Response to recommender

Dear authors

We regularly triage Stage 1 submissions before sending them out to expert reviewers to ensure various criteria for RRs are met. Your submission is already in a great shape but there are a several smaller issues that I thought merit fixing to avoid confusing reviewers.

OSF link

Please ensure that when you submit that the OSF link points to the manuscript directly, not the general OSF project. If you change or update the manuscript, it will update the link so the link may then be broken. This issue occurred in your previous submission - our team was able to salvage the correct link but this was only by luck. **Please ensure that the link to the manuscript works and points to the latest version when you submit.**

- Thank you for this cautionary note. We have accordingly made sure that the link is functional and points to the latest submission.
- The link to the latest version of the pre-registration is the following : https://osf.io/up743
- The link to the latest version of the supplementary material is the following : https://osf.io/ht9kb

Statements precluding outcome

Your manuscript is somewhat unusual for a Stage 1 RR in that there are several statements that seem to preclude the outcome. In fact, you have a whole *Discussion* and *Conclusions* section. These are fine because they can be replaced at Stage 2 (only Intro and Methods and Design is set at Stage 1). However, the second-to-last sentence in the *Introduction* also could be seen as precluding the outcome: "*Furthermore, we argue that the initial QA/QC on unprocessed data of neuroimaging studies must be critically carried out before defacing to avoid these biases*".

I realise that this is based on your pilot data and that you have a strong expectation that you will confirm those earlier results. Nevertheless, the results should not yet be known at this stage. Based on your description currently I judge the bias control level of this project to have a relative high risk Level 3 or 4 (see section 2.6 in the Guide for Authors) but your plan to use blinded, randomised rating should help mitigate this. Nevertheless, I advise you to be more circumspect in your expectations. You can certainly describe your expectations but in a way that requires no further changes to the Intro at Stage 2 if your results show the opposite.

- We thank the recommender for raising this concern and have deleted the last sentence of the abstract and the last two sentences of the introduction, which were not done in speculative terms, thereby precluding the outcome.
- We moved some speculative statements of expectations for the results from the discussion and the conclusion into the supplementary materials, and clearly indicated them as speculative. We also adapted the conclusion of the main paper to give it a more neutral tone and refer the reader to the drafted discussion in the conclusion (Page 9, ll. 193-194 and ll. 198-199).

> *This study is proposed to investigate whether manual and automatic aspects of QA/QC implemented in MRIQC are biased by the process of defacing data.*

> *Finally, a discussion has been included within the supplementary material, speculating the impact of this study should the hypotheses be verified.*

Why only 3T data?

You say you will only use the 3T for the manual rating. There are probably good reasons for that but I would suggest explaining them.

- We thank the recommender for this suggestion and have added an explanation to why we keep only the 3T site for the manual rating (Page 4, ll. 110-114).

> *This choice responds first to eliminate the field strength and other variability sources emerging from the specific scanning site. Second, images acquired with the 3T scanner are expected to showcase better signal-to-noise ratio (SNR), and likely yield better quality assessments on average by human raters independently of the defacing condition.*

Hypotheses 1 and 2

To my reading, the first two hypotheses are really part of the same. In RRs it is particularly useful to condense the preregistered plan down to the simplest statistical comparison (1-df test) necessary to answer the research question. In your case this seems to be a one-tailed paired t-test or non-parametric alternative on ratings between defacing statuses, plus your Bland-Aldman plots. Is the ANOVA/LMM analysis in Hypothesis 1 adding anything to that? If so, please explain.

- We thank the recommender for raising this concern and have updated the manuscript accordingly. First, we have merged the first two hypotheses into one. Now, the questions whether there is a bias, and that of the direction of the bias are separated under a single hypothesis in the study design template (Page 3, ll. 78-81 and Page 10-11).

> *1.Defacing influences trained experts' perception of quality, and their ratings will significantly vary between the defaced and the non-defaced conditions; Besides, because there is less information in the image after the removal of facial features, raters will assign more optimistic (better, on average) ratings in the defaced condition than in the corresponding non-defaced condition; and*

Hypothesis	*Question*
Defacing influences trained experts' perception of quality	*Do the quality ratings from human raters significantly vary between the defaced and the non-defaced conditions?*
	Are ratings in the defaced condition more optimistic (better, on average) than the corresponding ratings on the non-defaced condition ?

- We also thank the recommender for the suggestion that the simplest appropriate test should be used. We now make it more clear that the bias on the raters' perception is our effect of interest, and we have added a detailed justification in the paper as follows (Page 5, ll.134-141 and Page 1, ll.16-22):

> *We will test the influence of the defacing condition and the rater (within-subject factor variables) on the ratings (dependent variable) using rm-ANOVA, or linear mixed-effects models in case data do not meet rm-ANOVA's assumptions. As opposed to multiple t-tests, rm-ANOVA and linear mixed-effects models enable to disentangle the variability coming from the raters and the variability coming from defacing and to quantify the latters. Indeed, because we do not necessarily expect the ratings distribution of each rater to have the same mean, rm-ANOVA and LMM account for the baseline difference in ratings by adding the rater as a random effect in the model.*

> *By means of repeated-measures analysis of variance (rm-ANOVA), or linear mixed-effects models in case data do not meet rm-ANOVA's assumptions, we will determine whether four trained human raters' perception of quality is significantly influenced by defacing by comparing their ratings on the same set of images in two conditions "non-defaced" (i.e preserving facial features) and "defaced" (N=185 images per condition). Relatedly, we will also verify that raters are more optimistic about quality in the defaced set.*

Inconsistent power analysis

For a project like this, determining the minimal effect size for a prespecified power and alpha level makes sense. However, this seems to be inconsistently applied. For example, Figures 3 and 6 mention an alpha=0.02 but in the text and the Design Table the same power analyses are described as alpha=0.05. Moreover, it would be worth mentioning the power level in the text, not only the figure captions. Note that some RR-friendly journals expect an alpha=0.02 - if you plan to submit your final Stage 2 manuscript to one of these journals this is indeed the threshold you should set.

- We thank the recommender for spotting this important inconsistency and have accordingly adapted the significance threshold for the p-values to 0.02. Additionally, we have added the power level in the text reporting on the sensitivity analyses (Page 6, ll. 147-149 and Page 8, ll. 186-188) .

> *We determined using G*Power (Faul et al. 2009; see Figure 3) that with rm-ANOVA our experimental design can at worst identify effects of f=0.14 (i.e., a small effect) or greater with a power of 90%.*

> *We determined using G*Power (Faul et al. 2009; see Figure 6) that our experimental design can identify, with a 90% power, effects of f=0.16 (i.e., a small effect) or greater.*

Minor issues

- In first paragraph of *Introduction*: "...the ears themselves." The "themselves" doesn't seem to make sense to me (but I may be wrong, in which case ignore this comment)
- Figure 4: when describing the 95% confidence intervals I assume you mean "dotted" not "dashed" lines (the latter are the means)?
- Typo in Design Table, Hypothesis 1, Question: "bias" instead of "biases"
- Also in Design Table, all cells of Rationale column: reported "in" Figure

We thank the recommender for spotting those grammatical mistakes and have corrected them.