**Peer Review and Communication History**

**MS Title**: A Two–Stage Bayesian Sequential Assessment of Exploratory

Hypotheses

**Author Names**: Angelika M. Stefan, Lukas L. Lengersdor, Eric-Jan Wagenmakers

**Submitted:** Apr 11, 2022

**Editor First Decision**: Revise & Resubmit

May 26, 2022

Dear Prof. Stefan,
I have received two reviews for your manuscript “A Two-Stage Bayesian Sequential Assessment of Exploratory Hypotheses.” Both reviewers are very positive about your manuscript, and I share their enthusiasm as well.

However, the Reviewers made some suggestions that may help to clarify some aspects of your submission and further improve your contribution.

I, therefore, recommend you carefully consider each of the points raised by the reviewers and then submit a revised version.

Best,
MTL

# Reviewer 1

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  | ✔ |  |  |

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

I read this clear and well written paper with great pleasure and found the two-stage Bayesian sequential procedure a very promising and useful inferential tool.
Specifically, beyond the advantages described by the authors, I think that the proposed method is particularly suitable for concretely formalizing the hypotheses that are going to be tested in the second (i.e., confirmatory) stage of the procedure. Indeed, a precise formalization of hypotheses is one of the main issues that psychology has to address to increase the validity and the credibility of scientific results. Furthermore, if in the first stage several models are compared, the great advantage of this procedure is that information in the form of posterior distributions can be collected for all the model parameters involved in the estimation process.
In short, my opinion of the paper is very positive, and I have only a few points that the authors may consider to improve the quality of the paper.

1. Although not mandatory and not applicable in all cases, I think that the first stage could also be pre-registered (e.g., via Exploratory Reports) and published as a stand-alone paper for a number of reasons: 1) to describe the set of hypotheses that will be tested; 2) to justify the minimum sample size at which the sequential procedure will start ; 3) to indicate the a priori thresholds that define sufficient evidence for a hypothesis; and 4) to give the right credit to the important work done, thus encouraging researchers to conduct this important stage. I invite the authors to briefly discuss this option in their paper.
2. On page 2, second paragraph, the authors write :” (b) to sacrifice resources to an initial exploratory study that does not allow for (frequentist) statistical inferences “. Personally, I would delete the part “(frequentist)”; otherwise, the author(s) should briefly justify their choice.
3. To better familiarize the reader with the author(s)’ procedure, I think that an application (even on simulated data) should be presented This application could be briefly summarized in the main text and reported more in detail as Supplemental Material. I am convinced that presenting an application would greatly improve the quality of the paper. In any case, I leave this decision to the Editor on the basis of which kind of article is is expected by the journal (e.g., a theoretical article, a short communication or a full research article).

Gianmarco Altoè

# Reviewer 2

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  | ✔ |  |  |

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The manuscript proposes an analysis procedure composed of two stages: An initial, exploratory stage to generate hypotheses and a concrete analysis plan, and a subsequent confirmatory stage. The authors outline the basic concept of this procedure and discuss advantages and potential caveats.

The manuscript addresses a timely issue. Many researchers may shy away from preregistrations, despite their benefits, because they appear too constraining and require decisions that one may not feel comfortable to make before seeing the data. An explicit exploration stage gives room to explore the consequences of these decisions without compromising the integrity of a subsequent, critical confirmatory hypothesis test. Most importantly, researchers can develop a concrete analysis plan (ideally a complete analysis script) that they need only apply to independent data during the confirmatory stage. This may improve the quality and positive impact of preregistrations as it leads to fewer deviations from preregistered analysis (this preprint by Sarafoglou et al. may be relevant in this context: <https://psyarxiv.com/6dn8f/>). Therefore, I am sympathetic to the conclusion that a two-stage procedure would benefit (neuroscience/psychological) research. Nevertheless, I believe the manuscript has left some questions unanswered.

1. The authors argue that the procedure may solve the dilemma between either performing a confirmatory study or sacrificing resources on an exploratory study. It is not exactly clear to me how the dilemma is solved. Isn’t the two-stage procedure essentially a two-study procedure with a pilot study (where researchers use resources to do exploratory analyses) and a confirmatory main study? In other words, what is really new about the proposed procedure? I suggest the authors work out more clearly and explicitly how the procedure differs from a classical pilot-study design and how it requires fewer resources during the exploratory stage.
2. Relatedly, what should be my criterion during the exploratory stage? Let’s say I perform a t-test and I test a point null hypothesis against a bunch of alternatives. For all of them, I find evidence in favor of the null – for some stronger than for others. Which hypothesis should I preregister in Stage 2? The worst alternative, which (based on my data) is my best shot at finding strong evidence in Stage 2? The best alternative, which may be a more critical but quite likely also a more resource-demanding test of the null hypothesis? The authors note some restrictions on the “everything goes” on p. 5, but they seem to assume that the researcher always has two competing hypotheses (and not a bunch of competing possible alternatives). So, to what extent should hypotheses be fixed already in Stage 1? Is it even necessary to test hypotheses at that stage or could one start with an estimation approach, and put the resulting posterior to the test in Stage 2? To summarize this point, I believe that the manuscript in its current form may leave readers wondering what exactly they can/should do in Stage 1 and how Stage 1 results should influence how they analyze data in Stage 2. More detailed information on the practical application of the proposed procedure, and possibly a concrete example, would be helpful.
3. I am wondering, on a philosophical level, whether the “anything goes” during Stage 1 might foster what Lin et al. (2021; <https://doi.org/10.1177/1745691620974773>) refer to as a “mutual-internal-validity problem”. Researchers may use the exploratory stage to tailor their hypotheses to the studied paradigm, population, data-preprocessing strategy (and vice versa). Even if the procedure achieves its goal and leads to better preregistrations, fewer deviations from preregistered analyses, and, eventually, higher direct-replication rates, it may have a negative impact on conceptual-replication rates simply because the tested hypotheses are tied to specifics of the study/analysis design. I may be overstating this negative impact, but I believe it is a potential caveat that deserves some discussion.
4. On p. 4, the authors state that the design is “efficient in both stages”. This is certainly true for Stage 2 when a sequential test is employed and can be contrasted with a fixed-sample test, but I wonder how efficiency is evaluated in Stage 1. Without a concrete termination criterion, it is hard to compare whatever strategy the authors have in mind for Stage 1 (see Comment 2) to another strategy. For example, my strategy could be to sample N = 20 individuals in Stage 1. A competing strategy to find a Bayes factor of a certain strength may well be much less (or more) efficient – however, without a common criterion, the comparison wouldn’t make much sense. So, how can we say that the design is “efficient in both stages”?

To summarize, I believe that the two-stage procedure proposed in this well-written manuscript may indeed be a “valuable addition to the methodological toolbox”, especially by creating room for and clearly distinguishing between exploratory and confirmatory analysis. However, I feel that some important questions have not yet been addressed, and that doing so in a revision will improve the manuscript.

**Author Response**
Oct 12, 2022

See file “v2\_response\_letter.pdf”

**Editor Final Decision:** Accept

Oct 21, 2022

Dear Dr. Stefan,
I am glad to inform you that Reviewer #1 was satisfied with the new version of the manuscript and, therefore, I deem it suitable for publication as it stands.

Best,
MTL

# Reviewer 1

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  |  | ✔ |

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The authors responded to all my comments appropriately and brilliantly.
I think that now this nice paper is ready for publication.

Gianmarco Altoè