, 1961; Kass and Raftery, 1995; Wetzels et al., 2011). A Bayes factor between 10 and 30 indicates "strong" evidence, and a Bayes factor greater than 30 indicates "very strong" evidence. The decision threshold for the Bayes factor is set by practical considerations. Dienes (2016) used a Bayes factor of 3 for making decisions. Schönbrodt, Wagenmakers, Zehetleitner, and Perugini, (2016) recommended a threshold Bayes factor of 6 for incipient stages of research but a higher threshold of 10 for mature confirmatory research (in the specific context of a null hypothesis test for the means of two groups).

As concrete examples, consider the data corresponding to panel A of Fig. 5, which yield a Bayes factor almost 108 to 1 in favor of a particular default-alternative prior relative to the spike-null prior. If the prior odds were 50/50 then the posterior odds would be greater than 99/1 for the alternative hypothesis, but if the prior odds were 1/99 then the posterior odds would be only about 52/48. For the data corresponding to panel B of Fig. 5, a Bayes factor is more than 19 to 1 in favor of the spike-null prior relative to a particular default-alternative prior. If the prior odds were 50/50 then the posterior odds would be 95/5 for the null hypothesis, but if the prior odds were 1/99 then the posterior odds would be only about 16/84 (i.e., favoring the alternative hypothesis). For both examples we used the Bayes factor calculator provided by Rouder, Speckman, Sun, Morey, & Iverson, (2009, at http://pcl.missouri.edu/bf-one-sample) with its default setting for choice of alternative prior (i.e., $r = 0.707$).

There has recently been a flurry of articles promoting Bayes-factor tests of null hypotheses (e.g., Andraszewicz et al., 2014; de Vries, Hartogs, & Morey, 2014; Dienes, 2011, 2014; Jarosz &, Wiley, 2014; Masson, 2011; Morey & Rouder, 2011; Rouder, 2014; Rouder & Morey, 2012; Rouder, Morey, Speckman, & Province, 2012; Wagenmakers, 2007, among many others). Despite the many appealing qualities described in those articles, we urge caution when using model comparison for assessing null hypotheses (and Bayes factors in particular), for the following main reasons:

1. The magnitude and direction of a Bayes factor can change, sometimes dramatically, depending on the choice of alternative prior. To be clear, here we are referring to the shape of the prior distribution within the alternative model, shown as a broad distribution on μ in panel A of Fig. 6. Examples of the sensitivity of the Bayes factor to the alternative-hypothesis prior distribution are provided by Kruschke (2015, Ch. 12, 2011, and 2013, Appendix D), and by many others (e.g., Gallistel, 2009; Kass & Raftery, 1995; Liu & Aitkin, 2008; Sinharay & Stern, 2002; Vanpaemel, 2010). Proponents of the Bayes factor approach to hypothesis testing are well aware of this issue, of course. One way to address this issue is by establishing *default* families of alternative priors (e.g., Rouder & Morey, 2012; Rouder et al., 2012).
2. Default alternative priors often are not representative of theoretically meaningful alternative priors. For example, Kruschke (2011) showed a case of testing extrasensory perception in which the Bayes factor changed direction, from favoring the null to favoring the alternative, when the alternative prior was changed from a default to a distribution based on (idealized) previous results. Therefore, for a hypothesis test to be meaningful, the alternative prior distribution must be meaningful (and the prior on the other parameters in the null hypothesis also must be meaningful). Perhaps the most meaningful alternative prior distribution is one that is

obtainable as a Bayesian posterior distribution from previous data after starting with a diffuse proto-prior (Kruschke, 2015, pp. 294–295, 346). The same procedure can be used to make a meaningful null-hypothesis distribution. A similar approach was taken by Verhagen and Wagenmakers (2014) when applying Bayesian hypothesis testing to replication analysis. By contrast, alternative prior distributions that are centered at zero, or arbitrarily shaped such as half normals or restricted uniform intervals (Dienes, 2014), typically would not be obtained as a Bayesian posterior from previous data and a diffuse proto-prior.

3. The Bayes factor by itself does not use the prior probabilities of the hypotheses, hence does not indicate the relative posterior probabilities of the hypotheses, and therefore can be misleading. To be clear, here we are referring to the prior probabilities of the model indices, shown as the bars on M in panel A of Fig. 6. Consider, for example, the standard introductory example of disease diagnosis. Suppose there is a disease with a diagnostic test that has a hit rate of 95%, meaning that the probability of a positive test result for a diseased person is 95%. Suppose also that the diagnostic test has a false alarm rate of 5%. Lastly, suppose we test a person at random and the result is positive. The Bayes factor of the result is $0.95/0.05 = 19.0$ in favor of having the disease. By considering the Bayes factor alone, we would decide that the patient has the disease. But the Bayes factor ignores the prior probability of having the disease. If the disease were rare, with only 0.1% of the population having it, then the posterior probability of having the disease is only 1.9%. While the posterior probability is 19 times higher than the prior probability of 0.1%, it is still very low, and deciding that the person has the disease if there is only a 1.9% probability that s/he has the disease would be very misleading. Thus, using the Bayes factor to make decisions is dangerous because it ignores the prior probabilities of the models. When applied to null hypothesis testing, if either the null hypothesis or the alternative hypothesis has minuscule prior probability, then an enormous Bayes factor would be required to reverse the prior probabilities of the hypotheses.

   Sometimes it is argued that the Bayes factor is the same as the posterior odds when the prior probabilities of the hypotheses are set to 50/50, and it is reasonable to set the prior probabilities of the hypotheses to 50/50 as an expression of uncertainty about the hypotheses. On the contrary, setting the prior probabilities to 50/50 is not an expression of uncertainty; rather it is a strong assertion of equality, just as strong as setting the prior probabilities to 0.001/0.999 or any other values. If there is uncertainty in the prior probabilities it should be expressed in a more elaborate model structure; see http://tinyurl.com/ModelProbUncertainty.[4]
4. The Bayes factor indicates nothing about the magnitude of the effect or the precision of the estimate of the magnitude. In this way, using a Bayes factor alone is analogous to using a p value alone without a point estimate or confidence interval. The "ritual of mindless statistics" using p values could easily be replaced with a ritual of mindless statistics using default Bayes factors (Gigerenzer, 2004; Gigerenzer & Marewski, 2015). Not considering estimates of effect sizes and uncertainty violates a major emphasis of recent advice in statistical practice: "Always present effect sizes... . Interval estimates should be given for any effect sizes... . (Wilkinson, 1999, p. 599)" The emphasis on reporting and thinking in terms of effect magnitudes and uncertainties continues to this day (e.g., Cumming, 2012, 2014; Fritz, Morris, & Richler, 2012; Lakens, 2013). All the generic reasons offered in those sources to avoid p values and to examine effect magnitudes and precisions are well taken. But we go further and recommend that the goals are better achieved by Bayesian HDI's than by frequentist confidence intervals. Please see the companion article by Kruschke and Liddell (2017) for further discussion.
5. The Bayes factor can accept a null prior even when there is poor precision in the estimate of the magnitude of effect. In other words, the Bayes factor can accept the null prior even when an estimate of the magnitude indicates there is a wide range of credible *non*-null values for the effect. Examples are provided in Kruschke (2015, Ch. 12 and 2013, Appendix D).

In summary, we recommend to use the model-comparison approach to null hypothesis testing only when the null hypothesis has a plausible and quantifiable non-zero prior probability (point 3 above), and with a theoretically meaningful alternative prior distribution not only with a default alternative (point 2 above), and with a check of the sensitivity of the Bayes factor to reasonably different alternative priors when there is uncertainty or dispute about the priors (point 1 above), and with an explicit posterior distribution on the parameter values to examine the magnitude and uncertainty of the estimate (points 4 and 5 above).

Bayesian model comparison in general can be a richly informative procedure. It provides a coherent way to evaluate non-nested (or nested) models, with automatic accounting for model complexity. In general, Bayesian model comparison requires careful consideration of the prior distributions,

[4]The full URL is http://doingbayesiandataanalysis.blogspot.com/2015/12/lessons-from-bayesian-disease-diagnosis_27.html and a PDF version is available at https://osf.io/r9zfy/.

but when the prior distributions are carefully informed the results can be very useful. When Bayesian model comparison is applied to the special case of null hypothesis testing, the same caveats and promises apply. We raised the list of cautions above because Bayesian null hypothesis testing can too easily be done in a ritualized, default fashion that undermines the potentially rich information available in Bayesian model comparison.

Some proponents of the Bayes-factor approach object to the ROPE-with-HDI approach. We believe the objections center on how the term "null hypothesis" is allowed to be expressed mathematically, and on what probability statements are desired. If you believe that the only sensible mathematical expression of the term "null hypothesis" is a spike prior at a point value, then, by definition, the only sensible approach to null hypothesis testing is to compare the spike prior against an alternative prior distribution. Therefore be careful with terminology. When using the ROPE-with-HDI approach, we are considering a null "value" and its ROPE relative to the posterior HDI; we are not considering a null "hypothesis." In the ROPE-with-HDI approach, we can talk about the probability that the parameter (such as effect size) is practically equivalent to the null value, and we can talk about the magnitude and uncertainty of the parameter as revealed by its explicit posterior distribution. In the Bayes-factor approach, we can talk about the predictive probability of the null-hypothesis prior relative to a particular alternative-hypothesis prior, but the Bayes factor by itself does not provide the posterior probabilities of the hypotheses nor does the Bayes factor provide the magnitude of the effect or its uncertainty.

The prior-comparison approach and the ROPE-with-HDI approach are both based on mathematically correct Bayesian inference; the difference is merely in emphasis and interpretation. The two approaches are simply different levels in the unified hierarchical model that was shown in Fig. 6, and as further explained by Kruschke (2011, 2015, Ch's. 10 and 12). (See also the video at http://tinyurl.com/PrecisionIsGoal.[5]) In the hierarchical model of Fig. 6, the HDI-with-ROPE decision rule focuses on the continuous parameter estimate in the alternative model, that is, the posterior distribution on μ in panel B. The HDI-with-ROPE rule makes decisions based on the relations of sub-intervals in that posterior distribution. The Bayes-factor decision rule focuses on the higher-level model index M that points at the spike-null prior versus the alternative prior. Both the continuous parameter μ and the higher-level model index M are parameters with credibilities determined by Bayes rule, in a single Bayesian inference on the integrated hierarchical model. The decision rules focus on different levels of the model. Proponents of either decision rule agree on Bayesian inference as the recommended way to do data analysis.

Bayesian inference gets us from a prior distribution across parameter values to a posterior distribution across parameter values. Making a decision on the basis of the posterior distribution is a separate step. Don't confuse the Bayesian inference with the subsequent decision procedure. Bayesian inference provides the posterior distribution. The posterior distribution encodes the exact allocation of relative credibilities across all the parameter values. The posterior distribution embodies the complete quantification of uncertainty across the whole parameter space. This is the primary product and emphasis of Bayesian inference. Bayesian inference emphasizes quantification of uncertainty, embodied in the posterior distribution over parameters, from which we simply read off whatever information is relevant.

Prior distribution: innocuous, hazardous, beneficial

Newcomers to Bayesian data analysis are sometimes suspicious of using a prior distribution because they have heard rumors that a presumptuous prior can be smuggled into an analysis and thereby yield any desired posterior distribution (satirized by Kievit, 2011). Prior distributions have also been accused of giving unscrupulous analysts extra degrees of freedom for finagling questionable research practices (Simmons et al., 2011). These fears are mostly unfounded. Prior distributions should always be explicitly described and justified in any reported analysis (see essential points of reporting a Bayesian analysis in Ch. 25 of Kruschke, 2015). Priors are usually innocuous, sometimes importantly beneficial, and hazardous only if used carelessly in particular situations, as explained in the following subsections.

Vague priors for continuous parameter estimation are innocuous

Consider estimating continuous parameter values such θ in Fig. 2 or μ and σ in Fig. 3. When the prior is reasonably broad relative to the posterior, then virtually any broad prior will yield nearly the identical posterior distribution. For example, in Fig. 2, if the prior distribution were U-shaped or gently peaked instead of nearly flat, essentially the same posterior distribution would be produced (for an example see Figure 3, p. 305, of Kruschke, 2011). In Fig. 3, if the prior distribution on μ extended from 50 to 150, or instead from 0 to 500, the posterior distribution would be the same. Priors that are broad relative to the posterior distribution are called vague or only mildly informed by the typical scale of the data. We recommend the use of these innocuous broad

[5]The full URL is https://www.youtube.com/playlist?list=PL_mlm7M63Y7j641Y7QJG3TfSxeZMGOsQ4.

priors for typical data analysis in which the focus is on estimation of continuous parameters (Kruschke, 2013, 2015; Kruschke, Aguinis, & Joo, 2012).

Priors in model comparison must be handled carefully

On the other hand, when doing Bayesian model comparison (e.g., for Bayesian null hypothesis testing), the specific choice of prior distributions within models and for the model index is crucial. In model comparison, different choices of prior distributions within models can greatly affect the resulting Bayes factor. Consider for example, the situation of Fig. 3, for which the null hypothesis is $\mu = 100$. The posterior distribution on the effect size was shown in panel A of Fig. 5 and would be essentially unchanged whether the prior on μ was uniform from 50 to 150 or uniform from 0 to 500. But the Bayes factor is highly sensitive to the choice of alternative prior. Consider the Bayes factor of the null hypothesis prior versus an alternative uniform from 0 to 500, and call it BF[0,500]. Consider also the Bayes factor of the null hypothesis prior versus an alternative uniform from 50 to 150, and call it BF[50,150]. It turns out that BF[0,500] is *5 times larger* than BF[50,150]. The reason for the difference is that the prior credibility of μ values near the posterior mode of 110 is 5 times larger in the [50,150] prior than in the [0,500] prior, because the uniform prior must be diluted over a support that is 5 times wider in the [0,500] prior than in the [50,150] prior. In other words, when applied to Bayesian model comparison, and Bayesian null hypothesis testing in particular, different vague priors can yield very different Bayes factors. Many authors provide examples of prior sensitivity in Bayes factors (e.g., Gallistel,2009; Kass & Raftery, 1995; Kruschke, 2011, 2015; Liu & Aitkin, 2008; Sinharay & Stern, 2002; Vanpaemel, 2010).

Unlike in continuous parameter estimation for which any reasonably broad prior is innocuous, priors in Bayesian model comparison must be carefully established, justified, and examined for sensitivity (i.e., checked for whether the Bayes factor changes much when the prior changes reasonably). The sensitivity analysis helps put boundaries on the conclusions when there is uncertainty in choice of prior.

Various authors emphasize that the influence of the prior distribution in Bayesian model comparison is an importantly positive feature because it forces the theorist to acknowledge that the prior distribution on the parameters is a central aspect of the expression of the theory (Vanpaemel, 2010; Vanpaemel & Lee, 2012). For example, a theory that predicts that forgetting should occur within some small range of decay rates is quite different that a theory that predicts the decay rate could be anything. For model comparison, including null hypothesis testing, we recommend that the priors of all models begin with a vague proto-prior that is updated by a modest amount of representative data, to establish the mildly and equally informed priors used in the actual model comparison (for an example see Section 10.6.1, p. 294, of Kruschke, 2015).

Informed priors can be very beneficial

If previous data have established strong prior knowledge that can inform the prior distribution for new data, then that prior knowledge should be expressed in the prior distribution, and it could be a serious blunder not to use it. For example, consider random testing for illicit drug consumption. Suppose the correct-detection rate of the drug test is 95%, and the false-alarm rate is 5%. If we test a person at random and the test result is positive, what is the probability that the person uses the drug? The answer is not necessarily 95%/5% (the Bayes factor), because we must incorporate the prior probability of drug use in the population from which the person was randomly sampled. Suppose the prior probability of drug use is 1%. Then, even with a positive test result, the posterior probability of drug use is only 16.1% (which is considerably higher than the prior of 1%, but still quite low). To interpret the positive drug test only by its hit rate and false alarm rate (i.e., its Bayes factor), and not to incorporate the prior probability, would be a serious error (see http://tinyurl.com/ModelProbUncertainty[6]).

Informed priors can also be useful when trying to make strong inferences from small amounts of data. For illustration, consider a situation in which the Acme Novelty and Magic Company is known to manufacture two kinds of trick coins, one of which comes up heads 99% of the time, and the other of which comes up tails 99% of the time. Suppose moreover that the two types of trick coins are manufactured in equal numbers. That information constitutes strong prior knowledge about the possible values of the underlying probability of heads. Now, we are shown a coin from the company and our goal is to infer which kind it is. We flip the coin only once (i.e., data with $N = 1$) and observe a result of heads. From the single flip we can infer with very high probability that the coin is the head-biased type, because that result would happen extremely rarely from the tail-biased coin.

Prior distributions on parameters can also be useful as expressions of theory. For example, Vanpaemel and Lee (2012) showed how a theoretical notion, that attention should be allocated optimally in categorization (Nosofsky, 1986), could be expressed as a prior distribution over a formal model's attention parameter. Thus, rather than just estimating the parameter from a vague prior, a model that

[6]The full URL is http://doingbayesiandataanalysis.blogspot.com/2015/12/lessons-from-bayesian-disease-diagnosis_27.html and a PDF version is available at https://osf.io/r9zfy/.

assumes optimality can be tested by loading the prior distribution near the assumed value.

One must be careful to apply prior knowledge to new cases appropriately. Consider, for example, studying the effect of a "smart drug" on IQ scores. We know with great certainty from previous administration of many thousands of IQ tests to the general population that the mean IQ is 100 and the standard deviation is 15. Does that prior knowledge imply that the prior distribution for analyzing new smart-drug data should be sharply peaked over $\mu = 100$ and $\sigma = 15$? No, because that previous knowledge applies to *non*-smart-drug users, and we do not have much previous knowledge about IQ scores of smart drug users.

Hierarchical models as a special case of an informed prior

Bayesian methods and software are especially convenient for analyzing hierarchical models (also known as multilevel models). In hierarchical models, the probability structure can be expressed such that the probabilities of values of some parameters depend on the values of other parameters. A typical case is analyzing data from many individuals within a treatment group. There are parameters that describe each individual, and those individual parameters are modeled as coming from a higher-level group distribution, which has its own parameters. The data from each individual inform that individual's parameters, which in turn inform the group-level parameters, which in turn inform all the other individuals' parameters. Thus, all the other individuals in the present data act as "prior" knowledge for estimating any particular individual. The result is that estimates of individuals are rationally shrunken toward the group mode(s), which reduces spurious outlying estimates that would otherwise produce false alarms. For introductions to Bayesian hierarchical models, see Chapter 9 of Kruschke (2015), the chapter by Kruschke and Vanpaemel (2015), or the articles by Rouder and Lu (2005) and by Shiffrin et al. (2008). Vanpaemel (2011) discusses another way to construct informative priors using hierarchical models.

In summary, the prior distribution (a) is innocuous when it is broad relative the posterior in continuous parameter estimation, (b) is crucial and should express meaningful prior knowledge in model comparison such as null hypothesis testing, and (c) is useful when informed by previous data, concurrent data, or theory.

Two shifts of emphasis in data analysis

From frequentist to Bayesian

The mathematics of Bayesian analysis has existed for more than 250 years. The primary formula for Bayesian analysis is called *Bayes' rule*, in honor of the 18th century minister Thomas Bayes (1701–1761) who first wrote down the simple relation between marginal and conditional probabilities (Bayes and Price, 1763). It was the mathematician Pierre-Simon Laplace (1749–1827) who rediscovered Bayes' rule and developed it in sophisticated applications. Some historians argue that this branch of statistical methods ought to be named after Laplace (e.g., Dale, 1999; McGrayne, 2011).

As Bayesian mathematics has been around so long, and arguments in favor of Bayesian analysis have been made repeatedly over many years, why is Bayesian analysis only gaining ascendancy now? There are (at least) three main reasons: philosophical, sociological, and computational. In the early years, many people had philosophical concerns about the status of the prior distribution, thinking that the prior was too nebulous and capricious for serious consideration. But many years of actual use and real-world application have allowed reality to overcome philosophical anxiety. Another concern is that Bayesian methods do not control error rates as indicated by p values (e.g., Mayo & Spanos, 2011). This concern is countered by repeated demonstrations that error rates are extremely difficult to pin down because they are based on sampling and testing intentions (e.g., Kruschke, 2013; Wagenmakers, 2007, and references cited therein). Bayesian methods were also discouraged by sociological forces. One of the most influential statisticians of the 20th century, Ronald Fisher (1890–1962), was a vociferous and relentless critic of Bayesian methods (and of the people who used them). Relatively few people had the courage to investigate Bayesian methods in an era when they were belittled and dismissed. Perhaps the most important reason that Bayesian methods stayed in the background of statistical practice was computational. Until the advent of high-speed computers and associated algorithms, Bayesian methods could only be applied to relatively simple models for which formulas could be analytically derived. However, since the introduction of general-purpose Markov chain Monte Carlo (MCMC) algorithms in the 1990s, including general purpose software such as BUGS (Lunn et al., 2013), JAGS (Plummer, 2003), and Stan (Stan Development Team, 2016), Bayesian methods can be applied to a huge spectrum of complex (or simple) models with seamless facility (for histories of MCMC methods, see Gelfand, 2000; McGrayne, 2011). In summary, the practical results along with the rational coherence of the approach have trumped earlier concerns. The remaining resistance stems from having to displace deeply entrenched and institutionalized practices (e.g., McKee & Miller, 2015).

From hypothesis testing to estimation and uncertainty

Figure 7 shows a shift of emphasis from frequentist to Bayesian as a horizontal arrow across columns of a 2×2

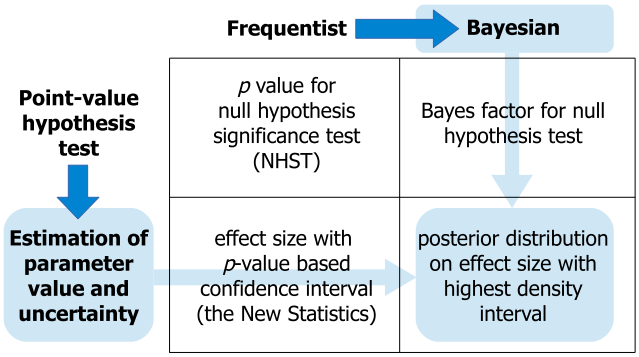

Fig. 7 Two shifts of emphasis in data analysis: From frequentist to Bayesian, marked across columns, and from point-value hypothesis testing to estimation of magnitude and uncertainty, marked across rows. The two emphases converge on the lower-right cell: Bayesian estimation of magnitude and uncertainty

grid. There is a second shift of emphasis in data analysis marked as a vertical arrow across rows in Fig. 7. This is the shift from focusing on null hypothesis testing to focusing on magnitude estimation and quantification of uncertainty. We already described one motivation for eschewing null hypothesis testing because it leads to Meehl's paradox. But there are many other reasons to avoid hypothesis testing as the default or only method of analysis. For decades, there have been repeated warnings in the literature that the artificial "black and white" thinking of null-hypothesis testing too easily leads to misinterpretations of data and biased publication in scientific journals. Recently the American Statistical Association issued warnings about misinterpreting p values (Wasserstein & Lazar, 2016). Gigerenzer (2004) has called NHST a "ritual of mindless statistics" and has also warned that merely replacing p-values with Bayes factors will not help dislodge analysts from ritualized black-and-white thinking (Gigerenzer and Marewski, 2015).

Instead of hypothesis testing, many people have promoted a focus on estimation of effect magnitudes and measures of the uncertainty in those estimates. Within the frequentist framework, this has recently culminated in the so-called New Statistics (Cumming 2012, 2014), as indicated in the lower-left cell of Fig. 7. While we applaud the goals of the new statistics, namely, to focus on magnitudes of effects, measures of uncertainty (such as confidence intervals), and cumulative science such as meta-analysis, we believe that all the goals can be better attained through Bayesian methods. There is not room here to explain the shortcomings of frequentist methods for these goals, and instead we refer the interested reader to other sources and specifically the companion article by Kruschke and Liddell (2017). The convergence of the two shifts of emphasis leads to the lower-right cell of Fig. 7, which is Bayesian estimation of parameters such as effect size and Bayesian quantification of uncertainty in the form of an explicit posterior distribution on credible parameter values. Examples of Bayesian estimation of magnitudes and uncertainty were provided earlier in this article in Figs. 2, 3, and 5.

Model comparison: From frequentist to Bayesian

The table in Fig. 7 assumes we are using a single descriptive model and nested versions of the model that involve restricting parameters to null values. This framework encompasses much of the traditional catalogue of statistical models, including the generalized linear model that spans varieties of regression and analysis of variance. For example, the lower row of the table in Fig. 7 might refer to estimation of parameters in multiple linear regression, while the upper row refers to null hypotheses about the regression coefficients in that model. But the framework of Fig. 7 does not directly embrace the broader structure of model comparison that was illustrated in Fig. 4. The table could be expanded with a third row that refers to model comparison, to suggest that model comparison is a structural generalization of a single traditional model. There are frequentist and Bayesian approaches to model comparison (e.g., Myung & Pitt, 1997).

Some things Bayesian data analysis is not

Bayesian data analysis is very attractive because it intuitively and explicitly reveals the probabilities of parametric descriptions of data, and because the methods are very flexible for complex and useful models. But Bayesian methods are not necessarily a cure for all problems.

Bayesian data analysis is not Bayesian modeling of mind

In the psychological sciences, Bayesian models have been used extensively not for data analysis but for models of cognition and brain function (e.g., Chater & Oaksford, 2008; Chater, Tenenbaum, & Yuille, 2006; Griffiths, Kemp, & Tenenbaum, 2008; Gilet, Diard, & Bessiére, 2011; Jacobs & Kruschke, 2010; Kording, 2014; Kruschke, 2008; Perfors, Tenenbaum, Griffiths, & Xu, 2011). Many newcomers to Bayesian data analysis may have previous exposure to Bayesian models of mind and brain. It would be a mistake to transfer evaluations of Bayesian models of mind/brain to Bayesian models of empirical data. Bayesian methods are the preferred approach to data analysis regardless of the viability of any particular Bayesian model of mind/brain (Kruschke, 2010).

Figure 8 shows different applications of Bayesian methods in psychological and other sciences. In all three panels of Fig. 8, there is a parameterized model that has

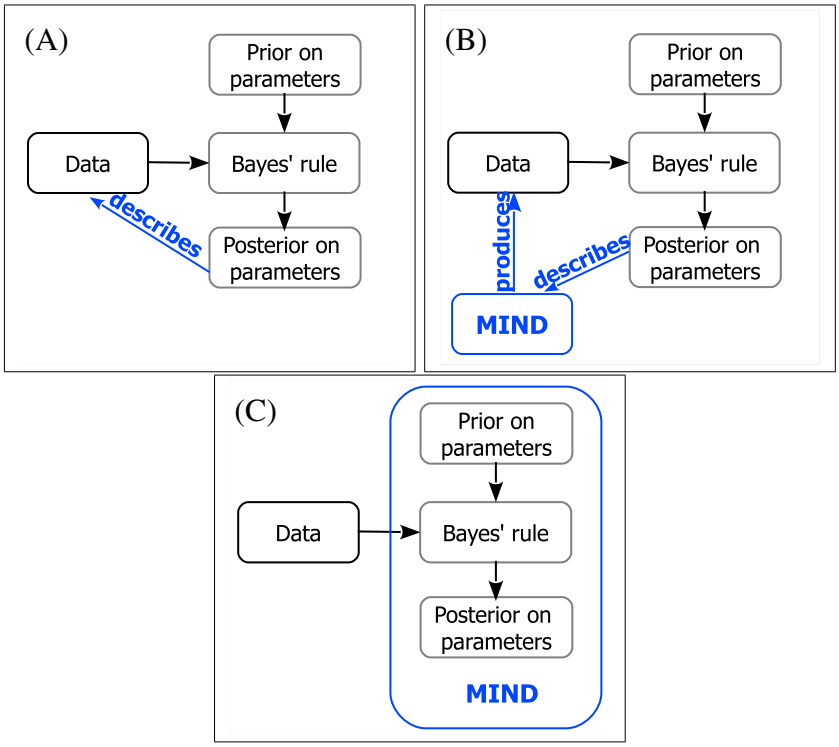

Fig. 8 Three meanings for parameters in Bayesian analysis. **A** Generic Bayesian data analysis: The model parameters meaningfully describe the data, without any necessary reference to the processes that produced the data. **B** Bayesian psychometric analysis: The data are produced by a mind, and the model parameters are supposed to describe aspects of the mind. **C** Bayesian model of mind: The mind itself is modeled as a Bayesian statistician, taking data from the world and updating its parameter values, which represent the knowledge of the mind

probabilities of its parameter values reallocated by Bayesian inference. What differs across panels is the semantic referent of the model and its parameters. (A slight variant of Fig. 8 originally appeared in 2011 at http://tinyurl.com/BayesianModels.[7])

Panel A of Fig. 8 depicts generic data analysis as emphasized in the present article. The model and its parameters describe trends in the data, without any necessary reference to the processes that generated the data. For example, the model parameters could indicate the slope of a linear trend in the data, or the magnitude of a difference between group means. Panel A indicates this property by the arrow labeled "describes" pointing from the posterior to the data.

Panel b of Fig. 8 depicts the case of psychometric models. In this case, the data are known to have been produced by a mind. For example, the data could be IQ scores, or response times, or survey choices. Moreover, the model is intended to describe aspects of the mind, such as scale value for general intelligence, or a diffusion rate in a model of response times, or a regression coefficient in an additive model of covariate influences in choice. Panel B indicates this property with an arrow from the posterior distribution pointing to a "mind" that produced the data. A few recent examples of Bayesian psychometric models include those reported by Fox et al. (2015), Lee and Corter (2011), Oravecz et al. (2014), Vandekerckhove (2014), and Vanpaemel (2009).

Panel C of Fig. 8 depicts the case of Bayesian models of mind. In these models, the mind itself is conceived as a Bayesian statistician, taking data from the world and updating its internal state by using Bayesian inference. The model and its parameters represent the knowledge of the mind, and Bayesian inference represents the processing through which the mind makes inferences, either quickly for perception or relatively slowly for learning. Panel C indicates this by enclosing the Bayesian processing inside a box labeled "mind." Bayesian models of mind assert that the mind is behaving like a Bayesian statistician, using some internal model of the world. There is an infinite family of possible models, so a failure of any one candidate Bayesian model to mimic human behavior does not disconfirm all possible Bayesian models. Moreover, no particular Bayesian model must explain how Bayesian inference is processed at the algorithmic or neural level.

Our central point of this section is that Bayesian models of data analysis (panel A), and Bayesian estimation of

[7]The full URL is http://doingbayesiandataanalysis.blogspot.com/2011/10/bayesian-models-of-mind-psychometric.html and a PDF version is available at https://osf.io/mxyck/.

psychometric models (panel B), are appropriate and useful regardless of whether or not any particular Bayesian model of mind (panel C) is an accurate description of human cognition (Kruschke, 2010). Whether or not you are an enthusiast of Bayesian models of mind, you would be well served to be a Bayesian analyst of data.

Bayesian data analysis is not a panacea

Bayesian data analysis is a coherent, cogent, and intuitive way to reallocate credibility across parameter values in descriptive models of data. Bayesian analysis solves a lot of problems in traditional frequentist analyses involving p values and confidence intervals (e.g., Kruschke, 2013; Ch. 11 of Kruschke, 2015; Kruschke & Liddell, 2017). But because frequentist methods are institutionally entrenched and difficult to displace, Bayesian methods have sometimes been presented with a degree of zeal that might inadvertently give newcomers the false impression that Bayesians claim to solve all problems. Bayesian analysis does not automatically create correct interpretations, just as computing an arithmetic average does not automatically create correct interpretations. In this section we briefly discuss how Bayesian analysis relates to the problems of false alarm rates and bias introduced by selectivity of data.

False alarm rates

Statistical practice in the 20th century was hugely influenced by Ronald A. Fisher, and in particular his work was central for establishing $p < .05$ as the criterion for declaring significance (Fisher, 1925), although Fisher might not have endorsed the ritual of checking p values that is institutionalized in present practice (Gigerenzer et al., 2004). The criterion of $p < .05$ says that we should be willing to tolerate a 5% false alarm rate in decisions to reject the null value. In general, frequentist decision rules are driven by a desire to limit the probability of false alarms. The probability of false alarm (i.e., the p value) is based on the set of all possible test results that might be obtained by sampling fictitious data from a particular null hypothesis in a particular way (such as with fixed sample size or for fixed duration) and examining a particular suite of tests (such as various contrasts among groups). Because of the focus on false alarm rates, frequentist practice is replete with methods for adjusting decision thresholds for different suites of intended tests (e.g., Maxwell & Delaney, 2004, Ch. 5) or for different stopping intentions (e.g., Sagarin, Ambler, & Lee, 2014).

Bayesian decisions are not based on false alarm rates from counterfactual sampling distributions of hypothetical data. Instead, Bayesian decisions are based on the posterior distribution from the actual data. But ignoring false alarms does not eliminate them. After all, false alarms are caused by random conspiracies of rogue data that happen to be unrepresentative of the population from which they were sampled. Nevertheless, Bayesian analysis can help mitigate false alarms. In particular, Bayesian software makes it relatively easy to implement and interpret hierarchical models that can use all the data to shrink outlying parameter estimates and reduce false alarms. In other words, Bayesian analysis reduces false alarms by letting the data and model structure inform the parameter estimates, instead of restricting false alarm rates through some arbitrary declaration of intended tests, post hoc tests, and so on (Gelman et al., 2012).

Biased data: Outliers, censoring, optional stopping, covariate selection, replication crisis

Data can be biased in many ways, including "questionable research practices" such as excluding inconvenient conditions or stopping data collection whenever the desired result is found (John et al., 2012). If the data are biased or unrepresentative of the population that they are supposed to represent, then no analysis can be certain of making correct inferences about the population. Garbage in, garbage out. However, analysts can attempt to model the bias and account for it in the estimation of the other meaningful parameters. Because of the flexibility of creating models in Bayesian computer software, and the direct interpretation of posterior distributions, some types of biased data might be usefully interpreted by Bayesian methods (e.g., Guan & Vandekerckhove, 2016).

Outliers

Most traditional models of metric data assume that the data are normally distributed. If the data have severe outliers relative to a normal distribution, conventional practice is to transform the data or to remove the outliers from the data (e.g., Osborne & Overbay, 2004, and references therein). Transforming data has limitations, and removing data is, by definition, selecting the data to fit the model. If the outlying values are authentic representations of the underlying population, removing them constitutes selective bias and artificially reduces the variance in the data. Instead of removing inconvenient data, Bayesian software makes it easy to use non-normal distributions to model the data. In particular, heavy-tailed distributions are seamlessly used, and outliers are naturally accommodated (e.g., Kruschke, 2013, 2015, and references cited therein).

Censoring

Many studies involve censored data. For example, in a response-time experiment, a stimulus appears and the subject must respond as quickly as possible. There may be trials during which the subject is "on task" but does not respond within the maximum duration that the researcher is willing to wait. The result of the trial is not a specific response time, but the trial does indicate that the response time is at least as large as the maximum allowed duration and therefore the datum should not be omitted from the analysis. Omitting censored data would be artificially biasing the data toward small values. Censored data also frequently occur in survival analysis, for which interest is in how long a patient survives after treatment. Many patients may still be alive at the time the researcher wants to do the analysis, and so the information from surviving subjects is censored. But the still-alive subjects do provide information, even if not specific survival durations, and therefore the censored data should not be omitted from the analysis, again because doing so could be selectively biasing the data. Bayesian software makes it easy to model censored data (e.g., Kruschke, 2015, Ch. 25). Essentially, the censored data values are imputed as if they were estimated parameters, with a constraint imposed by the known data cutoffs.

Optional stopping

Many researchers collect data until the results show a desired significant effect (John et al., 2012; Yu et al., 2014). That is, tests of significance are run as data are being collected, and data collection continues until the sought-after effect is found to be significant or until patience expires. Intuitively, there seems to be nothing wrong with this sort of "optional stopping," because the new data collected after previous data were tested should be utterly uninfluenced by the previous data or by the analysis. But this intuition fails to recognize that stopping will often be falsely triggered by randomly extreme data, and once the data collection has stopped, the termination precludes the collection of subsequent data that would compensate for the unrepresentative extreme data. Thus, collecting data until reaching significance biases the data toward extreme values.

Optional stopping might not be such a problem when the stopping criterion is based on Bayesian statistics. The Bayes factor, for comparing a null-hypothesis prior against an alternative-hypothesis prior, is not affected by the bias in the data introduced by optional stopping. The reason is that the null-hypothesis parameter value and any particular alternative-hypothesis parameter value would generate extreme data at the same relative rate regardless of whether you wait for those extreme data or not (Rouder, 2014). On the other hand, it is still true that an estimate of the magnitude of the effect can be affected by Bayesian optional stopping, although the bias may usually be small. See examples in Kruschke (2015, Ch. 13), Sanborn and Hills (2014), and Schönbrodt et al. (2016). The reason for bias in the estimates is that the termination of data collection could be triggered in small samples by unrepresentative extreme values that aren't yet compensated by values in the opposite direction, while termination of data collection for large samples can be biased by long sequences of unrepresentative modest values in the early trials (Schönbrodt et al., 2016). Thus, if you are interested in the magnitude of the effect, optional stopping may produce somewhat misleading estimates in single studies, even when analyzed by Bayesian methods. The biases should wash out across studies, but the effect of Bayesian optional stopping for meta-analysis across studies is a topic for ongoing research (Schönbrodt et al., 2016).

Bayesian analysis can make it easier to pursue goals other than hypothesis testing. In particular, because the Bayesian posterior distribution directly reveals the precision of the parameter estimate (e.g., by the width of the 95% HDI), Bayesian analysis makes it sensible to collect data until a desired degree of precision has been obtained. For most models, stopping when having achieved a desired precision does not bias the estimate of the parameters. The goal of precision is typical in political polling, for which pollsters sample a number of people sufficient to achieve a desired degree of precision in the estimate. This is also the motivation behind the accuracy-in-parameter-estimation (AIPE) approach to sample size planning (Maxwell et al., 2008; Kelley, 2013). (Here we conflate accuracy and precision merely for brevity.) Bayesian analysis makes the goal of precision particularly appealing because precision is measured by the posterior distribution instead of by frequentist confidence intervals, which suffer various infelicities that will not be recited here. Further discussion and examples can be found in Chapter 13 of Kruschke (2015) (see also the video at http://tinyurl.com/PrecisionIsGoal[8]).

Covariate selection

In multiple regression analysis, sometimes researchers explore many candidate covariates and report only those that appear to be significant. The selective exclusion of some covariates biases the data and can lead to excessive

[8]The full URL is https://www.youtube.com/playlist?list=PL_mlm7M63Y7j641Y7QJG3TfSxeZMGOsQ4.

false alarms. Bayesian analysis does not prevent a crafty person from fishing through a lot of candidate covariates and then reporting only the few that are strongly predictive. But Bayesian analysis does allow the forthright analyst to include all candidate covariates and implement hierarchical structure that shrinks regression coefficients, thereby reducing false alarms, and that simultaneously gives explicit inclusion probabilities of the various candidate regressors (e.g., Ch. 18 Kruschke, 2015, and references cited therein).

The crisis of replication

Some forms of bias in data can be usefully addressed with Bayesian analysis. But Bayesian analysis cannot magically undo all bias in the data that are delivered to it, and Bayesian analysis cannot prevent researchers from selectively biasing the data that get analyzed. In particular, Bayesian analysis by itself cannot solve the recently headlined "replication crisis" across the sciences (e.g., Ledgerwood, 2014; Pashler & Wagenmakers, 2012). The replication crisis is largely a consequence of biased selection of which data get analyzed and which data get published, while Bayesian analysis is primarily a means for analyzing data that have already been selected for analysis.

One important way that Bayesian analysis could help the replication crisis is by giving researchers tools to focus on precision (and accuracy) of parameter estimation as the main goal for data collection and as a key criterion for publication, instead of setting the goal to be rejecting or accepting a null value (as explained in Ch. 13 of Kruschke, 2015, and in the talk at http://tinyurl.com/PrecisionIsGoal[9]).

It is important to note, however, that the main reasons to do Bayesian analysis have little directly to do with solving the replication crisis. Many people have been promoting a transition away from null hypothesis significance testing to Bayesian methods for decades, long before the recent replication crisis made headlines.

Where to learn more

Some of the main attractions of Bayesian analysis are (a) the transparent interpretation of the posterior distribution in terms of most credible parameter values and their uncertainty, without using sampling distributions (i.e., without p values and p-value-based confidence intervals), and (b) the ability in Bayesian software to create flexible and meaningful models that are appropriate for describing the data. We hope you are convinced that Bayesian methods provide rich and intuitive interpretations of data, and that it would be worth your time to learn more.

You can learn more from articles, books, and workshops. Numerous workshops and short courses in Bayesian methods are offered throughout the year and around the world for audiences from various disciplines.

The special issue of this journal, in which the present article appears, has several tutorial articles on Bayesian analysis, including the companion to this article that explains frequentist and Bayesian analyses side by side (Kruschke and Liddell, 2017), along with Bayesian meta-analysis and power analysis. Other introductory articles that focus on parameter estimation and highest density intervals include Kruschke (2013), Kruschke et al. (2012), and Zyphur and Oswald (2015). A simple web app for comparing two groups, which runs in a web browser without needing to download any software, is available at http://www.sumsar.net/best_online/ (created by Rasmus Bååth). The app can be very instructive to use with your own data. You can watch the MCMC representation of the posterior distribution emerge, and then interpret the posterior distribution in terms of meaningful parameter estimates and their uncertainties.

If you are interested in Bayes factors for hypothesis testing (i.e., the upper-right cell in Fig. 7), there are online calculators and downloadable software available. Online Bayes-factor calculators include those by Rouder and colleagues at http://pcl.missouri.edu/bayesfactor and by Dienes at http://tinyurl.com/DienesBayesFactor.[10] Downloadable software for Bayes factors in the R language is available at http://cran.r-project.org/web/packages/BayesFactor/index.html, authored by Morey, Rouder, & Jamil. A package for frequentist and Bayesian analyses with pull-down menus, called JASP, is in development and available at https://jasp-stats.org/. A related nascent system, called Jamovi, invites contributed modules and is available at https://www.jamovi.org/.

Several textbooks are available for learning modern Bayesian methods. A thorough introduction that focuses on concepts and applications with a relatively small amount of mathematics is provided by Kruschke (2015). That book begins with very basic notions and gradually builds up to sophisticated models for real data analysis with an extensive suite of computer programs. The programs and exercise solutions are available at the book's web site, https://sites.google.com/site/doingbayesiandataanalysis/. More mathematically advanced books in applied data analysis, but still relatively accessible, are provided by Gelman et al. (2013), McElreath (2016), and Ntzoufras (2009). An introduction

[9]The full URL is https://www.youtube.com/playlist?list=PL_mlm7M63Y7j641Y7QJG3TfSxeZMGOsQ4.

[10]The full URL is http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm.

directed at cognitive scientists is provided by Lee and Wagenmakers (2014). For an accessible mathematical introduction, see the books by Bolstad (2009, 2016). Scholarly textbooks include those by Gill (2014) and Jackman (2009).

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*(3), 251–269.

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E. J. (2014). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, 1–23.

Arminger, G., & Muthén, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, *63*(3), 271–300.

Azevedo, C. L., Andrade, D. F., & Fox, J. P. (2012). A Bayesian generalized multiple group IRT model with model-fit assessment tools. *Computational Statistics & Data Analysis*, *56*(12), 4399–4412.

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S. *Philosophical Transactions*, *53*, 370–418. doi:10.1098/rstl.1763.0053

Bolstad, W. M. (2009). *Understanding computational Bayesian statistics*. Hoboken, NJ: Wiley.

Bolstad, W. M. (2016). *Introduction to Bayesian statistics*, 3rd edn. Hoboken, NJ: Wiley.

Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for a Bayesian cognitive science*. Oxford, UK: Oxford University Press.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Special issue: Probabilistic models of cognition. *Trends in CognitiveSciences*, *10*(7), 287–344.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd edn. Hillsdale, NJ: Erlbaum.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge, UK: Cambridge University Press.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G. (2014). The new statistics why and how. *Psychological Science*, *25*(1), 7–29.

Dale, A.I. (1999). *A history of inverse probability: From Thomas Bayes to Karl Pearson*, 2nd edn. Springer.

de Vries, R. M., Hartogs, B. M., & Morey, R. D. (2014). A tutorial on computing Bayes factors for single-subject designs. *Behavior Therapy*.

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

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

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. doi:10.1016/j.jmp.2015.10.003

Doyle, A.C. (1890). *The sign of four*. London: Spencer Blackett.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Fox, J. P., van den Berg, S., & Veldkamp, B. (2015). Bayesian psychometric scaling. In P. Irwing, T. Booth, & D. Hughes (Eds.) *Handbook of psychometric testing*. Wiley-Blackwell.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association*, *95*(452), 1300–1304.

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

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211.

Geweke, J., & Whiteman, C. (2006). Bayesian forecasting. *Handbook of Economic Forecasting*, *1*, 3–80.

Ghosh, J., & Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, *18*(2), 306–320.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.) *The sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421–440.

Gilet, E., Diard, J., & Bessière, P. (2011). Bayesian action–perception computational model: interaction of production and recognition of cursive letters. *PloS one*, *6*(6), e20387.

Gill, J. (2014). *Bayesian methods: Asocial and behavioral sciences approach*, 3rd edn. Boca Raton, Florida: CRC Press.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.) *The Cambridge handbook of computational psychology* (pp. 59–100). Cambridge, UK: Cambridge University Press.

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review*, *23*, 74–86.

Guglielmi, A., Ieva, F., Paganoni, A. M., Ruggeri, F., & Soriano, J. (2014). Semiparametric Bayesian models for clustering and classification in the presence of unbalanced in-hospital survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*(1), 25–46.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research*, *7*(1), 1–20.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. New York: Wiley.

Jacobs, R. A., & Kruschke, J. K. (2010). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 8–21.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1). doi:10.7771/1932-6246.1167

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.

Kary, A., Taylor, R., & Donkin, C. (2016). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology*, *72*, 210–219. doi:10.1016/j.jmp.2015.07.002

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

Kelley, K. (2013). Effect size and sample size planning. In T. D. Little (Ed.) *Oxford handbook of quantitative methods, Foundations* (Vol. 1, pp. 206–222). New York: Oxford University Press.

Kievit, R. A. (2011). Bayesians caught smuggling priors into Rotterdam harbor. *Perspectives on Psychological Science*, *6*(3), 313–313.

Kording, K. P. (2014). Bayesian statistics: Relevant for the brain? *Current Opinion in Neurobiology*, *25*, 130–133.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*(3), 210–226.

Kruschke, J.K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300. doi:10.1016/j.tics.2010.05.001

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Kruschke, J.K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi:10.1037/a0029146

Kruschke, J. K. (2015). *Doing Bayesian data analysis, Second Edition: A tutorial with R, JAGS, and Stan*. Burlington, MA: Academic Press/Elsevier.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*, 722–752. doi:10.1177/1094428112457829

Kruschke, J.K., & Liddell, T.M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, doi:10.3758/s13423-016-1221-4

Kruschke, J.K., & Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In J. R. Busemeyer, J. T. Townsend, Z. J. Wang, & A. Eidels (Eds.) *Oxford handbook of computational and mathematical psychology*. Oxford University Press.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, 963.

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, Retrieved from https://osf.io/preprints/psyarxiv/97gpc/

Lau, J. W., & Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, *16*(3), 526–558.

Ledgerwood, A. (2014). Introduction to the special section on advancing our methods and practices. *Perspective on Psychological Science*, *9*(3), 275–277.

Lee, J., & Corter, J. E. (2011). Diagnosis of subtraction bugs using Bayesian networks. *Applied Psychological Measurement*, *35*(1), 27–47.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, England: Cambridge University Press.

Lesaffre, E. (2008). Superiority, equivalence, and non-inferiority trials. *Bulletin of the NYU Hospital for Joint Diseases*, *66*(2), 150–154.

Liddell, T. M., & Kruschke, J. K. (2014). Ostracism and fines in a public goods game with accidental contributions: The importance of punishment type. *Judgment and Decision Making*, *9*(6), 523–547.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, Florida: CRC Press.

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*(3), 679–690.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*, 2nd edn. Mahwah, NJ: Erlbaum.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563.

Mayo, D. G., & Spanos, A. (2011). Error statistics. In P. S. Bandyopadhyay, & M. R. Forster (Eds.) *Handbook of the philosophy of science*. Philosophy of statistics (Vol. 7, pp. 153–198). Elsevier.

McCulloch, R. E., & Tsay, R. S. (1994). Bayesian analysis of autoregressive time series via the Gibbs sampler. *Journal of Time Series Analysis*, *15*(2), 235–250.

McElreath, R. (2016). *Statistical rethinking: Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.

McGrayne, S. B. (2011). *The theory that would not die*. Yale University Press.

McKee, R. A., & Miller, C. C. (2015). Institutionalizing Bayesianism within the organizational sciences: A practical guide featuring comments from eminent scholars. *Journal of Management*, *41*(2), 471–490.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.) *What if there were no significance tests* (pp. 395–425). Mahwah, NJ: Erlbaum.

Merkle, E. C., & Rosseel, Y. (2016). blavaan: Bayesian structural equation models via parameter expansion. arXiv:1511.05604

Mezzetti, M. (2012). Bayesian factor analysis for spatially correlated data: Application to cancer incidence data in Scotland. *Statistical Methods & Applications*, *21*(1), 49–74.

Morey, R. D., Hoekstra, R., Rouder, J. N., & Wagenmakers, E. J. (2015). Continued misinterpretation of confidence intervals: Response to Miller and Ulrich. *Psychonomic Bulletin & Review*, 1–10. doi:10.3758/s13423-015-0955-8

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419.

Mosteller, F., & Wallace, D. L. (1984). *Applied Bayesian and classical inference the case of the Federalist papers*. New York: Springer.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79–95.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, *115*, 39–57.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: Wiley.

Oravecz, Z., Vandekerckhove, J., & Batchelder, W. H. (2014). Bayesian cultural consensus theory. *Field Methods*, *26*(3), 207–222.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, *9*(6). Retrieved from http://pareonline.net/getvn.asp?v=9&n=6

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.

Peng, F., Schuurmans, D., & Wang, S. (2004). Augmenting naive Bayes classifiers with statistical language models. *Information Retrieval*, *7*(3-4), 317–345.

Perfors, A., Tenenbaum, J.B., Griffiths, T.L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing DSC*. (Vienna, Austria. ISSN 1609-395X).

Pole, A., West, M., & Harrison, J. (1994). *Applied Bayesian forecasting and time series analysis*. CRC Press.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(4), 731–792.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*(3), 293–304.

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*(2), 283–300.

Santos, J. R., Azevedo, C. L., & Bolfarine, H. (2013). A multiple group item response theory model with centered skew-normal latent trait distributions under a Bayesian framework. *Journal of Applied Statistics*, *40*(10), 2129–2149.

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2016). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*(1), 73–83.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren, & C. Lewis (Eds.) (pp. 199–228). Hillsdale, NJ: Erlbaum.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, *56*(3), 196–201.

Song, X. Y., & Lee, S. Y. (2001). Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *British Journal of Mathematical and Statistical Psychology*, *54*(2), 237–263.

Song, X. Y., & Lee, S. Y. (2012). A tutorial on the Bayesian approach for analyzing structural equation models. *Journal of Mathematical Psychology*, *56*(3), 135–148.

Stan Development Team (2016). Stan modeling language users guide and reference manual, version 2.14.0. Retrieved from http://mc-stan.org

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.

Vanpaemel, W. (2009). BayesGCM: Software for Bayesian inference with the generalized context model. *Behavior Research Methods*, *41*(4), 1111–1120.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology*, *55*(1), 106–117.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p* values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi:10.1080/00031305.2016.1154108

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, *32*, 741–744.

Westlake, W. J. (1981). Response to bioequivalence testing—a need to rethink. *Biometrics*, *37*, 591–593.

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

Wiens, B. L. (2002). Choosing an equivalence limit for noninferiority or equivalence studies. *Controlled Clinical Trials*, *23*, 2–14.

Wilkinson, L. (1999). The task force on statistical inference statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, *21*(2), 268–282.

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: a user's guide. *Journal of Management*, *41*(2), 390–420.