**Reviewer #1:**  
*1) Section 3.1.2, Equation 5 and “This equation states that the posterior odds of two hypotheses is obtained by updating the prior odds of these hypotheses with the Bayes factor, reflecting the information in the data. Subsequently, the posterior odds after observing a first data set can be used as the prior odds for new data.” Equation 5 holds only when the old data and old hypotheses’ parameters are independent of the new data and new hypotheses’ parameters.*

We have extended the section introducing BES considerably (pages 14-16), including an explicit explanation on the assumptions of BES (page 15). This also returns in the revised first part of the discussion section (page 24-25).

*The issue with the description (and the method itself) is that Bayesian update requires both the probability of the hypotheses and the parameter distributions of the hypotheses are updated by the data. In other words, data update our beliefs about the parameter distributions under each hypothesis as well as the probability of the hypotheses themselves. Updating both simultaneously guarantees all the desirable properties of Bayesian inferences (e.g., coherence when updating our results with a single observation at a time, in batches, or all together – as done in Bayesian meta-analysis/BSU).*

To serve as an updating approach providing the same results as BSU, it is indeed a requirement to also update the parameter distribution. We have added this information in the last paragraph on page 15. In addition, a small numerical illustration of simultaneous updating odds and parameter distribution is given on page 20.  
Furthermore, we explain that with different models (with different model parameters and parameter distributions) this is not feasible. In those cases, BES is suggested as an alternative. However, it is an alternative with different assumptions and different (interpretation of) results. This is now explicitly explained in the section introducing BES (pages 14-16) and re-iterated in the revised first part of the discussion section (page 24-25).

*The authors suggest BES as an alternative to Bayesian meta-analysis/BSU in cases when the data are in different formats and modeled with different parameters. Whether the different data formats and different model parameters differ enough to guarantee independence is a question that needs to be answered in each individual instance. One might, however, wonder whether the complete independence of parameters required to describe different data then relates back to a single hypothesis.*

We do belief that this is the case, and the paper includes a few examples of different informative hypotheses that investigate the same underlying common theory (page 14; and in the simulations in section 3.4). Further, see also our response to the (related) first two comments of this reviewer.

*2) Section 3.2.*

*I believe that it is fair to say that Bayesian meta-analysis/BSU provides the correct solution in the exact replication example (as one needs to update both the parameter distributions and the hypothesis probability). Any other result, i.e., the result obtained by BES, does not correspond to Bayesian evidence in favor/against the hypotheses. Then, I find the following comment a bit misleading:   
“When testing against the unconstrained model, BSU ‘suffers’ from the fact that the BF has a maximum.”,   
as if it were to suggest that BES overcomes a limitation of Bayesian meta-analysis/BSU (which, in fact, provides the correct answer here). I think that this section should be clarified to clearly state that Bayesian meta-analysis/BSU is a correct solution in this case and BES suffers from such and such shortcoming (i.e., splitting the data sets into different sizes will result in inconsistent results).*

We agree that this was phrased in a misleading way. The word “suffers” is removed and this sentence now reads (Page 17): “When testing against the unconstrained model, BSU still has a maximum value determined by the complexity of the constrained hypothesis, irrespective of the number of studies.”  
Also, we have added remarks about BSU being a method that is both consistent and coherent.   
Page 12: “The method is coherent, meaning that, with the same initial prior distribution, the posterior after adding one set of 100 observations is exactly the same as the posterior after sequential updating, for instance, after every tenth observation. Also, the order in which (subsets) of data come in does not affect the final resulting posterior.”   
Page 13: “When used for null hypothesis testing (i.e., BF1,0), with large enough sample sizes the SBF always converges to zero (if H0 is true) or infinity (when H1 is true), that is, it is consistent. This does not automatically generalize to comparing other types of hypotheses with Bayes factors. For instance, when used for testing an order constrained hypothesis against the unconstrained hypothesis (i.e., BFi,u), the SBF will converge to zero when Hi is not true, but to a maximum value, determined by the complexity of the informative hypothesis, when Hi is true (see also the previous section). However, it generally holds that by pooling the data from multiple studies, as is done with BSU, the overall power to detect true effects increases.”

*3) Section 3.3*

*“It is, for instance, not at all straightforward to compare the effect of a key variable on a continuous outcome with the effect of this same variable on a dichotomous outcome.”*

*This statement is not correct. The first case would result in using Cohen’s d/Spearman’s r and the second in log(OR), which can be, under some assumptions, converted into a common metric (Borenstein et al., 2009).*

Although the comparison in this section was between BSU and BES (not meta-analysis), we agree that this sentence can be misleading. We reformulated the section (3.4, page 21), including the addition of an explicit statement that meta-analysis might still be applicable by converting effect sizes.

*4) Discussion*

*I think that the aforementioned limitations and issues warrant revision of the recommendations.*

*“As long as the theory of interest can be formalized as an informative hypothesis, the support for the theory in each study can be expressed in terms of a Bayes factor, and BES can be applied to synthesize the evidence for this theory over studies.”*

*This is a very strong and unwarranted statement. Instead, I believe that Bayesian meta-analysis/BSU should be employed whenever possible. In cases when the different study designs do not allow for coherent Bayesian updating, BES presents a feasibly alternative for combining evidence from multiple studies.*

The first part of the discussion section is rewritten in line with the above comments (page 24-25).

*“BSU pools all available data and thus provides the support for the hypothesis of interest in all studies when taken together. This increases the total sample size and thus statistical power to find support for the true hypothesis, which is usually considered a strength of this method.”*

*The important thing is that BSU is the coherent way of Bayesian updating. Therefore, the strength of BSU is that it quantifies the evidence (instead of “usually considered”).*

The revision of the first part of the introduction (see also previous point) does no longer include similar statements. The coherency and consistency of BSU are explicitly mentioned, and BSU is recommended for those situations where it is feasible.

***Minor comments***

* *From the abstract “Bayesian model selection for informative hypotheses”, I would consider changing to Bayesian (informed) hypotheses tests, as it is more closely connected to the rest of the paper (model selection also invokes association with information criteria as LOO-CV and WAIC).*

We agree and changed to ‘Bayesian informative hypothesis testing’ in the abstract (page 1), and also at other places in the manuscript.

* *The authors use “Informative hypotheses” to denote only (order) constrained hypotheses. However, there are other types of informed hypotheses that can be characterized in other ways (i.e., non-local prior distributions). The outlined methodology can deal with those types of hypotheses too. It seems a bit of a missed opportunity to equate informed hypotheses with only (order) constraints as the NHT framework can also deal with some specific cases, e.g., one-sided and equivalence tests. I would suggest mentioning other forms of informed hypotheses (see Gronau and Wagenmakers, 2019, for an example from psychology).*

The intention of the paper is to add an aggregation method (BES) to our previous work on informative hypothesis testing, where informative hypothesis testing is defined as evaluating hypotheses with (equality, order, range) constraints on the model parameters or functions thereof. Extending to other types of informed hypothesis goes beyond the aim of this paper.   
Based on reviewers’ suggestions, the section introducing BES now mentions and includes references to related work by Stephan et al., Regenwetter et al., Klaassen et al. (page 15).

* *From Discussion: “A second issue that needs attention is that Bayes factors are not only affected by the fit of the data to the hypothesis (including the impact of effect sizes and sample sizes) but also to the complexity of the hypothesis. Hypotheses that are more specific (i.e., with a smaller complexity) will, with the same fit, provide larger Bayes factor values and thus have a stronger impact on the aggregated evidence. To what extent or under which circumstances this is desirable or problematic requires more research into different scenarios.” This is a principle of parsimony (Occam’s razor) which I believe is a generally accepted notion amongst scientists. Is there a particular reason why it should be problematic and require more research in different scenarios?*

As we explained in simulation 2, the strength of evidence for a hypothesis in a single study does not only depend on the size of the true effect (to what extent is the general theory supported in the population) and sample size but also on the operationalization of the general hypothesis into a study-specific informative hypothesis. Although it is not surprising that resulting BFs depend on the precise formulation of the tested hypothesis, the effect this has when aggregating over multiple studies with different informative hypotheses is yet unknown. The corresponding sentences have been rewritten to better explain this reasoning:  
Page 25:   
“A second issue that needs attention is that Bayes factors are not only affected by the fit of the data to the hypothesis (including the impact of effect sizes and sample sizes) but also by the complexity of the hypothesis. This principle of parsimony is generally known and understood when evaluating hypotheses with a single study. However, in the context of combining evidence levels from multiple studies, which do investigate the same underlying theory but may have different study specific hypotheses, the role of complexity is more complex. Study specific informative hypotheses that have a smaller complexity compared to other studies in the set will, with the same fit, provide larger Bayes factor values and thus have a stronger impact on the aggregated evidence. To what extent or under which circumstances this is desirable, or problematic requires more research into different scenarios.”

* *I would recommend adding a link to the GitHub repository directly in the text where it is mentioned.*

Done as suggested (page 4).

* *I find the critique of null hypotheses too strong “Since the null hypothesis is considered unrealistic (’the exact null is never true’) and usually does not represent a theoretical expectation of the researcher, it could be argued that it is not a good competitor for the informative hypothesis.”. There are many researchers who would disagree with the statement. Good counterexamples might include the null effect of nasal oxytocin (as it cannot physiologically cross the physiological barrier) or precognition.*

This no longer applies. Based on comments from reviewer 2 we considerably shortened Section 2 and 2.1 and this resulted in removing the part summarizing critique of NHST altogether. In the current manuscript, the only remaining (critical) remark about NHST has bene formulated more subtle as: “Although the null hypothesis has been criticized as a potentially unrealistic option (e.g., Cohen, 1994; Krueger, 2001; Lykken, 1991) and usually does not represent a theoretical expectation of the researcher, it is still a popular alternative.” (page 6)

* *“Finally, in contrast with the NHT approach, no penalty for multiple testing is required when computing the Bayes factor after each updating step (Schönbrodt et al., 2017).” I think it would be good to include quantification that this statement concerns the evidence, not the error rate, as some people often incorrectly claim.*

We elaborated on this point by extending the corresponding paragraph. Page 13: “This implies multiple or interim testing, that is, at several points in the data collection the SBF is calculated and based on the resulting value one either stops, or adds more data. Under the NHST framework such interim analyses require a correction for multiple testing. It is generally accepted that no penalty for multiple testing is required when computing the Bayes factor after each updating step. However, potential bias of Bayes factors in sequential designs is a topic of debate and long-term error rates can be affected by the SBF approach. Further elaboration is beyond the scope of this paper; interested readers are referred to Schönbrodt et al. (2017) and references cited herein.”

* *Simulation depicted in Figure 4/5 – I would find it informative to also visualize the behavior under the true null (i.e., R2 = 0).*

We understand this remark but decided against adding more simulations. Obviously, whatever choice you make for simulation settings, there are other scenarios that would add valuable information. However, adding simulations from a null population (R2=0) would in our opinion require multiple investigations: BES results when data come from the null and the hypothesis of interest is 1) also the null hypothesis, or 2) is not the null hypothesis. This would add several more simulations and figures and, we feel, does not fit with the introductory purpose of this paper.

Given that this is an introductory paper, the current simulations are, in our opinion, already highly informative for potential users. Our main messages to demonstrate with this paper are: 1) Aggregation using BES is feasible for diverse study designs, with one example varying the outcome, and a second example varying both the outcome and predictors. 2) How does BES perform when the studies come from a population in agreement with the informative hypothesis of interest, with increasing effect sizes from underpowered to well-powered settings.

* *It might be worth mentioning additional limitations of BES in contrast to the Bayesian meta-analysis/BSU, such as the inability to quantify the effect of moderators or adjust for publication bias.*

We have changed the section introducing BES to deal with a variety of remarks made by the reviewers, and this section now also includes remarks about limitations of BES (Page 16). Also in the discussion section, we start with a discussion of differences between BES, BSU, meta analysis and the limitations of BES compared to the more common alternatives (pages 24-25 ).

**Reviewer #2:**

*The paper reviews three main topics: modeling with informative hypotheses, Bayesian sequential updating (BSU), and Bayesian evidence synthesis (BES). While I enjoyed the review of informative hypotheses, this topic has already been covered at length in the literature, including a recent tutorial (Regenwetter & Cavagnaro, 2019).*

We considerably shortened the paragraphs on informative hypotheses, since, indeed, there are several papers already explaining this. This implies that section 2, until the introduction of the binomial example, has been completely rewritten (pages 4-8). We also included a reference to the R&C paper.

*Moreover, the review seems to lack some needed precision and formality. For example, it presents inequalities among population means, but it never formulates a model linking those means to data, nor does it discuss how those inequality constraints are encoded in the model via the prior. It is also somewhat vague in describing the quantities f\_i and c\_i from which one can compute the Bayes factors. The range of examples helps to convey the generality of the approach, but some of that should be described explicitly in the review, before things get more complicated with aggregating evidence across studies. What about other types of models like utility models, or process models? Incorporating informative hypotheses into different types of models than what have been covered frequently in the literature could also improve the novelty.*

This part has been shortened considerably as a reaction to the first point of the reviewer. This makes several points in this comment no longer applicable. We did not follow the suggestion to increase the level of novelty by explicitly discussing other types of models, because this will potentially make this part of the manuscript unnecessary complex, and the novelty is supposed to be in the third section on aggregating studies. Instead, we summarized how the software we used to compute BFs (BFpack) works (type of priors, estimation method) and state that this software can be applied for a wide variety of statistical models. We did not include too many technical details though (not all readers would appreciate them) but refer to the BFpack software manual instead (pages 5-8).

*The reviews of BSU and BES are also accessible and intuitive, but end up feeling somewhat superficial. Again, more formality in how to calculate the statistic for BES would be welcome. It seems to be just the product of Bayes factors across studies. This simplicity could be viewed as a strength, and presents an opportunity to connect the statistic to ideas of independence and joint probability.*

The section introducing BES is extended, in line with comments and suggestions made by all reviewers (pages 14-16). The computation of BES results can be found on page 14, a remark on the independence assumption on page 15.

*To elaborate further on that last point, the paper misses an opportunity to dive deeper into the differences between BSU and BES. The paper touches on statistical power, and also touts the flexibility of BES for being compatible with diverse studies, yet it is relatively silent on what the statistics actually mean. One important difference has to do with heterogeneity across studies. To wit, note that the Bayes factor is an evidence ratio. It tells us the weight of information for a target hypothesis (here, an informative hypothesis) relative to an alternative (here, either the unconstrained, the complement, or the null). In BSU, where data are pooled, both the target and the alternative assume that the parameter value is identical in all studies. It is not just that the same constraints hold in all studies, it is that the specific parameter is identical in all studies. The resulting Bayes factor in BSU tells how much more likely it is that this common parameter satisfies the constraints of the target model than it does those of the alternative. Thus, the BSU gains model selection power by combining observations to get a better estimate of that common parameter value.*  
*On the other hand, under BES, model parameters are estimated independently in each study. The specific parameter value of the model is allowed to vary from study to study. The BES statistic, which is essentially the product of Bayes factors across studies, is itself a Bayes factor. The numerator is the joint probability of the data from all studies, assuming that the constraints of the target model hold separately in every study. The denominator is the joint probability of the data under the assumption that the constraints of the alternative model (e.g., unconstrained) hold in every study. Thus, like the Bayes factor in BSU, it is also an evidence ratio, but it is one that compares different hypotheses than those considered in BSU. This is especially important when there are more than two hypotheses under consideration because, in this case, BES will select the model that is the most likely to hold in every study, even if that model is not the best model for any particular study.*  
We agree with everything the reviewer is pointing out here and we thank the reviewer for these constructive comments that helped us see what was missing or not yet clear in the previous version of the paper. The extensive rewrite of the manuscript should now explain these differences much better (most notably in section 3.1 and 3.2 and in first part of the discussion).

*The same methods, with different names, have been considered for aggregating evidence across participants in within-subjects analyses. In this context, BSU is known as the "pooled Bayes factor" and BES is known as the "group Bayes factor" (Stephan et al., 2009). Some differences between these approaches are discussed by Regenwetter et al. (2017).*  
Thanks for mentioning these references. This work is indeed relevant and thus now referred to in our paper (page 15).

*In sum, this paper provides an accessible review of some useful contemporary methods. There is not much novelty to it, but it does bring together some topics that go well together. It would benefit from more depth and formality. It would also benefit from elaboration on the relationships between these methods, which could add some novelty and help to make the whole more than the sum of its parts.*  
As stated earlier, the focus and thus novelty is in the part where we present and investigate BES (section 3). We feel this is now better achieved by shortening section 2 considerably and improving other parts in terms of depth and formality by following the advises by the three reviewers. We trust that the balance between (still) being accessible to a broad audience and level of novelty/formality is now much better.

**Reviewer #3:**

*Page 5, fourth paragraph: I did not see "NHT" defined anywhere. Presumably this is "null hypothesis testing." Because the authors cite Jacob Cohen's classic work, I might suggest using "Null Hypothesis Significance Testing (NHST)" to be consistent with his nomenclature.*Thanks for pointing out this omission. NHST is now defined on page 4; and throughout the manuscript NHT is replaced by NHST.

*Page 8, third paragraph: "…the option that all sample effects are likely to be chance results…" Would this not be equivalent to the informative hypothesis that the values in question are equivalent to one another, up to the resolving power of the sample size? Would that allow a more direct evaluation of the current method rather than a classical null hypothesis? There is some discussion of this issue at the bottom of page 12, but it is still not entirely clear to me.*  
We are not sure what the reviewer is asking here (nor which part on page 12 they are referring to). However, we slightly adapted the sentence on page 8 to better clarify its meaning. Page 6: “Often, researchers prefer to include and evaluate the option that there is no effect in the population and thus all sample effects are likely to be chance results. Therefore, they may want to compare the theoretical expectation with the hypothesis stating that there are no effects at all.”

*Page 9, second paragraph: "We usually assume both hypotheses equally likely before observing any data, that is, the prior odds equal one." If the authors mean here that scientists conducting hypothesis testing generally assume equally likely priors, this is an overgeneralization. If the authors simply plan to analyze this particular set of priors for the purposes of convenience in this study, that should be stated clearly.*  
We adjusted our text to express this choice along the line the reviewer suggested. Page 8: “A common choice, and also the default setting of BFpack for the computation of PMPs, is to assume that all hypotheses are equally likely before observing any data.” and a few lines below: “Note that different prior probabilities (i.e., prior odds) can be specified depending on the context. However, in the examples in this paper we will always use equal prior probabilities.”

*Page 10, first full paragraph: "It is easy to see that this distribution equals the likelihood in (3)." Conceptually this is straightforward because of Bayes' theorem and the specific case of the uninformative beta prior, but I do not think it is obvious in the general context of beta/binomial conjugate priors. Could the authors elaborate?*  
We are not sure what elaboration the reviewer is asking for. The first sentence of this paragraph provided the (general) relation between the beta prior, likelihood, and beta posterior. In the general case, the posterior is not equal to the likelihood, but it is equal to a beta distribution with hyperparameters that are a combination of information from prior and likelihood. The specific sentence that the reviewer refers to, applies this relation to the use of a uniform prior.   
However, to avoid similar confusion in other readers, we have changed the sentences into a more logical order. Page 8: “The unconstrained posterior distribution using the binomial likelihood and the Beta(α, β) prior is Beta(α + x, β + n − x). We use (4) with α = β = 1 as the prior distribution for a success probability θ without any constraints imposed, that is, Hu : θ. This is equal to the uniform distribution on the interval [0,1], i.e. p(θ) = 1. With this choice one states that, a priori, each value for θ between zero and one is considered equally likely. With this prior, the resulting posterior is the Beta(x + 1, n − x + 1) distribution. It is easy to see that this distribution equals the likelihood in (3), that is, the constant prior does not add any information about θ, and therefore the posterior is determined by the data only.”

*P17, third paragraph: I suggest "top row" instead of "first row" to indicate the plots containing the highest PMPs.*  
We have changed this as suggested (pages 18, 19)

*Page 21, last paragraph: "…a lack of power in the individual studies accumulates when using BES." This is an important limitation of BES and should be clearly stated in summary descriptions of the method.*  
The section introducing BES (3.2) has been rewritten and includes an explicit explanation on this issue on page 16 (last paragraph of 3.2). Also, in the discussion we list this as a limitation of BES (compared to data pooling methods). Page 24.

*Page 24, last paragraph: "…relatively novel…" Strike "relatively" and make an argument for why this approach is novel.*  
The beginning of the discussion section has been rewritten to deal with several comments and no longer includes the term “relatively novel”. We now start with a comparison of BES with the more common methods BSU and meta-analysis (pages 24-25).

*Table 1: The posterior for the last line can be read off Figure 1, but would it be advantageous to label or refer to the areas that give rise to the posteriors in the other hypotheses?*

We did not adapt the figure: since there are 4 areas in Figure 1 that are used in table 1 (both the prior and the posterior have a part below 0.6 and a part above 0.6), we feel it should be sufficiently clear with the vertical line at 0.6 (and that it will become rather cluttered/messy if we add more labels or, for instance, shade certain areas). However, we added an explicit reference to the areas in the figure in the text. Page 10: “For Hi this is the area to the right of the vertical line at θ = 0.6 in the prior and posterior distributions. For Hc we need the two areas to the left of this vertical line.“