Responses to reviewers, 2

We again wish to thank the reviewers for their informed and helpful criticism and suggestions. The reviewers agree that their previous concerns have been addressed, but note further aspects of the manuscript requiring optimization, particularly in how we lay out the results of our simulation study. In the following, we discuss how we have integrated their suggestions.

Reviewer 1

- The revision addresses the concerns expressed. The manuscript is poised to make a very nice contribution to the research community though I still find some points of confusion in this version.

We thank the reviewer for their encouragement, and for suggestions that hopefully have enabled us improve the manuscripts' clarity.

> Statement of Significance - In the first line here I think "regarding" should be changed to "used to assess"
>
> Pg 1, Abstract - The last sentence currently suggests testing is always in appropriate. I recommend adding "for assessing nuisance effects" after "philosophically".
>
> Pg 2, Section 1.1 - Line 9 of the first paragraph here should refer to the "related also incorrect intuition"
>
> Pg 3, near end of Section 1.1 - I'd recommend changing "The preferred solution" to "A preferred approach"

We agree, thank the reviewer for these observations, and have performed the suggested changes.

> Pgs 3-4, Section 1.2 - I found this section very confusing. The key point to me is that the analogous situation (concern about confounds) exists in clinical trials. It is sometimes dealt with there through randomization. In that case people sometimes do significance tests to assess whether the randomization worked appropriately - that community accepts this at times but has generally agreed

> it is not appropriate. When randomization is not possible, the clinical trials community has come up with other strategies (blocking, matching) that could be considered here.

We have reformulated and streamlined this section. It is now reduced to the references to the literature, and a very concise discussion of the key points.

Reviewer 1 and 2 on the simulation

Written originally because we agreed with Reviewer #2 that a quantification of the likely impact of the described procedures would be very helpful, both reviewers agree that the simulation – as presented in the previous version of the manuscript – is not exposed optimally. Reviewer #1 questions the general benefit of including a simulation analysis. Reviewer #2 discusses how in its current form, it is inadequately presented.

> Reviewer 1 pgs 5-6, Section 3 - I have two concerns about this section. First, I'm not convinced that it adds very much to the story you are telling. Thus it may not be necessary at all. Second, if there is to be a section on simulation, then the simulation methods need to be described much more clearly. I don't know what is meant by the "measured size of the confounding factor" and how you are addressing the correlation. Then I'm not clear on what your simulation procedures is doing. Help!!
>
> Reviewer 2 I do have an idea for an alternative, simpler way of conveying the results. It seems to me that the two most useful things to know from the simulation are (copy/pasted from my initial review): "What exactly is the statistical power to detect differences in confounding variables with different stimulus sample sizes and confounder effect sizes? And given this low degree of power, if one does rely on NHST for deciding whether to control for confounders, then what is the expected Type 1 error rate for rejecting the null of no difference on the focal/treatment variable when in fact the difference is entirely due to differences in the confounding variable?" So one idea to make the simulation results more clear and comprehensible is to – at least in the paper, although not necessarily on the app page – remove all of the other results and info, and instead only present the results for those two things as a function of the parameters varied in the simulation. If the authors really feel that all the additional info should be presented in the paper itself, then I wouldn't fight them on it, just as long as those results can be clarified a bit.

In response to this, we have decided to reduce the simulation aspect of the manuscript to its bare essentials, while referring via a web link to the full results, and the online application where the full simulation can be assessed and

manipulated. At this online site, the details of the simulation are exposed in great detail in the form of a well-documented online app, including code, the precise simulation procedure, and a demonstration of outcomes. We also note that the app is much more detailed and more clearly documented than the code which originally generated our simulation results; it more clearly addresses the question of interest, and can be employed by the user to investigate a broad spectrum of configurations.

We think that a detailed discussion of the simulation is beyond the scope of the manuscript, and such an interactive online presentation is much better suited. We do however think that this online app, and a reference to it, can help readers understand the nature of the problem, and for those readers who are not interested, the reference to the online app takes up little space.

Regarding a specific point, reviewer #2 has indicated that it might be helpful to present the power of stimulus confound inference testing. We have considered this, but eventually decided to not report it in the manuscript (although it can be readily simulated with the app), for the reason that it might confuse some readers regarding the power of what hypothesis test is specifically estimated; it is, after all, the power for a test that, so we argue, has only illusory relation to what researchers might be truly concerned with. It is not, after all, the power to detect a real confound! Our primary argument is not that the test has low power to detect real confounds, but that any of its error rates do not refer to the actual question the researcher is interested in (but to inference about a population the researcher is *not* interested in: the not tested stimuli). We think it is best to avoid this potential source of confusion. However, the rate of rejected stimulus sets for various stimulus set sizes and differences can be readily simulated with our app. Similarly, the rate of failures to detect "false positives" that are due to undetected stimulus confounds (as also requested by reviewer #2) can also be rapidly visualized with the app.

We thank the reviewers for their further encouragements, comments and criticism, again helping us in sharpening the focus of the manuscript. We hope the reviewers agree that further downsizing the manuscript was the appropriate way of dealing with their concerns, as it more clearly highlights the – uncontroversial – main questions.