Dear Dr. Liuzza,

My co-authors and I are happy to submit a revised version of our manuscript “A Two-Stage Bayesian Sequential Assessment of Exploratory Hypotheses” for publication in *Collabra: Psychology.*

We would like to thank both you and the reviewers for your positive feedback on our manuscript. In response to the suggestions for improvement we added an application example from the field of social neuroscience that illustrates the proposed procedure in practice, and we considerably extended our discussion section.

Below, we address each issue in turn, and note the changes that we have made to the manuscript. The action letter and reviews are in italics, our responses are in normal font. Within the revised manuscript, our revisions are in blue. We hope that these revisions meet the expectations you have set out in the action letter, and we would again like to thank you for the opportunity to revise our manuscript.

We look forward to your comments.

Kind regards,
Authors

Comments by the Editor

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.

[snip]

> **EDITOR-1:** We thank the editor for the positive evaluation of our manuscript. We have incorporated the suggested changes to the best of our ability and provide a detailed response to the points raised by the reviewers below.

Comments by Reviewer 1 (Gianmarco Altoè)

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.

> **R1-1:** We thank the reviewer for the kind words on our manuscript.

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.

> **R1-2:** We agree with the reviewer that these are important advantages of our proposed method that need to be emphasized more. We now mention that the proposed procedure “facilitates the specification of precise hypotheses and their translation to statistical models” on p. 9 of our revised manuscript and have added a corresponding phrase to the abstract. To emphasize that information from the first stage of the study can be integrated in prior distributions on all parameters in the second stage, we have reworded the first paragraph on p. 5 of our revised manuscript to say “information from the exploratory stage can be used to formulate

prior distributions on all model parameters in the confirmatory stage”. We also use the plural “prior/posterior distributions” to stress that information can be transferred for all parameters.

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.

R1-3: We thank the reviewer for this valuable suggestion. We added a new paragraph on p. 10 of our revised manuscript that describes our thoughts on the matter. In brief, in our view, the ideal publication format for a study conducted in the proposed design would be a registered report. Our reasoning is that the registered report format closely mirrors the two-stage design, such that the structure and goals of the publication format and study design are very well aligned. However, especially in the case of complex and/or extensive exploratory analyses, we agree with the reviewer that a separate publication of the first design stage may be beneficial. Potential publication formats could be exploratory reports (as mentioned by the reviewer) or study protocol papers, as used in the context of clinical trials. Whether a separate publication of stage 1 results is sensible, depends on the extent to which the exploratory results can make an independent contribution to the literature. On the one hand, a separate publication can highlight important contributions that might otherwise be overlooked, on the other, “salami slicing” of research results (the artificial compartmentalization of research projects for publication) should be avoided.

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.

R1-4: We agree with the reviewer that this point needed clarification. Our original formulation arose from the notion that frequentist inferences are invalidated by undisclosed multiple comparisons because error rates can no longer be controlled. Undisclosed multiple comparisons are impossible to avoid in exploratory analyses because the number of analyses is not predefined. For example, researchers may stop earlier if they discover promising trends or they may conduct further tests if their initial analyses are inauspicious. There is some debate about whether multiple comparisons affect Bayesian inference in a similar way (for a discussion, see de Jong, 2017,psyarxiv.com/s56mk/). Essentially, while optional stopping is unproblematic in the Bayesian framework (Rouder, 2014, doi.org/10.3758/s13423-014-0595-4), conducting multiple different hypothesis tests may still lead to multiplicity problems. We therefore agree with the reviewer that it is better to remove the term “(frequentist)” from the sentence. To make our argument even clearer, we replaced the phrase with “sacrifice resources to an initial exploratory study that does not allow for hypothetico-deductive inference (Jebb et al., 2017)”. We believe that the term “statistical inferences” may otherwise be interpreted too broadly to include also, for example, inferences from a visual examination of data or descriptive summary statistics, both being accepted methods of exploratory analysis.

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).

R1-5: We agree this is a good idea, and we have added an application example to our revised manuscript (pp. 5-8). The example illustrates a fictitious research scenario from the field of social neuroscience where researchers aim to identify the neural correlates of perspective taking. The exploratory stage is used to clearly define the region of interest in the parietal cortex, and the initial results are subjected to a confirmatory test in the second stage. We describe potential choices researchers need to make in both stages, and briefly discuss the operating characteristics of the design. Our online appendix (https://osf.io/z3ckm/) contains reproducible code as well as further details on our simulation setup.

Comments by Reviewer 2

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.

> **R2-1:** We thank the reviewer for the positive evaluation of our manuscript. Indeed, we would expect that a prespecification of analysis plans based on the exploratory analyses in the first stage of the design would lead to fewer unplanned adjustments in the confirmatory phase. We now mention this in two instances. First, we cite Sarafoglou et al. (2022) on p. 3 of our revised manuscript, where we first mention the risk of deviating from preregistrations if researchers are unfamiliar with the data environment. Second, we mention a reduced risk of deviations from preregistrations among the advantages of the Bayesian two-stage design on p. 9 of our revised manuscript.

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.

> **R2-2:** We thank the reviewer for raising this interesting point of discussion and we have added a new paragraph with a clarification on p. 9 of our revised manuscript. Indeed, the proposed design is very similar to a classic pilot study design where an exploratory pilot phase can be used to gain experience with the materials, procedure, and data environment, and a confirmatory trial is conducted in a second phase. The proposed procedure could be viewed as a modernized version of the classical pilot study design that capitalizes on the flexibility of Bayesian inference. In our view there are two main features that distinguish the proposed design from a classic pilot study design, both leading to increased efficiency. First, data collection is conducted in a sequential manner in both stages of the proposed design, leading to substantial efficiency benefits over "classic" pilot study designs where sample sizes (at least in the confirmatory stage) are fixed. Second, pilot study data are not discarded, but are used to enrich the tested hypotheses in the confirmatory phase. One could claim that this also happens in a "classic" pilot study design where the focal hypothesis test is selected based on the pilot study data (e.g., by selecting a certain dependent variable). However, the proposed design goes a step further by transferring the posterior information on all parameters from the exploratory phase to the confirmatory phase, thus making the hypothesis test in the confirmatory phase more informative and (potentially) more efficient with regard to expected sample sizes.

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?

> **R2-3:** We thank the reviewer for raising this interesting question. We agree that the described scenario, where all analyses conducted in the exploratory stage point towards the null hypothesis, is challenging to handle in the proposed design. We already mentioned this among the limitations in the original version of our manuscript (p. 5) and now expand a little more on it on pp. 10-11 of our revised manuscript as well as in our new application example section (pp. 5-8).
>
> In principle, we believe that the selection of analyses in the exploratory stage should be informed by substantive and technical considerations. For example: What model comparison would be the most severe test of the underlying theory? Would it be interesting to show evidence for/against the existence of an effect in a certain subgroup? Are assumptions of certain statistical models fulfilled? Is the data quality sufficient to run certain analyses? Strength of evidence is never the only criterion that sets different analysis options

apart. We now stress the importance of these substantive considerations more in our discussion on p. 11 of our revised manuscript.

In the particular case described by the reviewer, one important consideration is whether the consistent evidence for the null hypothesis would change a researcher's focus towards confirming the observed null result. If this is the case, the important substantive consideration becomes: How can I best show that the effect is *not* there? This may, for example, lead researchers to choose the analysis pathway that showed the most evidence for the alternative (the "worst alternative" in the reviewer's terms), thus potentially giving the alternative hypothesis the best possible shot. It could also lead researchers to focus on the whole population rather than only a subgroup in an effort to disprove the existence of an effect more generally. We now mention this scenario on p. 10f of our revised manuscript. However, the researcher's inferential goal does not need to change if tests in the exploratory phase show consistent evidence for the null. For example, researchers interested in whether a new therapy is effective may still decide to focus on the subgroup with the best outcome, regardless of whether the evidence pointed towards the null hypothesis, too, in this group.

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?

R2-4: We thank the reviewer for these thought-provoking questions. As it may have become clear from our response to **R2-3**, we do not assume that researchers have already narrowed down the field to two competing hypotheses or models. We now make this clearer on p. 3 of our revised manuscript where we added that researchers can "explore a variety of different [...] statistical models and hypotheses" in the exploratory phase. Of course, it will depend on the research environment to what extent models and hypotheses are still in flux in the exploratory stage. For example, researchers might simply be interested in exploring effects of different stimuli or data preprocessing steps in the exploratory phase, but they might not want to change the statistical models applied to the (preprocessed) data. We believe that our application example (pp. 5-8 of the revised manuscript) provides a good illustration of this case.

With regard to hypothesis testing vs. parameter estimation in the exploratory stage, we believe that researchers should use all their freedom to explore the data in the exploratory phase, and this includes pure estimation approaches. Indeed, the caveats formulated on p. 10f of our revised manuscript do not restrict the types of analyses researchers may conduct in the exploratory phase, but simply caution researchers not to throw general due diligence overboard in the interpretation of results and the (re-)use of data. Essentially, while anything is allowed in the exploratory phase, choices made based on these analyses have consequences, both statistically (e.g., prohibiting the re-use of data) and substantially (e.g., sharpening the tested hypotheses).

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.

R2-5: We thank the reviewer for this insightful suggestion. As mentioned earlier (**R1-5; R2-4**), we have now incorporated an application example from the field of social neuroscience in the revised version of our manuscript (pp. 5-8). We believe that this example provides a realistic illustration of the procedure in an application domain that could potentially benefit from the two-stage procedure. The example concreticizes the steps outlined in the theoretical part of our manuscript and demonstrates their implementation in practice. We also provide documented code in our online appendix (https://osf.io/z3ckm/) that may facilitate the uptake of our proposed methodology.

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.

R2-6: This is a very interesting question. We agree that psychology research tends to disregard external validity and we would find it worrisome if our proposed methodology was fostering the mutual internal validity problem described in Lin et al. (2021). If we understand it correctly, the primary concern of the reviewer is that exploratory analyses take place and inform the confirmatory stage. We would like to stress that exploratory analyses are not novel to the field of psychology. The only novel aspects of our proposed approach are that the exploratory analyses are (1) explicit, and (2) can be used to inform analyses in the confirmatory stage in a mathematically coherent way. Thus, to find out if our proposed methodology aggravates the mutual internal validity problem, we need to evaluate these two aspects. Personally, we believe that making exploratory analyses explicit may be beneficial, rather than detrimental. If exploratory analyses are explicitly reported, they showcase a broader selection of possible analyses and associated hypotheses and may therefore widen researchers' views on the tested theory and potential operationalization of constructs. As to the second aspect, the carry over of information, we agree that this may make hypotheses and analyses more context-specific. However, we do not claim that these specific analyses should be re-used for different studies. Whether this is sensible depends on the substantive research context. We also do not claim that our framework is tied to a one-shot design. Lin et al. (2021) claim that external validity can be achieved through triangulation. However, in our understanding, theoretical triangulation needs to happen within a research field, rather than within a single study. Therefore, we believe that researchers need to carefully consider internal and external validity when choosing an analysis pathway within our proposed study design, but that this is no different from any other study design. We now mention this on p. 11 of our revised manuscript and include a reference to Lin et al. (2021).

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"?

R2-7: We agree with the reviewer that "efficiency" is less clearly defined in the first stage of our proposed design than in the second stage. We therefore deleted the phrase "in both stages" from our manuscript to avoid misconceptions. In general, the first stage allows researchers to stop sampling whenever they have found a promising analysis method. This means efficiency in this case can be defined as obtaining necessary information about the (subjectively) best analysis pathway with as few observations as possible. Since the phase allows researchers to stop sampling whenever they choose, the stage is as efficient as it can be given the preferences (e.g., regarding minimum sample sizes) and skills (e.g., quick recognition of potential analysis pitfalls) of the research team. However, the reviewer is right in that it would be difficult to compare the efficiency of this phase to an alternative design. This is due to the fact that the goals of researchers (i.e., their definitions of "necessary" information for deciding upon the analysis pathway) may differ a lot between researchers and research scenarios.

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.

R2-8: We thank the reviewer for the positive evaluation and the constructive feedback. We have implemented the suggestions to the best of our ability, and we are confident that the changes have improved our manuscript.