Behavior Research Methods (2022) 54:1131–1147
https://doi.org/10.3758/s13428-021-01618-1

Biasing the input: A yoked-scientist demonstration of the distorting effects of optional stopping on Bayesian inference

Richard B. Anderson [1] · Jennifer C. Crawford [1] · Michael H. Bailey [1]

Accepted: 5 May 2021 / Published online: 7 September 2021

Abstract

Prior work by Michael R. Dougherty and colleagues (Yu et al., 2014) shows that when a scientist monitors the *p* value during data collection and uses a critical *p* as the signal to stop collecting data, the resulting *p* is distorted due to Type I error-rate inflation. They argued similarly that the use of a critical Bayes factor ($BF_{(crit)}$) for stopping distorts the obtained Bayes factor (*BF*), a position that has met with controversy. The present paper clarified that when $BF_{(crit)}$ is used as a stopping criterion, the sample becomes biased in that data consistent with large effects have a greater chance to be included than do other data, thus biasing the input to Bayesian inference. We report simulations of yoked pairs of scientists in which Scientist A uses $BF_{(crit)}$ to optionally stop, while Scientist B, sampling from the same population, stops when A stops. Thus, optional stopping is compared not to a hypothetical in which no stopping occurs, but to a situation in which B stops for reasons unrelated to the characteristics of B's sample. The results indicated that optional stopping biased the input for Bayesian inference. We also simulated the use of effect-size stabilization as a stopping criterion and found no bias in that case.

Keywords Bayesian statistics · Decision-making · Hypothesis testing · Statistical inference

Introduction

There is growing concern about *p* hacking (e.g., Nuijten et al., 2016; Simmons et al., 2011) wherein a researcher perhaps-unintentionally exploits sampling error to find spurious though statistically significant results. In one form of *p* hacking, known as optional stopping, a researcher achieves an artifactually small *p* value by monitoring the sample as it accumulates, and halting data collection when the *p* reaches a critical value (such as .05). It has been suggested that this practice contributes to what has been called a "replication crisis" in which the attempts to replicate published findings fail more often than they should (Pashler & Wagenmakers, 2012; also see Ioannidis, 2005). A variety of potential solutions have been proposed, ranging from the abandonment of *p* values and of classic null-hypothesis testing altogether (Cumming, 2012; Kruschke, 2013; Trafimow, 2014; Trafimow & Marks, 2015), to the adoption of the Bayes factor or other Bayesian measures as a statistical indices (The Psychonomic Society, 2020; Wasserstein & Lazar, 2016), to insisting that researchers compute a priori statistical power (Cohen, 1992) to decide on a sample size prior to starting data collection (The Psychonomic Society, 2020).

While a priori power analysis is a way to avoid optional stopping, its utility is limited. Unless the planned study is an exact replication of prior studies, a power calculation requires the researcher to assume that if the effect size for yet-to-be-conducted study is not zero (i.e., if the null hypothesis is false), then that effect size will have some specific non-zero value. However, because the study has not yet been conducted, the researcher does not know the size of the non-zero effect. The size can be only roughly estimated from prior, similar studies. Thus, as noted by McShane and Böckenholt (2016), a priori power analysis can easily produce a sample-size estimate that is too large or too small. As a response to this problem, McShane & Böckenholt devised a method to minimize bias in the estimate of effect size (also see Maxwell et al., 2008). However, we believe the problem of imprecision remains a fundamental logical and practical problem for a priori power analysis.

In practice, researchers do sometimes engage in optional stopping. In a study that included both behavioral and simulation data, Yu et al. (2014) found that researchers

✉ Richard B. Anderson
randers@bgsu.edu

[1] Department of Psychology, Bowling Green State University, Bowling Green, OH 43402, USA

adopted various stopping rules for data collection, including deciding ahead of time on a sample size, monitoring the *p* value for a critical low value, and monitoring the *p* value for a critical high value (see Yu et al. for details). The study's main purpose was to examine the impact of optional stopping on the validity of statistical inference. A series of simulations assessed the impact of such various stopping rules on outcome measures such as statistical decision errors, Bayes factors, and effect-size estimates. The results indicated a complex pattern of influence that depended on the effect size in the simulated population and on the particular stopping rule. The affected indices included the *p* value as well as the Bayes factor, thus demonstrating that optional stopping can indeed bias statistical decision-making, even in the context of Bayesian inference (see Bayes & Price, 1763).

Yu et al.' (2014) findings did not settle the debate over whether optional stopping is irrational from a Bayesian perspective. In Bayesian inference, decision-makers should stop, or at least pause after taking-in each bit of new data, to update their beliefs about the likelihoods of the relevant hypotheses. Earlier, Wagenmakers et al. (2012) endorsed optional stopping as an appropriate practice within the context of Bayesian statistical analysis (also see Schönbrodt et al., 2017). Later Sanborn and Hills (2014) identified circumstances (such as heterogeneous populations or diffuse statistical hypotheses) in which optional stopping can impact the Bayes factor, but they nevertheless maintained that "the Bayesian interpretation of the evidence does not depend on the stopping rule used and, thus, is correct no matter the reason used to stop the experiment." (p. 284). A similar argument was made by Rouder (2014).

Some thought experiments

Suppose there is a null statistical hypothesis that a country's two dominant ethnic groups do not differ in their mean

Fig. 1 For population effect size (*D*) = 0.0: The natural log of the obtained Bayes factor (log*BF*) as a function of the Bayes factor stopping criterion ($BF_{(crit)}$) and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the log*BF* that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the four levels of the stopping criterion, there were 1000 pairs of simulated experiments (though the criterion was often not met). In each simulated experiment, there was a maximum of 500 simulated participants (250 per group)

support for a particular social policy. The alternative hypothesis is that the two population means differ. A Bayes factor is calculated and properly interpreted to be the likelihood of the data given the alternative hypothesis, relative to the likelihood of the data given the null. However, the researcher in this scenario knows that the sample consists of the researcher's close friends, and thus not likely to be representative of the populations to which the statistical hypotheses pertain. In this situation, the Bayesian algorithm is correct, and yet the statistical inference is invalid because of a flawed input to the algorithm, i.e., an inappropriate sample. The validity problem, in this case, is due to bad data and is unrelated Bayesianism's non-frequentist treatment of probability.

Now consider a second thought experiment that situates optional stopping within the logic of statistical judgment. Suppose Scientist A (call her Adams) is collecting a random sample of data in an experiment to assess whether a particular drug is an effective treatment. Prior to starting the experiment, she had decided to halt data collection when the Bayes factor reaches a critical value of either 50 (which strongly favors the alternative hypothesis) or 1/50 (which favors the null). Suppose further that Scientist B (call her Burns), is independently sampling from the same population, testing the same hypothesis as Adams. However, Burns' stopping rule is to simply stop when Adams stops, thus guaranteeing that Burns and Adams

Table 1 For optional stopping based on $BF_{(crit)}$: obtained sample sizes (N), numbers of samples, proportions of simulated pairs of experiments in which the Scientist A's obtained Bayes factor exceeded those of Scientist B, and hypothesis test for those proportions

Criterion condition ($BF_{(crit)}$)	Obtained sample N			# of Samples	Prop. of Cases where $BF_A > BF_B$ (and $d_A > d_B$) [a]	Binomial test [b] for prop. ($BF_A > BF_B$) = 0.5	
	Median	25th %tile	75th %tile			p	Bayes factor [b]
For Pop. D = 0.0							
Never met	500 [c]	--	--	2818	0.496	.665	0.026
3	40	34	54	999	0.413	< .001	> 1,000
10	496	72.5	498	166	0.367	< .001	34.145
50	62	21.5	151	14	1.000	< .001	> 1,000
200	20	20	93	3	1.000	.250	2.000
For Pop. D = 0.2							
Never met	500 [c]	--	--	2349	0.421	< .001	> 1,000
3	42	34	60	1000	0.453	.003	3.291
10	238	100	356	386	0.912	< .001	> 1,000
50	320	208	429	180	0.978	< .001	> 1,000
200	340	210	434	85	1.000	< .001	> 1,000
For Pop. D = 0.5							
Never met	500 [c]	--	--	60	0.033	< .001	> 1,000
3	44	26	68	1000	0.588	< .001	> 1,000
10	108	58	174	999	0.632	< .001	> 1,000
50	176	108	252	985	0.578	< .001	> 1,000
200	206	134	297	956	0.591	< .001	> 1,000
For Pop. D = 0.8							
Never met	--	--	--	0	--		
3	26	14	44	1000	0.617	< .001	> 1,000
10	46	30	72	1000	0.603	< .001	> 1,000
50	78	51.5	108	1000	0.570	< .001	730.362
200	90	62	124	1000	0.590	< .001	> 1,000

Note. For each level of population effect size (D), and for each level of the criterion, there were 1000 pairs of simulated experiments (though the stopping criterion remained unmet in some of the experiments). In each simulated experiment, there was a maximum of 500 simulated participants (250 per group).

[a] The results indicated a perfect correspondence between whether Scientist A's obtained Bayes factor (BF_A) exceeded that of Scientist B (BF_B), and whether Scientist A's obtained Cohen's d (d_A) exceeded that of Scientist B (d_B).

[b] The values in these columns evaluate the simulation results and not the individual, simulated experiments.

[c] When $BF_{(crit)}$ was never met, the simulated sample necessarily reached its maximum size of 500 (250 per simulated group of participants).

have the same sample size. The Bayes factor for Adams' has crossed the "50" threshold, and so she stops. The sample size happens to be about 100.

Based on past experience, Adams knows that her use of a critical Bayes factor ($BF_{(crit)}$) as a stopping-criterion has likely made her effect size greater than the population effect size (some would call this a statistical bias), and higher than the effect size in Burns' sample, since Burns did not use $BF_{(crit)}$ as a stopping criterion. Moreover, because the sample's effect size determines the Bayes factor (all else being equal), Adams knows that her biased effect size has likely caused the Bayes factor to be higher for her sample than for Burns' sample.

Adams thinks she knows how to debias her results: Simply throw them away and substitute Burns' obtained effect size along with Burns's Bayes factor, calculated from a sample that has not been biased by $BF_{(crit)}$ optional stopping (or equivalently, simply substitute Burns' sample for her own and then re-calculate BF). Put differently, like the researcher in Thought Experiment 1 who sampled close friends rather than obtaining a representative sample, Adams does not think there is a problem related to the Bayesian algorithm or its non-frequentist character. Rather, she believes optional stopping has the effect of presenting a known-to-be-bad (or at least, likely-to-be-bad) sample to a Bayesian algorithm that is valid but that has no special ability to compensate for bad input.

Overview of the present studies

The present paper addresses the aforementioned issues by means of simulations of yoked-pairs of scientists. Each of the two scientists randomly samples from the same population, but only one—Scientist A—engages in optional stopping based on the sample characteristics. The other Scientist, B, simply stops when A stops. Consequently, Scientist B's random samples are representative of the population. If the characteristics of A's same-sized samples were to deviate systematically from

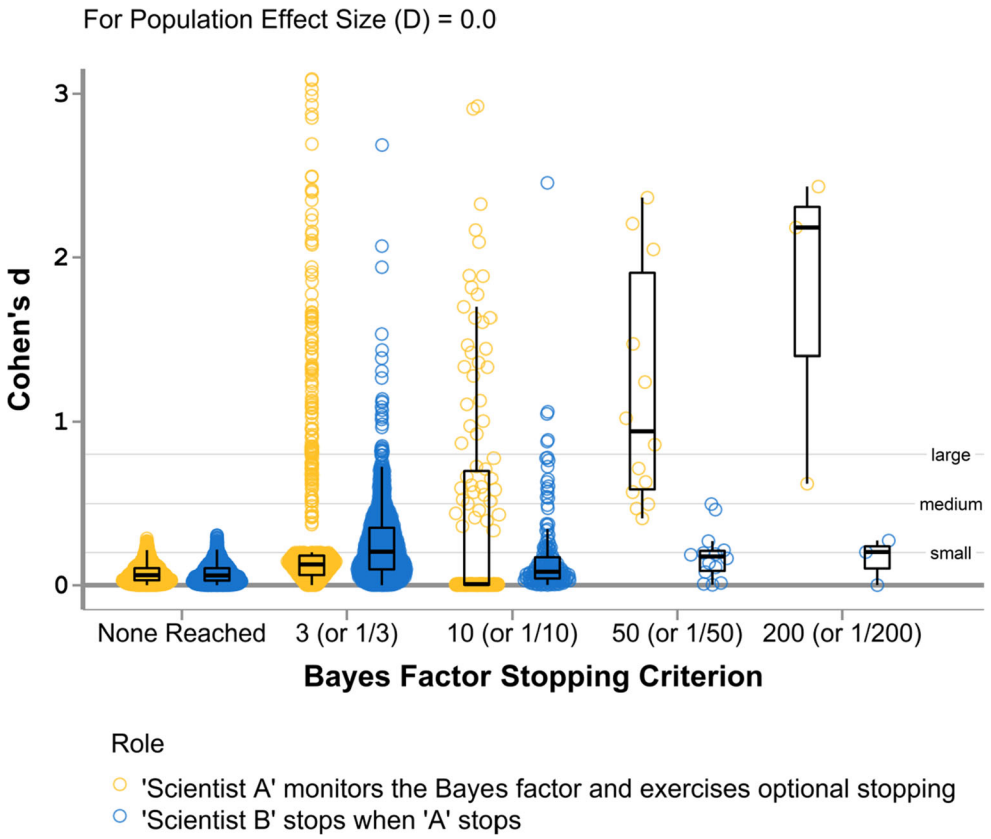

Fig. 2 For population effect size (D) = 0.0: The obtained effect size, Cohen's d, as a function of the Bayes factor stopping criterion ($BF_{(crit)}$) and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the logBF that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the four levels of the stopping criterion, there were 1000 pairs of simulated experiments (though the criterion was often not met). In each simulated experiment, there was a maximum of 500 simulated participants (250 per group)

B's (note that the yoking guarantees that A's and B's samples are the same size), such deviation would indicate that Scientist A's optional stopping has produced non-representative samples that capitalize on chance and are thus ill-suited to the task of drawing inferences about populations.

Study 1: Simulations of Bayes-factor hacking

The present study employed a simulation methodology in an attempt to clarify the nature of the distorting effect of optional stopping. We employ what we call a yoked-scientist design. Scientist A collects data incrementally, monitors the Bayes factor while the data accumulate, and stops when the Bayes factor reaches some pre-established criterion ($BF_{(crit)}$)—for example, when the Bayes factor reaches 3 or 1/3. Scientist B is yoked to Scientist A in that B randomly samples data from the same population (and at the same rate) as does A. However, whereas A monitors the Bayes factor calculated from A's own sample, and uses $BF_{(crit)}$ as a stopping criterion, B's stopping rule is simply to stop when A stops. Moreover, A has no knowledge of the characteristics of B's sample (other than the size). Thus, the two scientists obtain identical sample sizes (though the other aspects of their samples are non-identical), but only Scientist A uses the characteristics of a sample (A's own sample) to decide when to stop. Thus, the methodological contrast is not one of optional stopping versus a hypothetical situation in which data collection continues further. Instead, the contrast is that of Scientist A using a sample-derived statistic (A's Bayes factor) to apply a stopping rule, versus Scientist B using a different stopping rule ("stop when A stops") that does not depend on the characteristics of B's sample.

We reasoned that optional stopping would often allow initial, extreme data to trigger the shutting-out of subsequent,

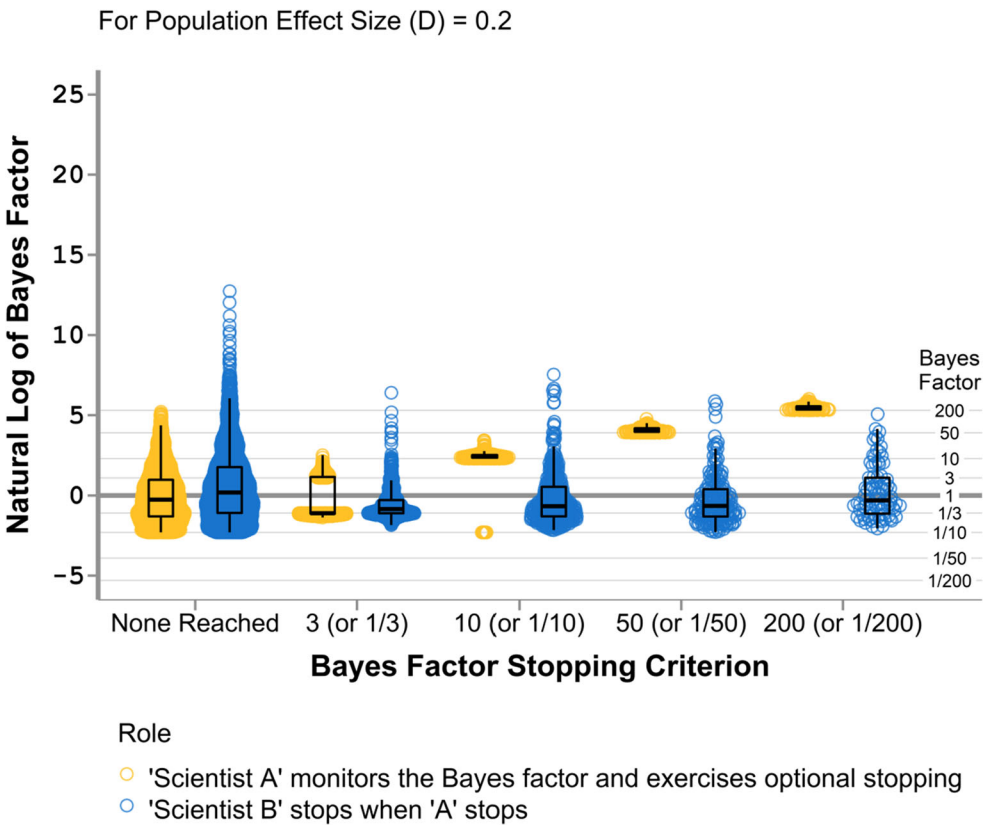

Fig. 3 For population effect size (D) = 0.2: The natural log of the obtained Bayes factor (logBF) as a function of the Bayes factor stopping criterion ($BF_{(crit)}$) and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the logBF that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the four levels of the stopping criterion, there were 1000 pairs of simulated experiments (though the criterion was often not met). In each simulated experiment, there was a maximum of 500 simulated participants (250 per group)

less-extreme data. A consequence would be that some samples of size *n*—specifically, those consistent with an extreme Bayes factor—are more likely to have been obtained than are other samples of size *n*. Thus, Scientist A's sample (arrived at via optional stopping) would end up being a non-representative sample, and would therefore differ systematically from the unbiased though same-sized sample obtained by Scientist B, to whom Scientist A is yoked.

Method

The simulations were conducted as follows. On each of many occasions, Scientist A collected data in a two-group experiment, sampling from two normally distributed populations. For each simulated experiment, there was a maximum of 250 subjects per group ($N = 500$).

Scientist A stopped collecting data at pre-established, critical values of the Bayes factor. Each stopping criterion was two-sided: Data collection could stop when the Bayes factor reached 3 or 1/3, 10 or 1/10, 50 or 1/50, or 200 or 1/200. The scientist conducted 1000 experiments under each of the four stopping-criterion conditions. For each experiment Scientist A conducted, there was another, concurrent experiment conducted by Scientist B. Scientist B's experiment was always identical to A's, sampling from the same populations as A, but Scientist B was yoked to Scientist A in that B stopped collecting data when A did. Thus, A and B always ended up with samples of the same size. If an experiment's sample size reached 500 without the criterion having ever been met, data collection stopped anyway.

The entire procedure was repeated four times, once for each of the four population effect sizes (Cohen's $d = 0.0$, 0.2, 0.5, and 0.8). Thus, with the stopping-criterion variable also having four levels, and with 1000 pairs of experiments per level, there were a total of 32,000 simulated experiments ($4 \times 4 \times 1000 \times 2$).

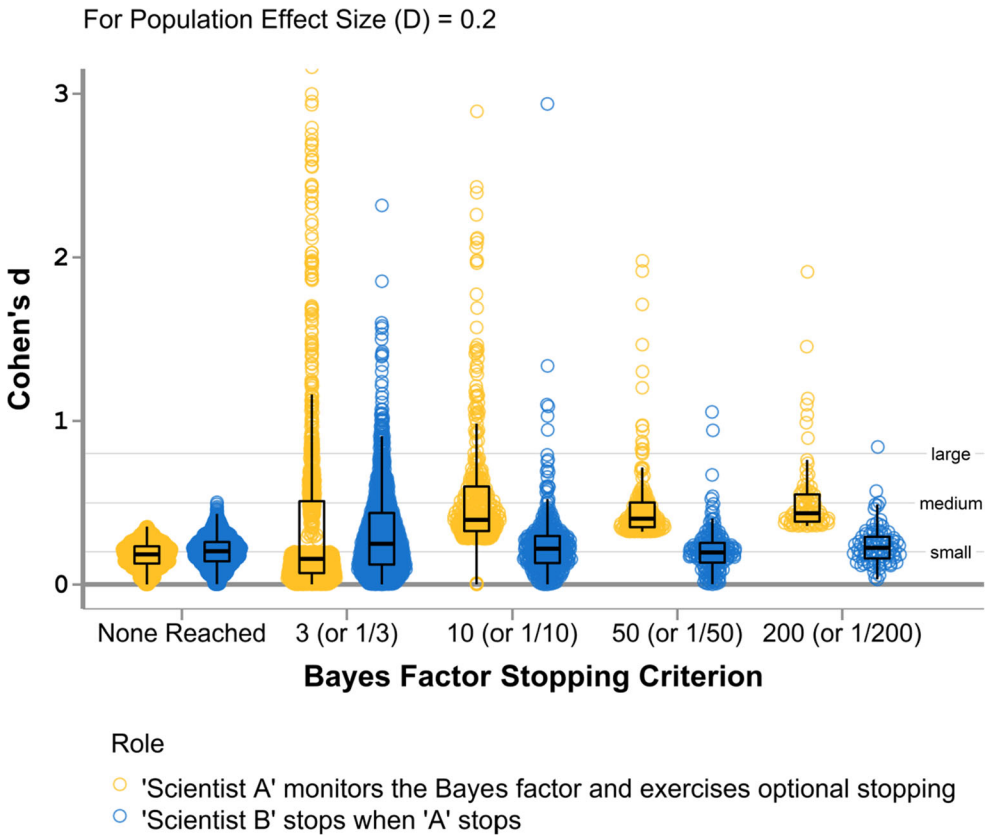

Fig. 4 For population effect size (D) = 0.2: The obtained effect size, Cohen's d, as a function of the Bayes factor stopping criterion ($BF_{(crit)}$) and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the logBF that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the four levels of the stopping criterion, there were 1000 pairs of simulated experiments (though the criterion was often not met). In each simulated experiment, there was a maximum of 500 simulated participants (250 per group)

We used the R package, *BayesFactor* (Morey & Rouder, 2018) to compute the Bayes factors. See Anderson et al. (2021) for the R code for the simulations.

Results

Throughout, we use the terms "Bayes factor" (*BF*) and "log Bayes factor" (log*BF*) to refer to the strength of the evidence for the alternative hypothesis relative to the null hypothesis. $BF_{(crit)}$ refers to the critical value of *BF* that Scientist A uses to determine when to stop collecting data. We plot log*BF* rather than simply *BF* because, for the former, a given magnitude of evidence—either favoring the alternative hypothesis or against the alternative hypothesis—extends the same distance above as below the neutral point. For example, if $BF = 3$, then the strength of the evidence favoring the alternative hypothesis is three times the strength of the evidence favoring the null hypothesis (and indeed, if the prior probabilities are equal, the likelihood of the alternative hypothesis is three times the likelihood of the null hypothesis). However, if the reverse is true, if the strength of the evidence favoring the null hypothesis is three times the strength of the evidence favoring the alternative hypothesis, then *BF* is only 1/3. Thus, in *BF* units, the deviations from the neutral point (1.0) are unequal (3.0 versus 0.333) , but with log*BF*, the values are 1.099 and –1.099, respectively, constituting equal absolute deviations (1.099 and 1.099) from the neutral point of 0.0.

Figure 1 shows that optional stopping distorted the distribution of the log*BF*. Specifically, with the criteria "3 or 1/3" or "10 or 1/10," optional stopping produced bimodality (the data points within a condition are clustered into two subsets on the vertical axis), whereas with more extreme criteria, log*BF* was higher with optional stopping than without. The inflation was further analyzed by computing the proportions of times the obtained log*B*F was greater for Scientist A than for Scientist B, and then performing binomial tests to assess the significance of each proportion's deviation from 0.5. Table 1 shows that when $BF_{(crit)}$ was met, the deviation was significant, and

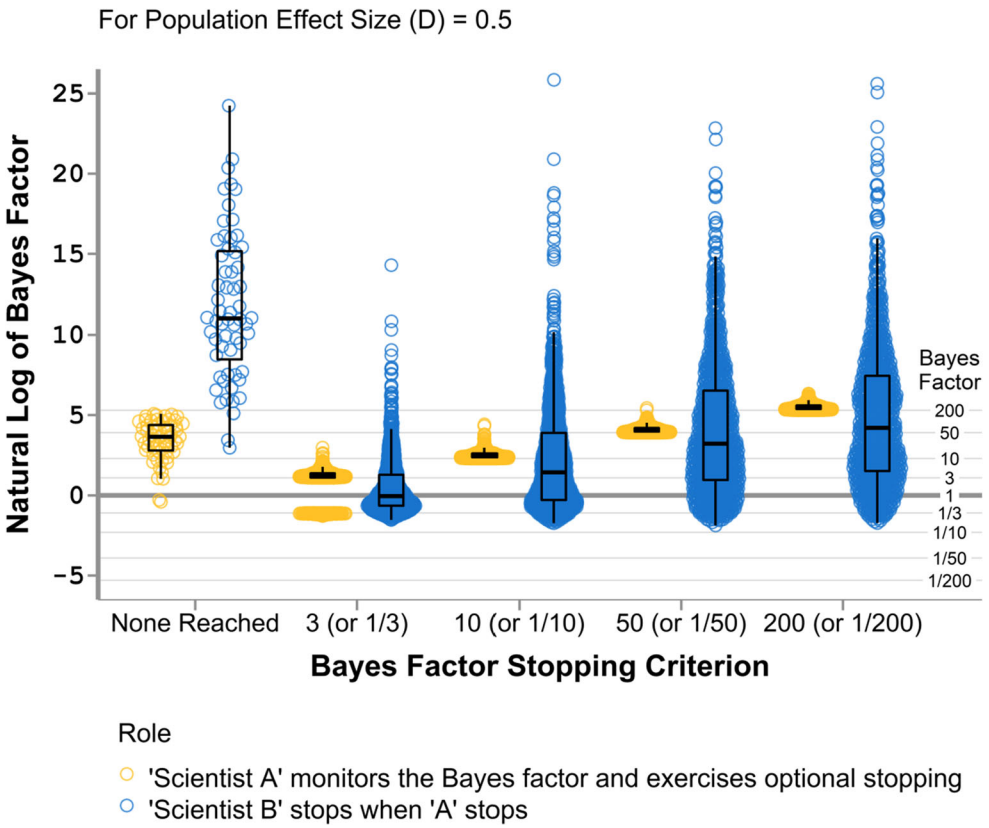

Fig. 5 For population effect size (*D*) = 0.5: The natural log of the obtained Bayes factor (log*BF*) obtained in 1000 pairs of simulated experiments as a function of the Bayes factor stopping criterion ($BF_{(crit)}$) and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the log*BF* that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the four levels of the stopping criterion, there were 1000 pairs of simulated experiments. In each simulated experiment, there was a maximum of 500 simulated participants (250 per group)

the Bayes factor was substantial (the Bayes factor for the analysis of the simulation results—not for a particular, simulated experiment) in all conditions except the "200 (or 1/200)" condition. For the latter condition, the lack of significance is not surprising since that condition contained only three data points, that is, only three pairs of experiments that met the "200 (or 1/200)" stopping criterion.

To examine the results further, Fig. 2 plots not log*BF* but the effect sizes (Cohen's *d*), revealing a distortion similar to that already described for log*BF*: For cases in which the stopping criterion was met, the distribution of the effect sizes was bimodal when the stopping criterion was "3 (or 1/3)" or "10 (or 1/10)," and was inflated (i.e., the effect sizes tended to be greater for Scientist A than for Scientist B) when the stopping criterion was more stringent. Notably, if Scientist A's obtained Bayes factor (or log Bayes factor) exceeded that of Scientist B, then without exception Scientist A's effect size also exceeded Scientist B's. Consequently, the distorting impact of optional stopping on the simulated scientists' obtained Bayes factors is explainable by the distortion of the scientists' obtained effect sizes.

Thus far, we have focused on conditions in which the null hypothesis was true—that is, in which the effect size in the population was 0.0. Figures 3, 4, 5, 6, 7, and 8 show the simulation results when the null hypothesis was false—in particular, when the population effect size was 0.2, 0.5, or 0.8. Those figures show a pattern of results similar to that shown in the previous figures. Less stringent levels of the stopping criterion $BF_{(crit)}$ yielded bimodal distributions of the Bayes factors and of the effect sizes, whereas more stringent levels produced inflation of the Bayes factors and of the effect size estimates (see Table 1 for accompanying descriptive statistics and for the binomial test results).

Note, however, that when the population effect size was 0.5 and when the stopping criterion was never met (Figs. 5 and 6), Scientist A's Bayes factors and effect sizes were lower than those

Fig. 6 For population effect size (*D*) = 0.5: The obtained effect size, Cohen's *d*, as a function of the Bayes factor stopping criterion ($BF_{(crit)}$) and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the log*BF* that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the four levels of the stopping criterion, there were 1000 pairs of simulated experiments (though the criterion was often not met). In each simulated experiment, there was a maximum of 500 simulated participants (250 per group)

of Scientist B—not higher. This is perhaps explainable as a consequence of a relationship between random sampling error and failure to reach the stopping criterion. When Scientist A fails to reach the stopping criterion, it may be because the effect size in the sample is small relative to the population effect size. Indeed, in Fig. 6, for experiments in which the criterion was not reached, Scientist A's median effect size was below 0.25, whereas B's was close to the population value of 0.5. Of course, the potential for random sampling error to produce a too-small estimate of the population effect size must decrease as the population effect size decreases, and that pattern is evident in Figs. 2, 4, and 6 (Note that in Fig. 8, the population effect size is quite large [$D = 0.8$] and the stopping criterion was always met, regardless of the stringency of that criterion Fig. 7.

Also of note is that, when the population effect size (D) exceeded 0.0, and especially when it was 0.5 or 0.8, the median logBF for the yoked scientist, B, appeared to increase with the stringency of A's stopping criterion. A likely explanation is that in these conditions, there was a trend in which the median sample size, and therefore the strength of the evidence favoring the true, alternative hypothesis (that D does not equal 0.0), increased with the stringency of the stopping criterion (Table 1). Consistent with this interpretation, the Cohen's d effect sizes for Scientist B, in conditions where the population effect size exceeded 0.0, showed no trend clear trend of rising with the stringency of the stopping criterion (Figs. 4, 6, and 8).

Note that overall, regardless of optional stopping, the distributions of logBF were highly skewed, creating much greater potential to find strong evidence favoring the alternative hypothesis than to find strong evidence favoring the null.

See Anderson et al. (2021) for files containing the simulations' raw output.

Discussion

The present simulations employed a yoked scientist procedure that experimentally controlled for the size of the sample, thus establishing a standard for assessing the proposition that

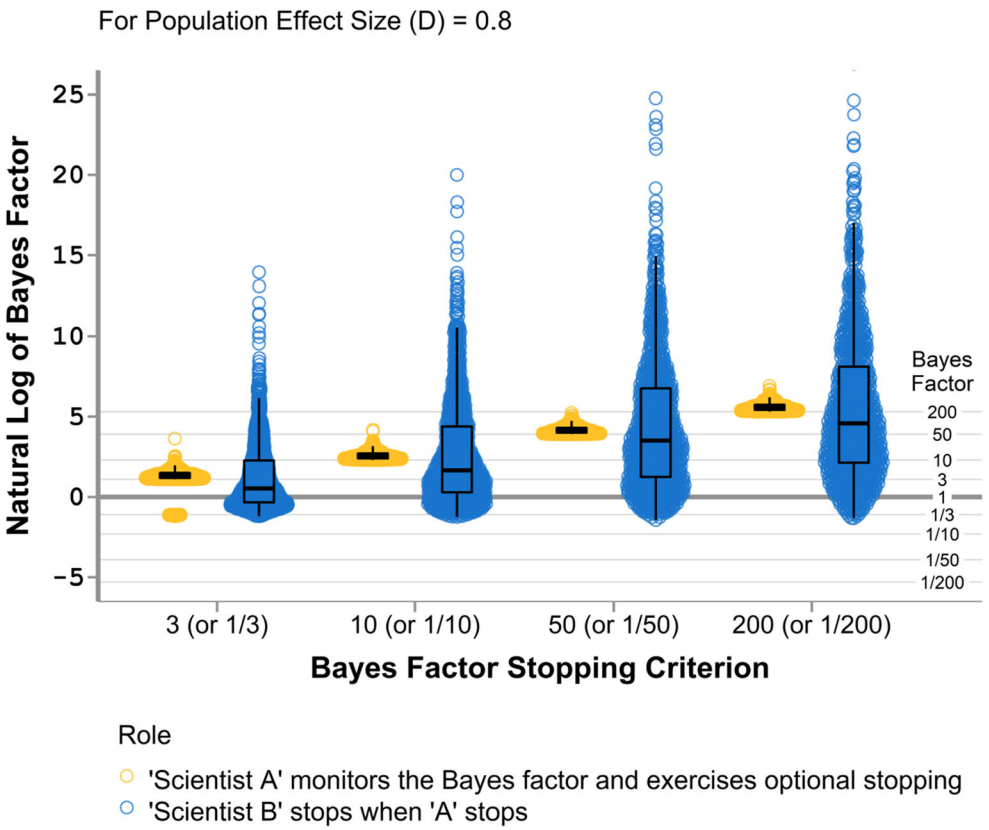

Fig. 7 For population effect size (D) = 0.8: The natural log of the obtained Bayes factor (logBF) as a function of the Bayes factor stopping criterion ($BF_{(crit)}$) and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the logBF that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the four levels of the stopping criterion, there were 1000 pairs of simulated experiments, with a maximum N, per experiment, of 500 (250 simulated participants per group)

optional stopping produces non-representative samples, and thus, biased inferences. The results showed that optional stopping can indeed inflate the obtained Bayes factor (due to distorted input to the Bayes-factor calculation)—as well produce bimodality, depending on the value of $BF_{(crit)}$ and the value of the population effect size—relative to a yoked standard. The simulations also implicate effect-size distortion as a principal contributor to distorting the obtained Bayes factor.

Prior to the present work, there was controversy (e.g., Sanborn & Hills, 2014; Yu et al., 2014) concerning whether establishing a critical value of the Bayes factor ($BF_{(crit)}$) as the criterion for stopping data collection renders the subsequently obtained Bayes factor an incorrect description of the likelihood of the data given the statistical hypotheses. The present findings indicate that the interpretation of the Bayes factor remains correct under optional stopping, but that the validity of the Bayes factor is compromised by bad input: Optional stopping distorts the input-sample causing it to be non-representative, thereby impacting the output.

The present findings are also relevant to the issue of replication. We have described our yoking procedure as involving *concurrent* data collection (in two simulated experiments). However, time, per se, plays no computational role in the simulations. Without changing any aspect of the simulation or the situation results, one can conceptualize the two scientists as being yoked *across* time, with Scientist B conducting an exact replication of Scientist A's experiment (i.e., using the same methods, and gathering the same sized sample from the same population). Therefore, the simulations demonstrate that optional stopping distorts effect sizes and Bayes factors not only with respect to a concurrent yoked experiment but also with respect to a subsequent attempt at exact replication.

Replication is often assessed within a meta-analytic framework focused on the estimation of effect-size. The present

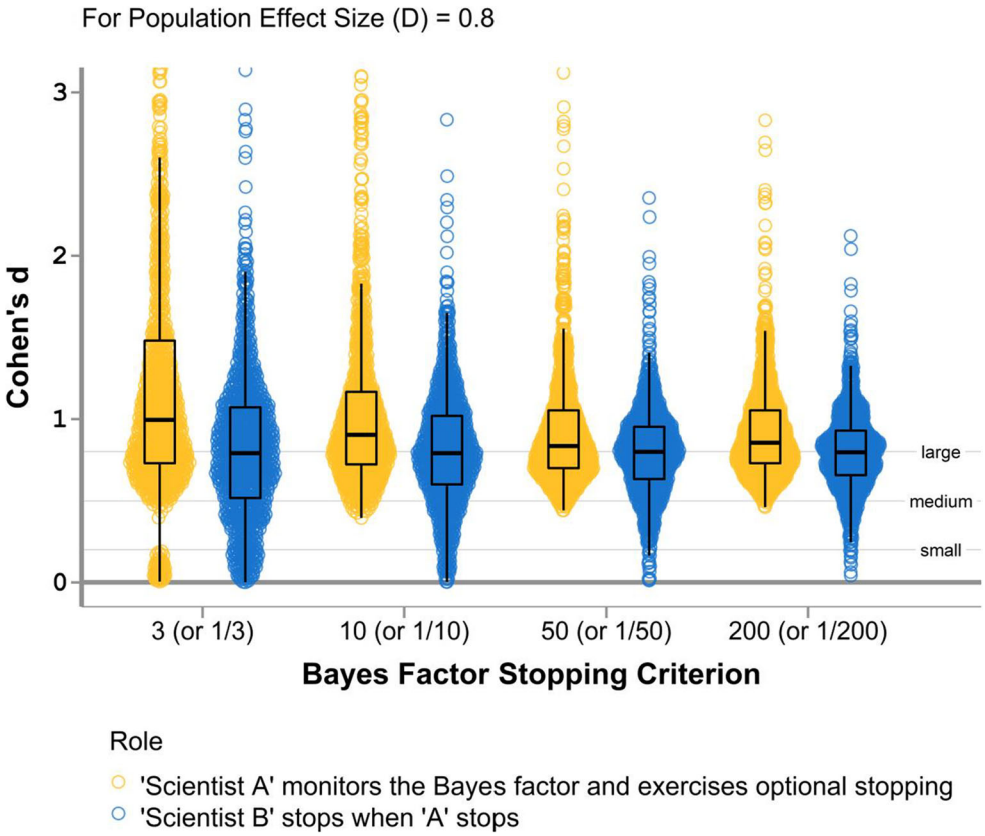

Fig. 8 For population effect size (D) = 0.8: the obtained effect size, Cohen's d, as a function of the Bayes factor stopping criterion ($BF_{(crit)}$) and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the logBF that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the four levels of the stopping criterion, there were 1000 pairs of simulated experiments, with a maximum N, per experiment, of 500 (250 simulated participants per group)

findings show that such effect-sizes can be distorted by optional stopping, even in the context of Bayesian analysis. Thus, the problem optional stopping poses for replicability (see Yu et al., 2014) is not solved by the use of Bayesian methods.

Study 2: Effect-size stabilization monitoring

Given Study 1's demonstration that the use of $BF_{(crit)}$ for optional stopping distorts the inputs to the Bayes-factor calculations, we performed a second set of simulations to assess whether a different kind of stopping criterion, effect-size stabilization, would avert the distortion caused by $BF_{(crit)}$. Some prior research has addressed the question of how much data are required for effect-size stabilization. In particular, Schönbrodt and Perugini (2013) investigated this question with respect to the stabilization of correlation coefficients. They found that there is no simple answer: Stabilization depends on how one defines stabilization as well as on the effect size in the population. In the present set of simulations, we defined stabilization as q consecutive minimal changes (plus or minus 0.05) in the value of Cohen's d. We reasoned that, whereas the use of $BF_{(crit)}$ biased the samples (in Study 1) to favor inclusion of extreme data points, thereby distorting the sample effect-sizes and thus the inputs to the Bayes-factor calculations, stabilization should not favor extremeness since extremity would run counter to stability. We therefore hypothesized that a stopping criterion defined as the observed stabilization of the effect size would lead to unbiased effect-size estimates and thus unbiased inputs to the Bayes-factor calculations.

Method

The method differed from that of Study 1 only insofar as there were three levels of the stopping criterion rather than four, and

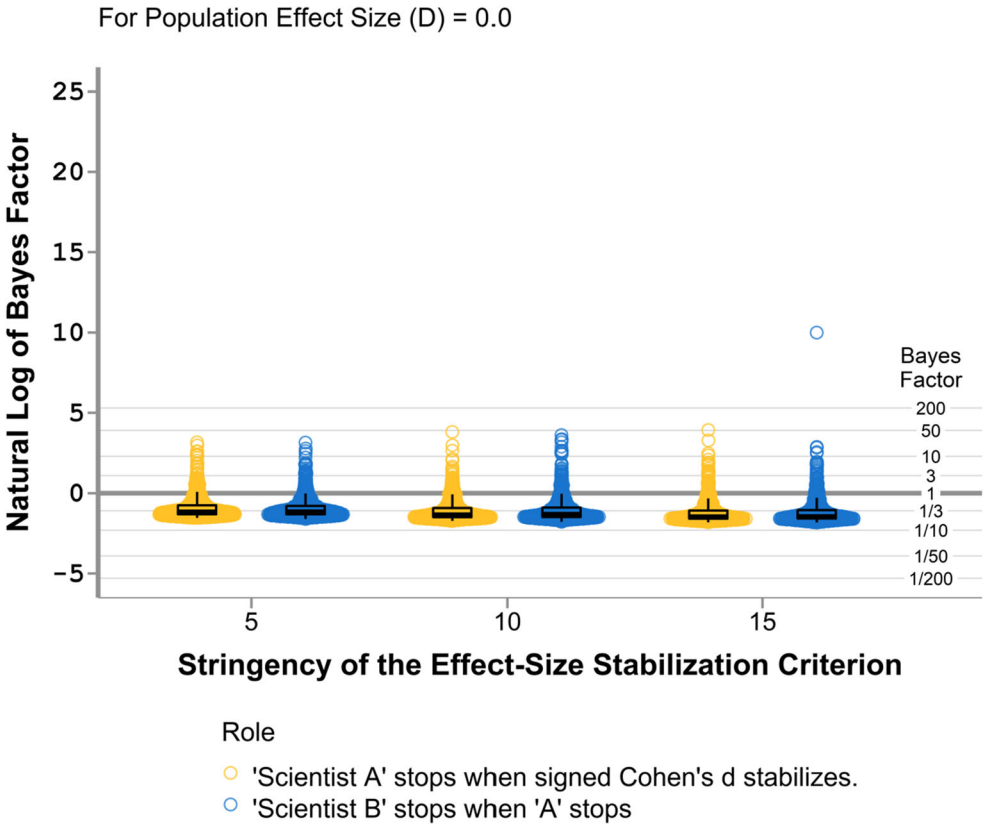

Fig. 9 For population effect size (D) = 0.0: the natural log of the obtained Bayes factor (logBF) as a function of the effect-size stabilization criterion for stopping and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the logBF that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the three levels of the stabilization criterion for stopping, there were 1000 pairs of simulated experiments, with a maximum N, per experiment, of 500 (250 simulated participants per group). Note that for the stopping criterion, the effect size (signed Cohen's d) was considered stabilized when there were 5, 10, or 15 consecutive instances in which adding a new data pair produced a minimal (< 0.05) absolute change in signed Cohen's d

insofar as those three levels were effect-size stabilization criteria rather than values of $BF_{(crit)}$. For each simulated experiment, the effect size was computed each time a new data pair was added to the sample. That effect size was calculated as Cohen's *d*, but with a positive or negative arithmetic sign, depending on whether the mean of the simulated Group 1 was higher or lower than the mean of the simulated Group 2. The sign was necessary in order to accurately index the change in effect size. For example, if signed Cohen's *d* were to change from 0.1 to – 0.1, that would indeed be a change, whereas it would be no change if the effect size were sign-less. With the addition of each new data pair, the simulation computed a change score equal to the absolute value of the change in signed Cohen's *d*. (Note that mathematically, Cohen's *d* can only be calculated when the number of data pairs is at least 2.) The absolute-value transform was necessary since stabilization is characterized by a sustained, minimal change in effect size—whether that change constitutes an increase or a decrease. The change score counted as minimal if it was less than 0.05. The stopping criterion was met when the number of consecutive minimal changes reached a critical value. Across conditions of the simulation, that critical number was 5, 10, or 15 consecutive minimal changes.

There were 1000 yoked pairs of experiments at each of the three criterion levels, and four levels of the population effect size, *D* (*D* = 0.0, 0.2, 0.5, and 0.8), yielding a total of 24,000 simulated experiments. See Anderson et al. (2021) for the R code for the simulations.

Results

The use of effect-size stabilization as a criterion for optional stopping had no discernable, distortive effect on the obtained Bayes factors: The distributions shown in Figs. 9, 10, 11, and 12 are nearly identical for the simulates scientist who

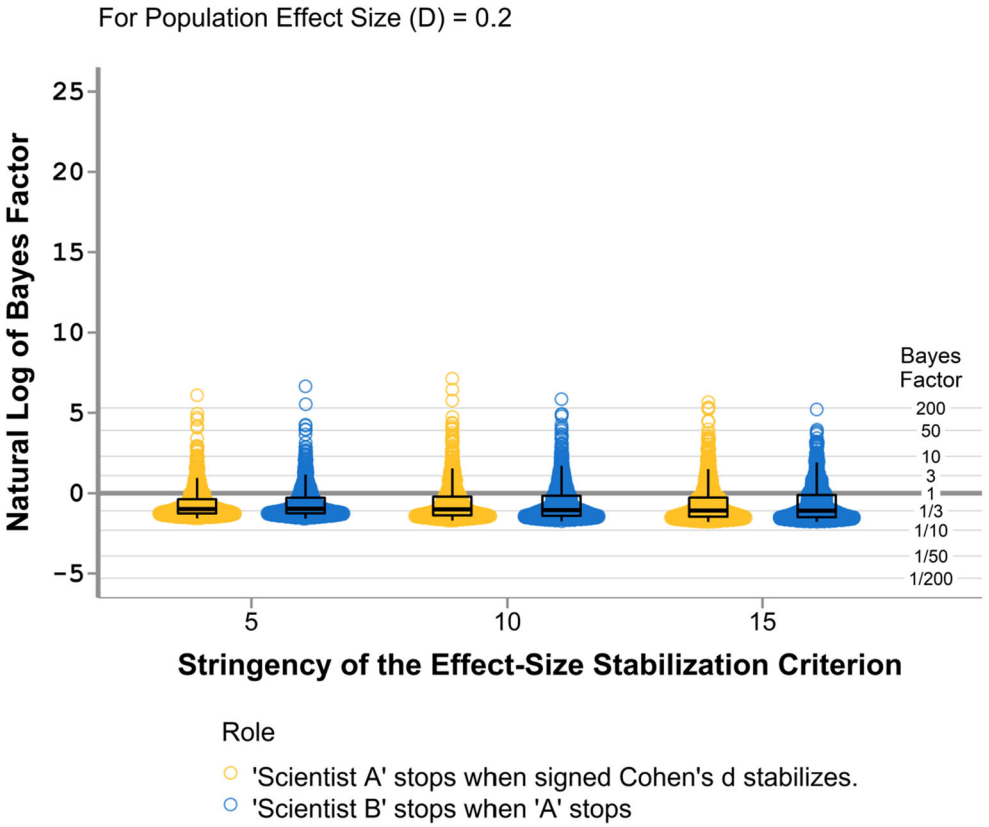

Fig. 10 For population effect size (*D*) = 0.2: the natural log of the obtained Bayes factor (log*BF*) as a function of the effect-size stabilization criterion for stopping and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the log*BF* that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the three levels of the stabilization criterion for stopping, there were 1000 pairs of simulated experiments, with a maximum *N*, per experiment, of 500 (250 simulated participants per group). Note that for the stopping criterion, the effect size (signed Cohen's *d*) was considered stabilized when there were 5, 10, or 15 consecutive instances in which adding a new data pair produced a minimal (< 0.05) absolute change in signed Cohen's *d*

optionally stopped (Scientist A), and for the yoked scientist. The binomial tests in Table 2 are consistent with the data patterns shown in the figures. Moreover, and as in Study 1, the results indicated a perfect correspondence between whether Scientist A's obtained Bayes factor (BF_A) exceeded that of Scientist B (BF_B), and whether Scientist A's obtained Cohen's d (d_A) exceeded that of Scientist B (d_B) Figs. 10, 11 and 12.

When the sampling stopped, the median sample size ranged from 68 (34 per group) to 122 (61 per group), depending on the stringency of the stopping criterion (Table 2). Notably, regardless of whether it was Scientist A or the yoked Scientist B, the median obtained effect size (upon stopping) was very close to the population effect size (Table 3). Additionally, the 25th and 75th percentiles of the effect-size distributions were similar across the two scientists (Table 3). See Anderson et al. (2021) for files containing the raw output for the simulations.

Discussion

The results of Study 2 showed that when optional stopping was defined as monitoring the sample effect size as the data accumulate, and stopping data collection once the effect size has stabilized, the distributions of obtained Bayes factors were unaffected by such stopping. Moreover, when the population effect size was medium-to-large (D = 0.5 or 0.8), the most stringent degree of stabilization produced the largest Bayes factors and therefore the strongest statistical evidence for the alternative relative to the null hypothesis.

In contrast to the type of optional stopping examined in Study 1 (stopping based on $BF_{(crit)}$), optional stopping based

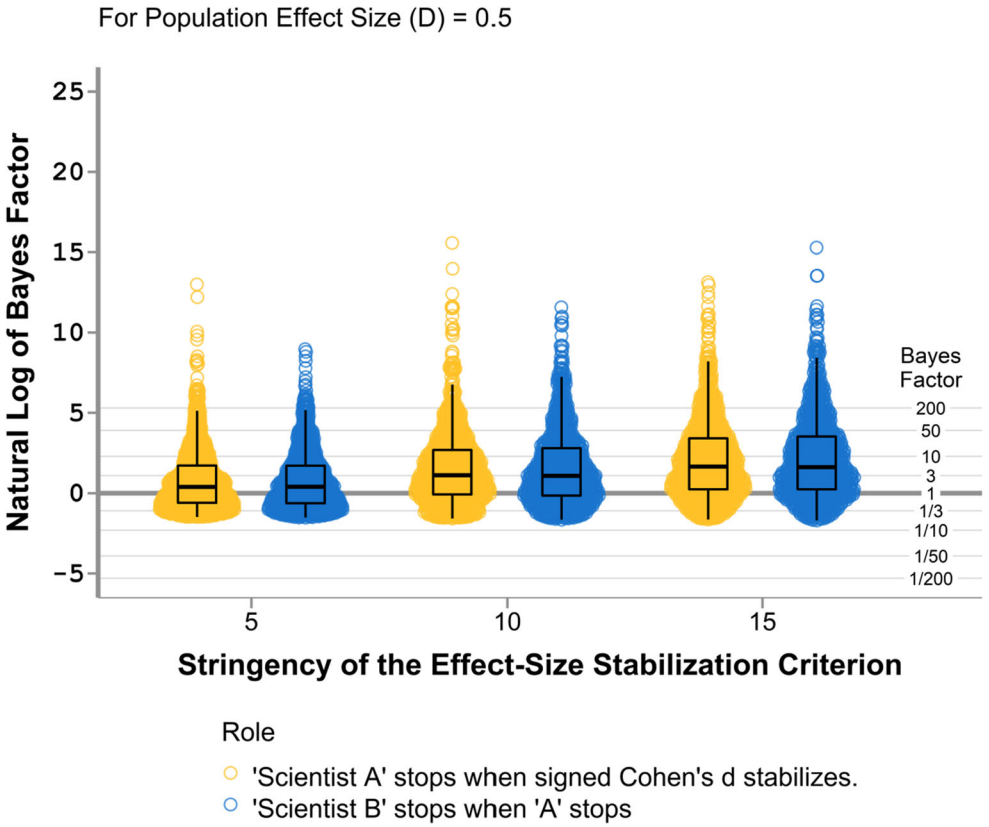

Fig. 11 For population effect size (D) = 0.5: the natural log of the obtained Bayes factor (logBF) as a function of the effect-size stabilization criterion for stopping and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the logBF that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the three levels of the stabilization criterion for stopping, there were 1000 pairs of simulated experiments, with a maximum N, per experiment, of 500 (250 simulated participants per group). Note that for the stopping criterion, the effect size (signed Cohen's d) was considered stabilized when there were 5, 10, or 15 consecutive instances in which adding a new data pair produced a minimal (< 0.05) absolute change in signed Cohen's d

on effect-size stabilization had no apparent impact on the representativeness of the samples, since there was no impact on the distribution of Bayes factors. A plausible explanation is that, in contrast to optional stopping based on $BF_{(crit)}$ in which the samples were distorted, the principal impact of stabilization-based stopping may be on the ultimate *size* of the obtained sample, with larger samples tending to be more stabilized than smaller ones. Thus, given that Scientist A and the yoked Scientist B necessarily obtain the same sample size, and given that stabilization-based stopping produces no apparent distortion of samples (Table 3 indicates no distortion of the obtained effect sizes), the distributions of obtained Bayes factors tend to be similar for the two scientists.

In summary, the present results demonstrate that optional stopping based on effect-size stabilization lacks the shortcomings of optional stopping based on $BF_{(crit)}$, and that therefore the former is to be preferred over the latter.

General discussion

The present studies employed a yoked-scientist simulation procedure to demonstrate that if a researcher monitors Bayes factor as the sample grows, and stops collecting data when the Bayes factor reaches a critical value, the result is a non-representative sample in which the effect size—and consequently the input to the Bayes-factor calculation—is distorted. It is perhaps instructive to consider a logical implication of such non-representativeness within the yoked-scientist paradigm: Imagine that Scientist A has stopped collecting data in an experiment—based on a critical value of the Bayes factor—and is asked, *"Now that both you and the yoked scientist, B, have finished your experiments, whose Bayes factor should the public trust more: Yours, or Scientist B's?"* The rational answer would be that B's is more trustworthy, since B's is based on a representative sample whereas A's is based on a sample that was made unrepresentative by optional stopping.

Fig. 12 For population effect size (D) = 0.8: the natural log of the obtained Bayes factor (logBF) as a function of the effect-size stabilization criterion for stopping and of the simulated scientist's role in the yoked sampling procedure. Each *circle* is a data point indicating the logBF that results from a particular, simulated experiment. At each level on the vertical axis, the random horizontal displacement of the circles indicates the density of the data at that point (see Clarke & Sherrill-Mix, 2017). Thus, the *pattern of the circles* shows the distribution of the data. The *rectangles* are box plots: The *top of the box* indicates the 75th percentile of the data, the *bottom of the box* indicates the 25th percentile, and the *horizontal bar* between the top and the bottom indicates the median. The *whiskers* on each box extend to the most extreme observation whose distance from the box's edge is less than or equal to 1.5 times the interquartile range (IQR)—where the interquartile range is the range delineated by the 25th and 75th percentiles. For each of the three levels of the stabilization criterion for stopping, there were 1000 pairs of simulated experiments, with a maximum N, per experiment, of 500 (250 simulated participants per group). Note that for the stopping criterion, the effect size (signed Cohen's d) was considered stabilized when there were 5, 10, or 15 consecutive instances in which adding a new data pair produced a minimal (< 0.05) absolute change in signed Cohen's d

Table 2 For optional stopping based on effect-size stabilization: Obtained sample sizes (*N*), numbers of samples, standard deviations (SD) of BF_A and of BF_B , proportions of simulated pairs of experiments in which the Scientist A's obtained Bayes factor exceeded those of Scientist B, and hypothesis tests for those proportions

Stabilization criterion stringency [a]	Obtained sample *N*			Prop. of cases where $BF_A > BF_B$ (and $d_A > d_B$) [b]	Binomial test [c] for prop. ($BF_A > BF_B$) = 0.5	
	Median	25th %tile	75th %tile		*p*	Bayes factor [c]
For Pop. D = 0.0						
5	68	56	78	0.48	0.217	0.088
10	98	86	112	0.49	0.591	0.047
15	122	108	136	0.49	0.393	0.059
For Pop. D = 0.2						
5	68	54	78	0.49	0.359	0.062
10	100	88	114	0.52	0.195	0.096
15	122	106	134	0.52	0.359	0.062
For Pop. D = 0.5						
5	68	56	78	0.51	0.591	0.047
10	100	88	112	0.50	0.975	0.040
15	122	108	136	0.48	0.268	0.076
For Pop. D = 0.8						
5	70	60	78	0.51	0.548	0.048
10	98	88	112	0.50	0.924	0.040
15	120	106	136	0.49	0.728	0.043

Note. For each level of population effect size (D), and for each level of stringency of the criterion, there were 1000 pairs of simulated experiments. In each simulated experiment, there was a maximum of 500 simulated participants (250 per group).

[a] For the stopping criterion, the effect size, signed Cohen's *d*, was considered stabilized when there were 5, 10, or 15 consecutive instances in which adding a new data pair produced a minimal (< 0.05) absolute change in signed Cohen's *d*.

[b] The results indicated a perfect correspondence between whether Scientist A's obtained Bayes factor (BF_A) exceeded that of Scientist B (BF_B), and whether Scientist A's obtained Cohen's *d* (d_A) exceeded that of Scientist B (d_B).

[c] The values in these columns evaluate the simulation results and not the individual, simulated experiments.

Table 3 For optional stopping based on effect-size stabilization: Distributions of Signed Cohen's *d*

Stabilization criterion stringency*	Signed Cohen's *d*					
	Scientist A			Scientist B		
	Median	25th %tile	75th %tile	Median	25th %tile	75th %tile
For Pop. D = 0.0						
5	0.00	– 0.18	0.16	0.01	– 0.17	0.17
10	0.01	– 0.13	0.14	– 0.01	– 0.15	0.14
15	0.00	– 0.12	0.11	0.00	– 0.12	0.11
For Pop. D = 0.2						
5	0.20	0.03	0.36	0.20	0.03	0.38
10	0.21	0.08	0.34	0.20	0.07	0.35
15	0.19	0.08	0.32	0.20	0.07	0.34
For Pop. D = 0.5						
5	0.50	0.33	0.66	0.50	0.32	0.67
10	0.49	0.37	0.62	0.49	0.36	0.64
15	0.49	0.37	0.61	0.49	0.37	0.61
For Pop. D = 0.8						
5	0.80	0.63	0.98	0.80	0.63	0.95
10	0.80	0.66	0.94	0.79	0.66	0.93
15	0.79	0.67	0.92	0.80	0.68	0.93

Note. For each level of population effect size (D), and for each level of stringency of the criterion, there were 1000 pairs of simulated experiments. In each simulated experiment, there was a maximum of 500 simulated participants (250 per group).

* For the stopping criterion, the effect size, signed Cohen's *d*, was considered stabilized when there were 5, 10, or 15 consecutive instances in which adding a new data pair produced a minimal (< 0.05) absolute change in signed Cohen's *d*.

As discussed in the thought experiment in this paper's introduction, the problem with optional stopping is not within the algorithm for Bayesian updating of beliefs in hypotheses. Bayes theorem prescribes what data are most consistent with which hypotheses, and which hypotheses are more likely given the data, but does not prescribe rules for the stopping of data collection. Thus, in light of the present findings, the problem with using a critical Bayes factor for optional stopping lies in the biasing of the input to Bayesian inference, not in the inference algorithm itself.

In summary, the present findings argue against the recommendation that scientists monitor the Bayes factor and use a critical value of the Bayes factor for optional stopping, and argue in favor of the recommendation to avoid such a practice.

However, the present findings also indicate that a different alternative form of optional stopping, one based on effect size stabilization, lacks the shortcomings of optional stopping based on monitoring either the *p* value (see Nuijten et al., 2016; Simmons et al., 2011; Yu et al., 2014) or the Bayes factor. We recommend that effect-size stabilization, for optional stopping, be employed in conjunction with—rather than instead of—the calculation of a priori power. As discussed earlier in the paper, a priori power estimates have limited precision given that researchers have only limited knowledge of the true effect size for the phenomena being studied. Nevertheless, an a priori power estimate can provide a rough plan for the size of the research sample, with optional stopping based on effect size stabilization serving as a means to fine-tune the study's efficiency by averting the collection of too little or too much data. We think the approach outlined here has the potential to improve the replicability of research findings by avoiding situations (i.e., optional stopping based on critical *p* values or critical Bayes factors) that give rise to non-representative samples, and by promoting adequate and finely tuned statistical power (via effect size stabilization) for replicating effects.

Author Note We thank Michael E. Doherty for his critical reading of the manuscript.

References

Anderson, R. B., Crawford, J. C., & Bailey, M. H. (2021, January 28). *The distorting effects of optional stopping on Bayesian inference: Public files*. Open Science Foundation. https://www.osf.io/9jd6g

Bayes & Price (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 (1683–1775), 53*, 370–418. Retrieved January 17, 2021, from http://www.jstor.org/stable/105741

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge.

Clarke, E., & Sherrill-Mix, S. (2017). *ggbeeswarm: Categorical Scatter (Violin Point) Plots* (R package Version 0.6.0) [Computer Software]. The R Foundation. https://cran.r-project.org/web/packages/ggbeeswarm/ggbeeswarm.pdf

Ioannidis, J. A. (2005). Why most published research findings are false. *Plos Medicine, 2*(8), 696–701. https://doi.org/10.1371/journal.pmed.0020124

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

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. https://doi.org/10.1146/annurev.psych.59.103006.093735

McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods, 21*(1), 47–60. https://doi.org/10.1037/met0000036

Morey, R. & Rouder, J. (2018). *Computation of Bayes Factors for Common Designs* (R package Version 0.9.12-4.2) [Computer Software]. The R Foundation. https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48,* 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

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. https://doi.org/10.1177/1745691612465253

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review, 21*(2), 283–300. https://doi.org/10.3758/s13423-013-0518-9

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

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. https://doi.org/10.1177/0956797611417632

The Psychonomic Society (2020). *Statistical guidelines.* https://www.springer.com/journal/13423/submission-guidelines?detailsPage=aboutThis#Instructions%20for%20Authors_STATISTICAL%20GUIDELINES

Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology, 36*(1), 1–2. https://doi.org/10.1080/01973533.2014.865505

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1–2. https://doi.org/10.1080/01973533.2015.1012991

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638. https://doi.org/10.1177/1745691612463078

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. https://doi.org/10.1080/00031305.2016.1154108

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. https://doi.org/10.3758/s13423-013-0495-z