|  |  |
| --- | --- |
| Reviewer | To Do |
| Reviewer 1 |  |
| I don't think it makes a very strong case for the final recommendation. In particular, I am not convinced that their recommendation actually makes sense. |  |
| 1) That the simulations show little difference between p<.05 and p<.005 criterion seems to be a side effect of including more stringent analyses (BF or OOM). My guess is that the two curves would be nearly the same without any NHST test, because the other tests are limiting Percent Agreement. | This graph is literally only done on the nhst and bf info… |
| 2) I think there can be reasons to be prefer different analyses for different situations, but it seems hard to justify using a variety of analyses for (nearly) every situation. If someone wants a procedure that controls the Type I error rate, then why bother with the Bayes Factor analysis (indeed, they will lose control of the Type I error rate)? If someone cares about behavior of individuals (OOM) then why bother with an analysis based on means (NHST)? | Clarify in discussion that we are not suggesting every researcher do this |
|  |  |
|  |  |
|  |  |

3) More generally, the conclusion seems to promote a one-size-fits all approach that introduces arbitrary thresholds with multiple analytically approaches. Using these different approaches simultaneously does not get away from the one-size-fits all approach. I think it is ill advised to advocate decision criteria that do not consider utilities (gains and losses) of different decisions. To be fair, the text is more clear that the analyses can be (and should be) considered without criteria, but the simulations and conclusions emphasize such decisions. I found this contrast rather jarring.   
  
4) The figures are rather confusing and hardly described in the main text.   
  
5) I struggled to figure out what new message is being presented here. Comparing p-values and BF values has been done many times. The suggestion to utilize multiple analytic methods has also been suggested many times (even though I think it makes little sense). I guess what is new here (to me, anyhow) is the simulations comparing OOM with p-values and BFs. It was interesting, but I often felt like it was comparing apples and oranges. p-values and BFs are analyses that focus on means, so they are at least discussing similar issues. OOM focuses on rather different issues (individuals or distributions of individuals), so I do not think there is a clear comparison with the other analytical approaches.   
  
6) Most of the history of the different methods is already well known, but maybe I am not representative of the manuscripts target audience.