Revision round #2
Decision for round #2 : Revision needed
Invitation to revise Stage 1 RR
First, I sincerely apologise for the delay in returning this decision to you. I have been waiting for a promised review from one of the original reviewers, but this has not materialised. Because this reviewer's comments were extensive, I wanted to get his input on the changes, but it is just taking too long to achieve this. Indeed, I should probably have terminated the assignment earlier.

In any case, the revised manuscript has been evaluated by Reviewer#1, and I have looked over the changes myself. In general, I think that you have addressed the comments well, and in cases where there might be outstanding concerns about aspects of the design, you have been sufficiently explicit in your reasoning that readers will be able to evaluate these aspects transparently.

Therefore, I think IPA could be issued for this study, pending consideration of a few minor issues. These are the issues raised by Reviewer#1, particularly, **the critical issue of inclusion/exclusion criteria.**

In addition, I think that you should provide greater clarity on your thresholds of Bayesian evidence. You seem to configuring this study as a hypothesis testing experiment that will allow you to make theoretical claims. If so, then you should make it clear what threshold BF will be taken to confirm the alternative hypothesis (or the null).

At the moment, you refer to Lee & Wagenmakers (2013) scheme for describing levels of evidence. Here you state that "When BF10 equals 1-3,it is commonly inferred that the evidence for H1 over H0 is anecdotal; when BF10 is close to or equals 10 it means that the evidence is moderate; lastly, when BF10 is higher than 10 signifies strong evidence for the hypothesis of interest (Lee & Wagenmakers, 2013)."

It would be more accurate to say that BF10 3-10 corresponds to 'moderate'. You should make it clear that this is the specific scheme you are adopting ('commonly inferred' gives the impression that this is the most-often adopted scheme, which may not be the case).

Critically, as noted, you should state what your threshold level is to claim support for the hypothesis (assuming that this is what you intend to do). Your BFDA suggests that the threshold you are adopting is 10, and that your data could be expected to pass this threshold 78% of the time if the hypothesis is true. This is fine, but note that the choice of threshold of evidence (and sensitivity to detect that level of evidence), may influence the list of eligible journals for you to publish the paper in.

Of course, you are not obliged to place the final manuscript in any journal, but if this is your aim, then you should make sure from the outset that your design is compatible with your target outlet.

Again, apologies for the delay in this decision, but I hope you can move forward quite swiftly with your study from here. Given the time already taken, I see no requirement for further external review.

by Robert McIntosh, 18 Mar 2024 11:59
Manuscript: https://osf.io/7vtdk?view_only=59385b6256b5492791f6882705c20424
version: 7

Dear Prof. McIntosh,
Thank you for allowing us to submit the revised manuscript of our Stage 1 Registered Report for the second round. We appreciate the time and effort that you and the reviewer dedicated to providing constructive feedback on our manuscript that significantly improved our study. We have incorporated the suggestions made by you and the reviewer. Those changes are highlighted/marked as tracked changes within the manuscript. Before providing our point-by-point responses to the reviewer's comments, we would also like to address the concerns that you raised.

Following your recommendation, we have adjusted our descriptions of levels of evidence. The updated description now is as follows: "We adopted the scheme by Lee and Wagenmakers (2013) to describe the levels of evidence. When BF10 is between 1 and 3, it corresponds to anecdotal evidence in favour of H1 over H0; when BF10 is between 3 and 10, it corresponds to moderate evidence; lastly, when BF10 is higher than 10, it corresponds to strong evidence." (this excerpt can be found in our manuscript on page 20).

Thanks for your warning about the journals' requirement. We actually used the most conservative approach of BF10 = 10, in order to comply with the high requirements of journals such as Nature Human Behavior and Nature Communication. That said, we are aware that some other prestigious journals like Cortex demand BF10 = 6. Therefore, we believe our calculation will fulfil the demands of both groups of journals.

Review by anonymous reviewer 1, 06 Dec 2023 11:09
In the revised version of the Registered Report, the authors have significantly improved the quality of the manuscript and have well addressed my comments. I believe that the changes made to the experimental paradigm and analyses make the paper easier to follow and would provide a more in-depth exploration of the research question.

We would like to thank the reviewer for their positive feedback and constructive comments. Below we will try to answer the outstanding concerns of the reviewer.

I have some outstanding points for the authors, which I think should be addressed before performing the study:

- The authors now include HHb within the analyses, but I think some information is still missing in relation to how this signal will be used in the GLM. For example: "The highest coefficient of HbO for each condition on each region of interest (ROIs; bilateral parietal regions) will be used to test whether CP-knowers exhibit higher bilateral parietal activation, defined by increased HbO/decreased HHb, specifically in the left parietal region, relative to subset-knowers (Hypothesis 1). " This analysis doesn't account for HHb, as one would need to select the lowest coefficient of HHb for each condition on each region of interest to examine whether a significant decrease in HHb is present. If this is correct, the authors should carefully check their analyses throughout.

We apologise for the remaining ambiguity of using HHb in the analyses and hypotheses. We have elaborated on the related parts of the manuscript as follows: "The highest coefficient of HbO or the lowest coefficient of HHb for each condition on each region of interest (ROIs; bilateral parietal regions) will be used to test whether CP-knowers exhibit higher bilateral parietal activation, defined by higher HbO or lower HHb, specifically in the left parietal region, relative to subset-knowers (Hypothesis 1). Note that neurophysiologically, increased HbO or decreased HHb represent increased brain activation (Sholkmann et al., 2014)." (this excerpt can be found in our manuscript on page 18-19).

We further use HHb changes for the quality check. We added this point to the revised manuscript:

"Therefore, we must expect opposite directions in HbO and HHb changes, in a way that if one goes up, the other goes down (Sholkmann et al., 2014). Accordingly, we will check HbO and HHb changes in each channel. The channels will be excluded if both HbO and HHb significantly increase or decrease as they are most probably contaminated by uncorrected artefact."(this excerpt can be found in our manuscript on page 19).

- I think that the inclusion criteria of "a minimum of one clean channel per ROI and per participant for inclusion" is too lenient. I fully appreciate the challenges associated with testing infants and children, but it would be important to ensure that any activation found is a true reflection of a genuine response. I believe that in the developmental fNIRS literature, a common objective criterion is to exclude a participant from further analyses if > 30% of all channels had to be excluded (e.g. due to weak or noisy signal). It is also quite common for studies to not consider single isolated channels further in the analysis. This is because a channel is deemed reliably active if at least one spatially contiguous channel is also significant (see for example Lloyd-Fox's fNIRS studies). So I think it would be best to change the inclusion/exclusion criteria accordingly (or at the bare minimum consider any isolated channel response with caution). You might also want to make sure that each channel contains valid data in both conditions to be included in further analysis.

We acknowledge the concerns regarding the inclusion criteria. While some fNIRS studies may have used an arbitrary threshold of 30% for participant inclusion, the field is yet to find a generally agreed pipeline for preprocessing (e.g., FRESH project, the recent multi-site fNIRS project led by Yucel, Luke and von Luhmann, that our lab contributed as well). This is more challenging when it comes to developmental studies, due to shorter experiments, higher movement artefacts, etc. that leads to much lower SNR. We systematically searched fNIRS studies in preschoolers, and surprisingly most of the studies did not clarify what constituted noisy data and what was their exclusion criteria (e.g., Perlman et al., 2014; Ding et al., 2023; that are from established labs using fNIRS in preschoolers). Other papers mention excluding channels without specifying the number of channels, or excluding participants based specifically on the quality of data (e.g., Li et al., 2021). Finally, some papers mention nothing about data quality check at all (e.g., McKay et al., 2022; Xie et al., 2021; Barreto et al., 2021), and none were found to report the number of included channels

per ROI. As open research advocates, we urgently believe that inclusion of that information is vital, however, we are mindful that taking restricted inclusion criteria might lead to a huge discard and accordingly low-powered and non-replicable findings, especially in the field that does not yet have a practice of reporting actual quality of data that goes into the analysis.

Please also note that although some infant studies are more informative about its data quality (e.g., Lloyd-Fox et al., 2010; Blasi et al., 2007), the insights from infant or school children studies cannot be readily extended to studies in preschoolers for two reasons: variability of the hemodynamic response during the childhood (Gemignani & Gervain, 2021) and the very different challenges associated with each age stage (e.g., the biggest challenge in infant studies is ensuring they look at the screen, whereas the biggest challenge in preschoolers studies is to ensure they do not move too much and are happy to wear the cap).

However, we agree with the reviewer that the quality of data should be rigorously controlled. According to our literature search and a very relevant fNIRS study in preschooler that came out meanwhile, we applied the following modification that we hope would be convenient for the reviewer. We suggest our solution to the dilemma (please find these points reflected in page 17 of the revised manuscript):

1. First and foremost, we will report the quality of data (quantified by the number of channels per ROI included) for both groups and individually for each participant. Similarly, each analysis will contain information on the included number of participants, as well as how many of these participants had only one channel per ROI included for transparency. Similarly to Ding et al. (2023), we will exclude participants separately for each analysis. This allows us to not entirely discard datasets that could be included in one analysis but not in another, which leads to better SNR. For instance, when a participant has several good channels in the left parietal area, but not in the right; instead of completely discarding the participant, they can be included in the first hypothesis testing about activity in the left parietal area, but not in the second hypothesis testing about the bilateral parietal connectivity.
2. One of the rare fNIRS studies in preschoolers that applied QT-NIRS was published recently by Bulgarelli et al. (2023). While they also do not provide information about the included number of channels per participant/on average per group, they were the first in the field who applied QT-NIRS data quality checking on a preschoolers' fNIRS data, to the best of our knowledge. As they have set a precedent for future papers, we decided to follow their suggestion and slightly updated our QT-NIRS parameters accordingly: SCI = 0.6; Qth = 40%; PSP = 0.06.
3. We have also added an additional data quality check that was used as the only method by Xie et al. (2022). This includes considering the recording software for the data collection (Oxysoft) that allows checking the quality of data during data collection. Thus, we started data recording in each child once at least medium quality is reached in at least one channel per ROI. The quality was then monitored throughout the entire experiment and corrected if needed during the rests.

To conclude, even though we think it is better not to restrain further the inclusion criteria, our solution eliminates most of the situations where only one channel per ROI is viable. As for the cases where such outcome cannot be prevented for various

reasons (e.g., bad coupling between the scalp and the optodes), such cases will be reported in a table, and will be perceived with caution. We believe that by applying these additional measures, we will comply with the reviewer's recommendation, as well as be the first study in our field to provide extensive information on the quality of data, which we hope will lead to more studies being more transparent about the quality of data as well.