Five Bayesian Intuitions for the Stopping Rule Principle

Eric–Jan Wagenmakers[1], Quentin F. Gronau[1], and Joachim Vandekerckhove[2]

[1] University of Amsterdam

[2] University of California, Irvine

Correspondence concerning this manuscript should be addressed to:
E.-J. Wagenmakers
University of Amsterdam, Department of Psychology
Nieuwe Achtergracht 129B
1018VZ Amsterdam, The Netherlands
E–mail may be sent to EJ.Wagenmakers@gmail.com.

Abstract

Is it statistically appropriate to monitor evidence for or against a hypothesis as the data accumulate, and stop whenever this evidence is deemed sufficiently compelling? Researchers raised in the tradition of frequentist inference may intuit that such a practice will bias the results and may even lead to "sampling to a foregone conclusion". In contrast, the Bayesian formalism entails that the decision on whether or not to terminate data collection is irrelevant for the assessment of the strength of the evidence. Here we provide five Bayesian intuitions for why the rational updating of beliefs ought not to depend on the decision when to stop data collection, that is, for the Stopping Rule Principle.

> I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right.
>
> Leonard 'Jimmie' Savage, 1962

The Stopping Rule Principle (SRP; Berger & Wolpert, 1988, pp. 74-88) holds that our statistical conclusions ought to be independent from the choice of when to terminate data collection. A direct consequence of the SRP is that "It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience." (Edwards, Lindman, & Savage, 1963, p. 193).

To many researchers—and especially those with a solid background in frequentist statistics—the SRP seems too good to be true. Surely interim peeks at the data induce a

multiple-comparisons problem that needs to be addressed? Surely researchers who wish to demonstrate evidence for their favorite theory use the SRP to mislead themselves and their peers? The impression that the SRP sanctions statistical cheating is engendered by the fact that the standard frequentist p-value crucially depends on the stopping rule. Specifically, if the null hypothesis is true then the p-value will meander randomly on the interval from 0 to 1 as the number of observations increases; consequently, persistent researchers can bide their time and conduct a new analysis after every new batch of data arrives – if the null is true, the random fluctuations of the p-value guarantee that at some point statistical significance will be achieved, for any level of α greater than 0. The practice of monitoring the p-value until it dips below α is known as "sampling to a foregone conclusion" or "optional stopping".

Despite decades of research, the SRP remains the topic of considerable statistical controversy. Part of the reason is that the stakes are so high. After all, if the SRP is accepted, this means that (a) researchers gain substantial freedom in conducting their experiments; (b) the core tenets of frequentist inference are found wanting (as frequentism violates the SRP). On the other hand, if the SRP is rejected this means that (a) researchers are required to state their sampling plan in advance of data collection and adhere to it during data collection; (b) the core tenets of Bayesian inference are found wanting (as Bayesianism implies the SRP[1]).

Over the years, *Psychonomic Bulletin & Review* has featured several papers on the SRP (e.g., Sanborn & Hills, 2014; Wagenmakers, 2007; Wagenmakers et al., 2018; Yu, Sprenger, Thomas, & Dougherty, 2014). Of particular relevance here is the article by Rouder (2014): "Optional stopping: No problem for Bayesians" and the preprint by de Heide and Grünwald (2018): "Why optional stopping is a problem for Bayesians".[2] The disagreement that is evident from these titles should give anybody pause: here are influential statisticians/methodologists—intelligent, mathematically strong, well aware of the literature on the topic—who appear to take opposing viewpoints on a crucial issue that seems simple enough: should the SRP be accepted or rejected?

The Stopping Rule Principle

Berger and Wolpert (1988) mention that "The Stopping Rule Principle was first espoused by Barnard (1947, 1949), whose motivation at the time was essentially a reluctance to allow an experimenter's intentions to affect conclusions drawn from data." (p. 74). Other references of interest include Anscombe (1963); Barnard, Jenkins, and Winsten (1962); Bartholomew (1967); Basu (1975); Berger (1985); Berger and Berry (1988); Bernardo and Smith (1994); Birnbaum (1962); Cornfield (1966b); Edwards et al. (1963); Good (1991); Kadane, Schervish, and Seidenfeld (1996a, 1996b); Kerridge (1963); Lee (2012); Lindley (1957); Pratt (1965); Raiffa and Schlaifer (1961); Royall (2000); and Wagenmakers (2007).

The Bayesian take on the SRP is summarized by O'Hagan and Forster (2004, p. 123):

> "Another notable context in which the stopping rule affects classical methods is sequential inference. (...) There we consider at various stages during an experiment deciding whether to continue the experiment by obtaining more data, or to stop and make an inference or decision based on the data available

[1] But see Steel (2003).

[2] `https://arxiv.org/abs/1708.08278`, version 3.

> up to that point. If the decision is to stop at a point where n observations have been made, then the inference is based upon the posterior distribution of the unknown parameters, based on the n observations, and is exactly the same as would have been obtained if a non-sequential experiment had been conducted, with the intention from the outset having been to take exactly n observations.
>
> However, classical inference based upon the same data would be different if the non-sequential experiment were performed. If a hypothesis test is required, for instance, the sequential experiment results in a lower degree of significance from the same data, because the probability of the first kind of error is inflated by the chance of rejecting the null hypothesis when it is true at some other stage of the sequential experiment. The difference between the classical and Bayesian inference in this context can be quite striking. To a Bayesian it seems *absurd* [italics ours] that classical inference when the experiment has stopped after n observations depends not only on whether a decision was taken at some earlier stage not to stop the experiment then, but also on whether the decision at this stage might have been to continue and defer inference to a later stage."

The intuitive appeal of the SRP can be clarified with concrete examples (see Berger & Wolpert, 1988, for an entertaining collection) and general arguments. Below we discuss five intuitive arguments to support the general conclusion drawn by Rouder (2014).

Intuitions for the Stopping Rule Principle

Intuition 1: Why the Rouder Simulation Works

In order to clarify the SRP to an audience of psychologists, Rouder (2014) argued as follows (pp. 303–304):

> "Posterior odds are the probability of competing hypotheses given data. If updating through Bayes factor is ideal and if the prior odds are accurate, then the posterior odds should be accurate as well. If a replicate experiment yielded a posterior odds of 3.5-to-1 in favor of the null, then we expect that the null was 3.5 times as probable as the alternative to have produced the data. We can check this interpretation with simulations as follows: In repeated simulations, we can select all those replicate experiments that yield the same posterior odds—say, 3.5-to-1 in favor of the null—and tally how many of these selected experiments came from the null truth and how many came from the alternative truth. If the posterior odds are interpretable as claimed, then about 3.5 times as many of these selected experiments should come from the null than from the alternative."

This simulation setup appears compelling, and it also forms the basis of the preprint by de Heide and Grünwald (2018). However, a more careful inspection suggests that this setup contains a distinctly non-Bayesian element. Specifically, the Rouder simulation does not condition on *what is known* (i.e., the data that have been observed) but instead conditions on *the value of the Bayes factor*. When multiple data sets (all but one of which are hypothetical) can produce the same Bayes factor, this could mean that the simulation results are affected by imaginary data sets whose potential for realization depends on the

stopping rule. Hence, Rouder's simulation itself could be thought to violate the SRP, the very principle that it was designed to support.

When we discovered this possible flaw in the Rouder simulations we set out to demonstrate the problem with concrete examples. To our initial surprise, we came up short every time. For instance, we would compare two sampling scenarios; in scenario A, the Bayes factor was monitored until it exceeded either 3 or 1/3. In scenario B, the exact same rule was followed until the 11th observation, after which the evidence threshold at 1/3 was replaced with one at 1/100. We then imagined a data set consisting of 10 observations and a BF in favor of the null just exceeding 3. Clearly, the probability of these data (under $\mathcal{H}_0$ versus $\mathcal{H}_1$) is the same under scenarios A and B; but what about the proportion of Bayes factors coming from $\mathcal{H}_0$ that just exceed 3, taken across all of the hypothetical data that could be observed? We expected this proportion to differ between scenario A and B, but it did not. The intuition for this invariance is as follows. For each hypothetical data set that yields a Bayes factor of 3 in favor of $\mathcal{H}_0$, the data are 3 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$. Changing the sampling plan changes the prevalence of these hypothetical data sets, but as each of them has a Bayes factor of 3, the end result is unaffected: when the same number is averaged, the averaging weights are irrelevant.

In sum, despite its dependence on hypothetical data sets that depend on the sampling plan, the Rouder simulation is nevertheless consistent with the SRP.

Intuition 2: Learners Always Ignore the Stopping Rule

At its core, Bayesian inference is a theory of learning. All organisms learn from experience[3], and this must be done by updating knowledge in light of prediction errors: gross prediction errors necessitate large adjustments in knowledge, whereas small prediction errors require only minor adjustments. In general terms, we then have the following rule for learning from experience:

$$\begin{array}{c}\text{Present uncertainty}\\ \text{about the world}\end{array} = \begin{array}{c}\text{Past uncertainty}\\ \text{about the world}\end{array} \times \begin{array}{c}\text{Predictive}\\ \text{updating factor}\end{array}$$

The principle of learning from experience can be made more precise using Bayes' rule:

$$\underbrace{p(\theta \mid \text{data})}_{\substack{\text{Posterior beliefs}\\ \text{about } \theta}} = \underbrace{p(\theta)}_{\substack{\text{Prior beliefs}\\ \text{about } \theta}} \times \underbrace{\frac{p(\text{data} \mid \theta)}{p(\text{data})}}_{\substack{\text{Predictive}\\ \text{updating factor}}} . \tag{1}$$

Here, the Greek letter θ ('theta') represents some aspect of the world about which we are unsure; depending on context, it can be known as a 'parameter', a 'hypothesis', a 'model', or, in philosophers' jargon, a 'proposition'. The equation shows how our prior beliefs are transformed to posterior beliefs by the predictive updating factor: values of θ that predicted the data better than average receive a boost in plausibility, whereas values of θ that predicted the data worse than average suffer a decline (see also Wagenmakers, Morey, & Lee, 2016 and Figure 1).

[3] Either individually or as a species, through evolution.

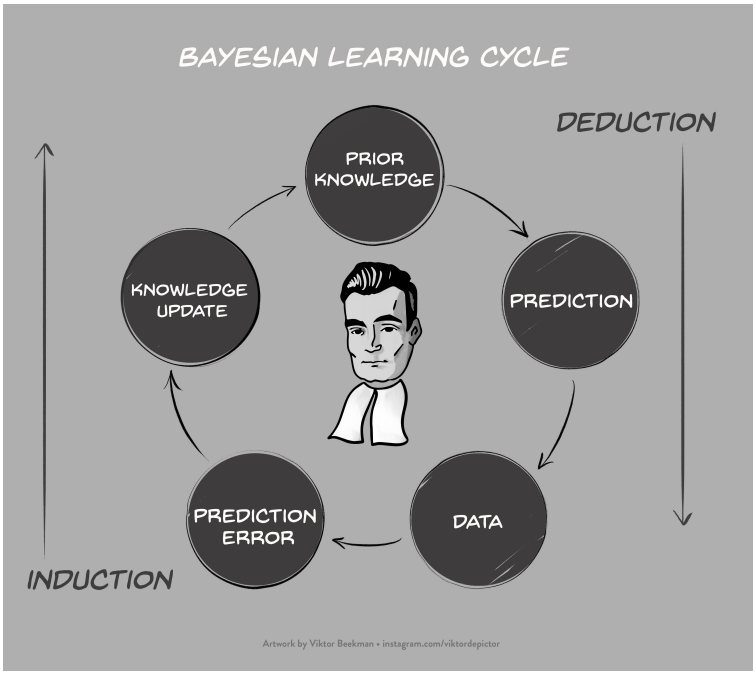

Figure 1. Bayesian learning can be conceptualized as a cyclical process of updating knowledge in response to prediction errors. The prediction step is deductive, and the updating step is inductive. For a detailed account see Jevons (1874/1913, Chapters XI and XII). Figure available at `BayesianSpectacles.org` under a CC-BY license.

For concreteness, let proposition $\mathcal{H}_B$ denote 'the butler murdered the family guest' and proposition $\mathcal{H}_H$ denote 'the housekeeper murdered the family guest'. Assume that we restrict our inference to these two propositions. When we rewrite Bayes' rule in its odds form we have:

$$\underbrace{\frac{p(\mathcal{H}_B \mid \text{data})}{p(\mathcal{H}_H \mid \text{data})}}_{\substack{\text{Posterior odds} \\ \text{for } \mathcal{H}_B \text{ vs. } \mathcal{H}_H}} = \underbrace{\frac{p(\mathcal{H}_B)}{p(\mathcal{H}_H)}}_{\substack{\text{Prior odds} \\ \text{for } \mathcal{H}_B \text{ vs. } \mathcal{H}_H}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_B)}{p(\text{data} \mid \mathcal{H}_H)}}_{\substack{\text{Predictive} \\ \text{updating factor}}}. \tag{2}$$

Suppose we know, from earlier experience in similar cases, that butlers are ten times more likely to murder family guests than housekeepers. Hence the prior odds are 10:1 in favor of the butler being the murderer. We could decide to ignore this information, but one would have to explain why. Regardless of whether one sets the prior odds at 10:1 (to incorporate prior knowledge) or 1:1 (to avoid prejudice), these odds are updated by the relative degree to which the data are compatible with the hypotheses under consideration. For instance, on day 1 of the investigation, the murder weapon is found – it is a heavy candlestick, one that the brawny butler could wield with ease, but the slight housekeeper would find difficult to use with the force required to strike a deadly blow. If the butler is 100 times more likely than the housekeeper to use a heavy candlestick for a murder, this updates the odds to $10 \times 100 = 1000$ in favor of the butler being the murderer. On day 2, it is discovered that the only fingerprints on the candlestick belong to the butler. This

is only modest evidence – the butler has handled the candlestick before, so the presence of those fingerprints is not surprising; the probability of the butler's fingerprints being on the candlestick is only somewhat higher under $\mathcal{H}_B$ ('the butler is the murderer') than under $\mathcal{H}_H$ ('the housekeeper is the murderer'). Let's say the predictive updating factor is 8. Hence the odds after day 1 are adjusted based on the information after day 2 to yield a posterior odds of $1000 \times 8 = 8000$. On day 3, we learn that, at the time of the murder, the butler was in surgery at a nearby hospital, as a result of having been accidentally shot by the family guest during a fox hunt earlier that afternoon. Under hypothesis $\mathcal{H}_B$, the fact that the butler was being operated upon in the hospital during the time of the crime is highly surprising (i.e, $p(\text{butler in hospital at time of the murder} \mid \mathcal{H}_B) \approx 0$), much more surprising than under the hypothesis that the housekeeper committed the murder. In fact, we may learn that the housekeeper and the butler are childhood friends, so that the butler's shooting provides the housekeeper with motive, and this again changes the odds in favor of the hypothesis that the housekeeper is the murderer.[4]

Crucially, at no point during the investigation would a detective take into account the stopping rule in order to adjust his assessment of the evidence. This utter disregard for the stopping rule is not unique to detectives solving murder mysteries; it was also on display, for instance, in Thorndike's cats when they sought to escape his puzzle boxes; it was there in the alphaGo program when it taught itself to play Go; it is present in the spam filters that make email a usable technology; and it is evident in children who learn to speak. For their survival, almost all living creatures need to update their knowledge based on a continual stream of feedback from the environment. No real-life learner has ever given a moment's thought as to how a stopping rule ought to adjust the evidence obtained thus far. The only organisms who seem to care about stopping rules are frequentist statisticians.[5]

Intuition 3: There Can be Only One Posterior and Only One Bayes Factor

The Bayesian process of knowledge updating occurs automatically and yields a single posterior distribution and a single Bayes factor. This holds at any point before, after, and during data collection. Complaints about the result of a Bayesian analysis need to be directed to the elements whose deterministic combination gave rise to it: the prior distribution ($p(\theta)$; e.g., Lindley, 1993), the likelihoods of the various models under consideration ($p(\text{data}|\theta)$; e.g., Etz, 2018), and the data. With the models completely specified, the connection to the data drives a knowledge update that is dictated by the rules of probability theory. Figure 2 tries to convey the impression that the updating process proceeds in a way that is unavoidable; Bayes' theorem "is to the theory of probability what Pythagoras's theorem is to geometry." (Jeffreys, 1931, p. 19; see also Jevons, 1874/1913).

Intuition 4: Evidence Accumulates Towards the Truth

Can a researcher cheat by monitoring the Bayes factor until it indicates sufficiently compelling evidence in favor of the researcher's pet hypothesis? The Bayesian learning

[4] In real murder cases, the learning process will rarely if ever take place with quantitative precision.

[5] There is an exception to this rule. *Informative* stopping rules affect the kernel of the likelihood function (i.e., the part that involves the parameters) and they do affect Bayesian inference. Informative stopping rules are relatively rare; for details see Berger and Wolpert (1988).

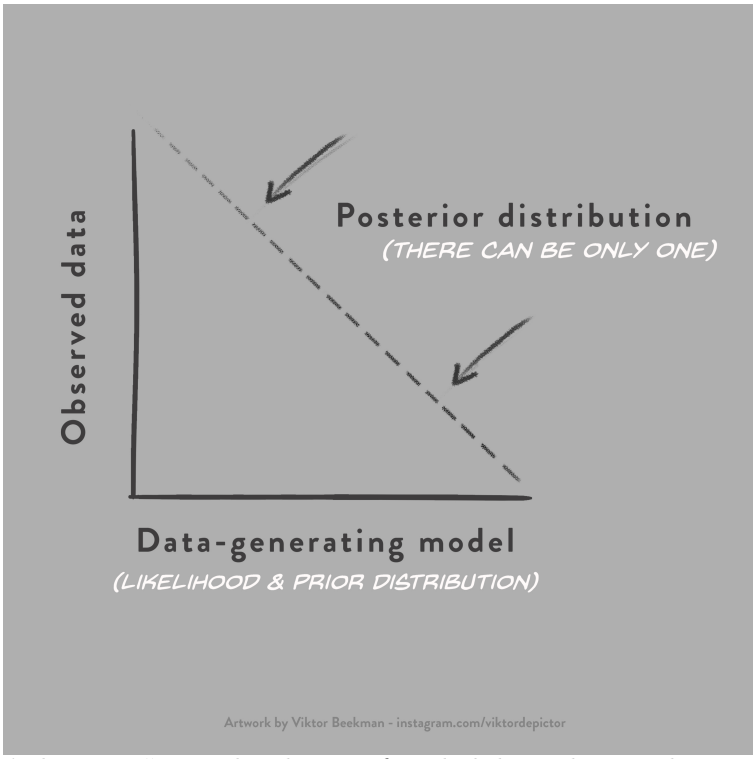

Figure 2. Bayes' theorem "is to the theory of probability what Pythagoras' theorem is to geometry" (Jeffreys, 1931, p. 19). Given the specification of a data-generating process (i.e., the prior distribution $p(\theta)$ and the likelihood $p(\text{data}|\theta)$), the observed data give rise to a single posterior distribution.

cycle shown in Figure 1 already suggests that this is not the case; when we learn about the predictive adequacy of, say $\mathcal{H}_0$ versus $\mathcal{H}_1$, we will discover that one hypothesis does better than the other – collecting more data generally serves to reinforce the correct impression.

Assume that $\mathcal{H}_0$ is true. In such a case, monitoring the p-value is akin to releasing a toy sailboat in a stagnant pond. Over time, random gushes of wind push the sailboat around so that it ends up visiting every position in the pond. Waiting for the sailboat to visit a particular area is therefore a strategy that is certain to succeed and therefore meaningless ("sampling to a foregone conclusion"). In contrast, monitoring the Bayes factor is akin to releasing the toy sailboat in a flowing river. The sailboat will tend to travel downstream, suggesting more and more support for the true $\mathcal{H}_0$; one may decide to wait until the boat ends up traveling upstream in support of $\mathcal{H}_1$, but, instead of resulting in certain success, this strategy is doomed to fail (Edwards et al., 1963).

The intuition about the sailboat can be made precise. It is well known that the sequential monitoring of Bayes factors is subject to a *universal bound* on the frequency of obtaining misleading evidence (e.g., Cornfield, 1966a; Good, 1991; Kerridge, 1963; Royall, 2000; Sanborn & Hills, 2014). This universal bound states that if one of the two hypotheses

under consideration is true[6] and the Bayes factor is monitored until it reaches a level of k in favor of the incorrect hypothesis, the frequency of this happening in repeated use is no more than $1/k$. For instance, in case the null hypothesis $\mathcal{H}_0$ is true and one monitors the Bayes factor BF_{10} until it reaches 20 in favor of the incorrect alternative hypothesis $\mathcal{H}_1$, the frequency of this happening in repeated use is no more than .05. Similarly, in case the alternative hypothesis $\mathcal{H}_1$ is true and one monitors the Bayes factor BF_{01} until it reaches 20 in favor of the incorrect null hypothesis $\mathcal{H}_0$, the frequency of this happening in repeated use is also no more than .05. As summarized by Good (1991, p. 192):[7]

> "*Suppose that a sample of any kind whatever can be continually enlarged and that an experimenter decides that he will continue to enlarge the sample until he obtains a Bayes factor of at least B against a true theory or hypothesis. As soon as he achieves this goal he stops (perhaps pretending that he has to catch a train). Then the probability that he will ever attain his goal is no greater that [sic] $1/B$.*" [italics in original]

Intuition 5: Model Misspecification Can Make Bayes Factors Vulnerable to Optional Stopping

The curse of model misspecification affects all methods of inference, and the Bayes factor is no exception. The Bayes factor compares the predictive performance of two models, say $\mathcal{H}_0$ and $\mathcal{H}_1$. If neither model is true, the Bayes factor will eventually favor the model that is closest to the true model (e.g., Chatterjee, Maitra, & Bhattacharya, in press). Consequently, monitoring the Bayes factor is still akin to releasing the toy sailboat in a flowing river; however, since neither $\mathcal{H}_0$ nor $\mathcal{H}_1$ is true, the sailboat will travel downstream not towards the true model, but towards the one that is closest to it. Therefore, even under

[6] We say that a hypothesis $\mathcal{H}$ is "true" if the data are generated from the distribution $p(\text{data} \mid \mathcal{H})$. Consider the hypothesis that a binomial success probability θ is equal to 0.5, that is, $\mathcal{H} : \theta = 0.5$. In this case, $p(\text{data} \mid \mathcal{H})$ corresponds to a binomial distribution with success probability 0.5 and we say that $\mathcal{H}$ is "true" if the data are generated from this binomial distribution. In case $\mathcal{H}$ features a vector of free parameters θ, we still say that $\mathcal{H}$ is "true" if the data are generated from $p(\text{data} \mid \mathcal{H})$. However, $p(\text{data} \mid \mathcal{H})$ is now obtained by integrating out the parameter vector θ with respect to its prior distribution, that is, $p(\text{data} \mid \mathcal{H}) = \int_\Theta p(\text{data} \mid \theta, \mathcal{H})\, p(\theta \mid \mathcal{H})\, \text{d}\theta$. For instance, consider the hypothesis that does not fix a binomial success probability to a specific value but assigns it a continuous prior distribution $p(\theta \mid \mathcal{H})$. In this scenario, in general, one cannot expect the universal bound to hold in a simulation study where θ is fixed to a particular value θ_0 and data sets are generated repeatedly using only this one value θ_0. The reason is that this procedure does not generate data according to $p(\text{data} \mid \mathcal{H})$. In contrast, the universal bound holds when, in each repetition of the simulation, one (1) generates a value for θ from its prior distribution $p(\theta \mid \mathcal{H})$ and (2) uses this θ-value to generate data from $p(\text{data} \mid \theta, \mathcal{H})$.

[7] According to Cornfield (1966a), the earliest mention is by Edwards et al. (1963, p. 239) who stated that "(...) if you set out to collect data until your posterior probability for a hypothesis which unknown to you is true has been reduced to .01, then 99 times out of 100 you will never make it, no matter how many data you, or your children after you, may collect." However, Barnard already mentions the bound in earlier work; for instance, in a comment on Smith (1953), Barnard (1953) states: "To put it another way, if we interpret the phrase 'more extreme result' to mean 'result giving a smaller likelihood ratio,' then if we obtain, for instance, a likelihood ratio of 1/100, we can say that in rejecting the hypothesis tested on the basis of such a result, or a more extreme one, the odds of error will be less than 1/100. This result will be true regardless of whether or not sampling has been sequential, fixed sample size, or whether we have simply taken what observations we can."

model misspecification, the Bayes factor is in general immune to optional stopping. Nevertheless, there are special cases of model misspecification in which the Bayes factor may behave erratically and become vulnerable to optional stopping.

For example, it has long been known that a one-sided p-value can be given a Bayesian interpretation as an approximate test for directionality (see Marsman & Wagenmakers, 2017). Specifically, the one-sided p-value can be viewed as a Bayes factor test for $\mathcal{H}_- : \delta < 0$ versus $\mathcal{H}_+ : \delta > 0$. But the p-value is affected by optional stopping, and the Rouder simulations suggest that Bayes factors are unaffected by optional stopping. This paradoxical situation is exemplified in the three panels from Figure 3.[8] In each panel, the three different lines represent the result of a two-group comparison for three different simulated data sets (denoted by "1", "2", and "3") created under $\mathcal{H}_0$: a true group difference of exactly 0. The upper panel displays the fluctuations of the right-tailed one-sided p-value of an independent samples t-test as a function of the number of observations n.[9] Because the data were generated under $\mathcal{H}_0$, the one-sided p-value meanders randomly. The middle panel displays the corresponding Bayes factors for directionality, BF_{-+}; just as the one-sided p-value, the Bayes factor for directionality also fluctuates randomly.[10] The lower panel displays the corresponding two-sided Bayes factors, BF_{10}, for testing whether or not an effect is present. As the number of observations increases, the two-sided Bayes factor provides more and more evidence for the true null hypothesis.

In sum, the upper and middle panel of Figure 3 demonstrate that under $\mathcal{H}_0$, both the one-sided p-value and the Bayes factor for directionality will meander randomly; contrary to what we have stated earlier, this Bayes factor allows "sampling to a foregone conclusion". The paradox is resolved by noting that it is critical that the data are assumed to come from the point null hypothesis $\mathcal{H}_0 : \delta = 0$ (see also Kadane et al., 1996a, p. 1234). For the Bayesian test of directionality, this means that neither $\mathcal{H}_-$ nor $\mathcal{H}_+$ is true: the truth is literally in the middle, and our flowing river of evidence has been reduced to a stagnant pond. Consequently, Bayes factors start to drift randomly, just as p-values do.

The interpretation of the Bayes factor is still correct: at any point during data accumulation, there is only one posterior distribution and only one Bayes factor, which informs us about the relative predictive performance of $\mathcal{H}_-$ versus $\mathcal{H}_+$; however, when the data are generated by the point null, researchers can now bide their time and be certain to eventually collect compelling evidence for their favored direction. Of course, when this strategy is followed the posterior distribution will likely show that the effect is very, very small. In contrast, the Bayes factor that tests the null hypothesis $\mathcal{H}_0$ against the alternative $\mathcal{H}_1$ is not misspecified, and the lower panel of Figure 3 shows that for our example trajectories, the evidence increasingly supports the true model $\mathcal{H}_0$.

[8] The corresponding `R` code is available at `https://osf.io/w5kah/`.

[9] Data were generated according to a balanced design (i.e., an equal number of participants in each group), which was ensured by generating observations in alternating fashion.

[10] One-sided p-values higher than .5 are associated with Bayes factors for directionality in favor of the opposing hypothesis; for instance, if a one-sided p-value of .01 maps unto a Bayes factor of $.99/.01 = 99$ in favor of $\mathcal{H}_+$ over $\mathcal{H}_-$, then a one-sided p-value of .99 maps unto a Bayes factor of $.99/.01 = 99$ in favor of $\mathcal{H}_-$ over $\mathcal{H}_+$. This evidential symmetry holds for one-sided p-values, not for two-sided p-values.

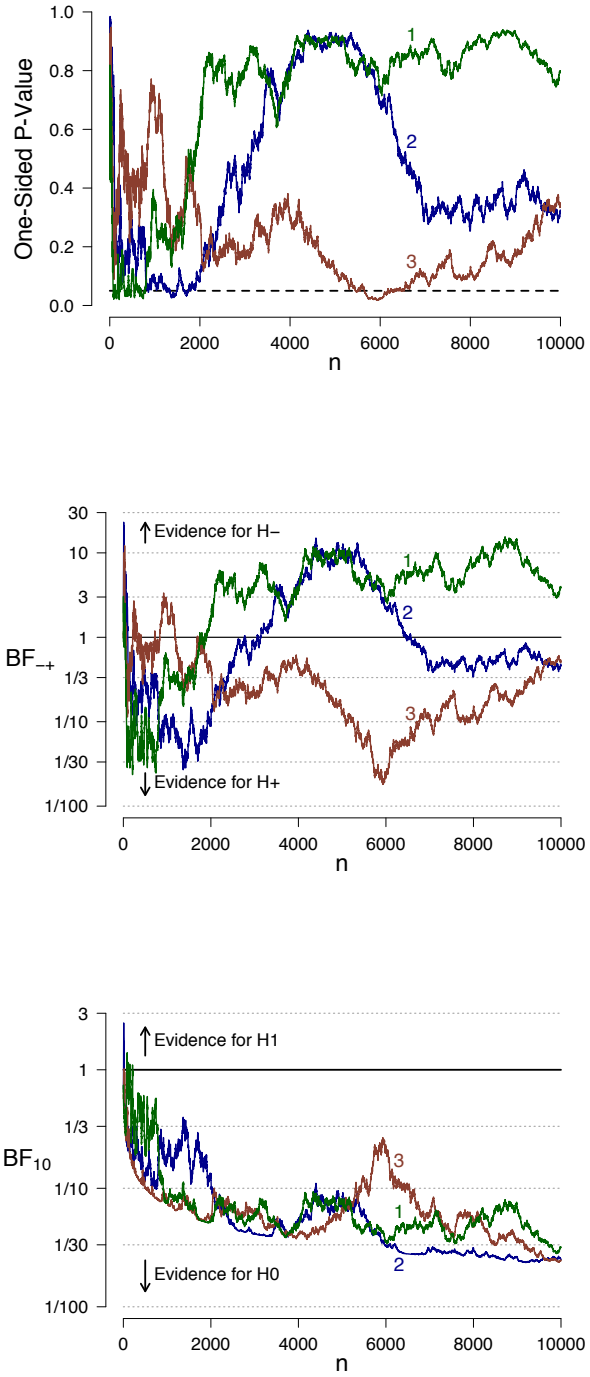

Figure 3. Model misspecification can make a Bayes factor vulnerable to optional stopping. In each panel, the different lines represent a different data set for a two-group comparison, simulated under $\mathcal{H}_0$. The upper panel displays the right-tailed one-sided p-value for an independent samples t-test as a function of sample size n. The middle panel displays the corresponding (misspecified) Bayes factor for directionality, BF_{-+}, and the lower panel displays the two-sided Bayes factor, BF_{10}, for testing whether an effect is present or absent. Figure available at `https://tinyurl.com/yccjy5h9` under CC license `https://creativecommons.org/licenses/by/2.0/`.

Concluding Comments

We have provided intuitions for why Rouder's optional stopping simulation works, for why learners universally ignore the stopping rule, for the inescapable nature of Bayes' rule, for the notion that evidence (but not the p-value) accumulates towards the truth, and that optional stopping might be a concern for Bayesians when there is a specific form of model misspecification. For the sake of brevity we did not discuss how Bayesian inference can be designed to have frequentist guarantees (e.g., Schönbrodt & Wagenmakers, 2018; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017), and how Bayes factors can be interpreted as an accumulation of one-step-ahead prediction errors (e.g., Wagenmakers & Grünwald, 2006).

It may well be that de Heide and Grünwald (2018) agree with most or even all of the points mentioned above. The main argument of de Heide and Grünwald is that optional stopping is a problem for a specific subset of Bayesian analyses, a subset for which the prior conflicts with the Stopping Rule Principle. Hence, for subjective Bayesians, optional stopping is not a problem, and neither should it be a problem for objective Bayesians when the prior is one that a subjective Bayesian could possibly entertain. But objective priors that violate the SRP are potentially problematic, at least from a philosophical perspective – in practice, it may not matter much and the violation of the SRP may only be minor (Berger & Wolpert, 1988). Nevertheless, a violation of the SRP suggests that, for the problem at hand, the search for advisable priors should continue. For an objective Bayesian analysis, where the specification of prior distributions is based on general desiderata, it may happen that adherence to the SRP makes it difficult to fulfill other important desiderata such as scale invariance.

In sum, we welcome the further debate on the importance of stopping rules for Bayesian inference. For now, we conclude that while there exist scenarios in which Bayesian inference is affected by optional stopping policies, these scenarios are relatively uncommon and rely largely on the presence of model misspecification. In any case, the Bayes factor retains its canonical interpretation as *the amount of evidence in the data at hand*. The fact that the Bayes factor depends only on the data and the specification of the competing models, and not on how the data were obtained, is a feature that is present by design: Unavoidable dependence on the stopping rule would all but rule out meta-analysis, the use of naturally occurring data, or even most forms of retrospective analysis including the most trivial case of reading a published study. In that context, we welcome the SRP as well – the practice of statistics would be severely hamstrung without it.

References

Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association*, *58*, 365–383.

Barnard, G. A. (1947). A review of sequential analysis by Abraham Wald. *Journal of the American Statistical Association*, *42*, 658–669.

Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society B*, *11*, 115–149.

Barnard, G. A. (1953). Comment on "the detection of linkage in human genetics" by C. A. B. Smith. *Journal of the Royal Statistical Society B*, *15*, 187.

Barnard, G. A., Jenkins, G. M., & Winsten, C. B. (1962). Likelihood inference and time series. *Journal of the Royal Statistical Society A*, *125*, 321–372.

Bartholomew, D. J. (1967). Hypothesis testing when the sample size is treated as a random variable. *Journal of the Royal Statistical Society. Series B (Methodological)*, *29*, 53–82.

Basu, D. (1975). Statistical information and likelihood. *Sankhya A*, *37*, 1–71.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer.

Berger, J. O., & Berry, D. A. (1988). The relevance of stopping rules in statistical inference. In S. S. Gupta & J. O. Berger (Eds.), *Statistical decision theory and related topics: Vol. 4* (pp. 29–72). New York: Springer Verlag.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)*. Hayward (CA): Institute of Mathematical Statistics.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, *53*, 259–326.

Chatterjee, D., Maitra, T., & Bhattacharya, S. (in press). A short note on almost sure convergence of Bayes factors in the general set-up. *The American Statistician*.

Cornfield, J. (1966a). A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *Journal of the American Statistical Association*, *61*, 577–594.

Cornfield, J. (1966b). Sequential trials, sequential analysis, and the likelihood principle. *The American Statistician*, *20*, 18–23.

de Heide, R., & Grünwald, P. D. (2018). Why optional stopping is a problem for Bayesians. *Manuscript submitted for publication*.

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

Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, *1*, 60–69.

Good, I. J. (1991). C383. A comment concerning optional stopping. *Journal of Statistical Computation and Simulation*, *39*, 191–192.

Jeffreys, H. (1931). *Scientific inference* (1st ed.). Cambridge, UK: Cambridge University Press.

Jevons, W. S. (1874/1913). *The principles of science: A treatise on logic and scientific method*. London: MacMillan.

Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1996a). Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, *91*, 1228–1235.

Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1996b). When several Bayesians agree that there will be no reasoning to a foregone conclusion. *Philosophy of Science*, *63*, S281–S289.

Kerridge, D. (1963). Bounds for the frequency of misleading Bayes inferences. *The Annals of Mathematical Statistics*, *34*, 1109–1110.

Lee, P. M. (2012). *Bayesian statistics: An introduction* (4th ed.). Chichester, UK: Wiley.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Marsman, M., & Wagenmakers, E.-J. (2017). Three insights from a Bayesian interpretation of the one–sided p value. *Educational and Psychological Measurement*, *77*, 529–539.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.)*. London: Arnold.

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society B*, *27*, 169–203.

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory.* Boston: Harvard Business School.

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

Royall, R. (2000). On the probability of observing misleading statistical evidence (with discussion). *Journal of the American Statistical Association*, *95*, 760-?780.

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339.

Smith, C. A. B. (1953). The detection of linkage in human genetics (with discussion). *Journal of the Royal Statistical Society B*, *15*, 153–192.

Steel, D. (2003). A Bayesian way to make stopping rules matter. *Erkenntnis*, *58*, 213–227.

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

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing. *Psychological Science*, *17*, 641–642.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

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