Response to Reviewers

HANDLING EDITOR
Dr. Dorothy Bishop

Assessment of reliability

1. The question of how to assess reliability has raised questions by reviewers 1 and 2: given that even experts are finding this confusing, my recommendation is that you could take the opportunity here to expand the introduction a little to say a bit more about how the different methods work. You do cite excellent primers – I was pleased to be pointed to the papers by Liljequist et al (2019) and Noble et al (2021), but I think that a point that does not come across, and can be confusing, is the key role of systematic change across sessions. I suggest you don't mention the six ways of computing ICC: as Liljequist notes, most are irrelevant for the current context. An easier way to explain the ICC is to say that in a test-retest design it is equivalent to the Pearson correlation if there are no systematic differences between the two sessions (i.e. no practice effects). The larger the mean differences between sessions, the greater the difference between ICC and Pearson r, with the ICC being lower. If introduced this way, then I think it may be easier for people to grasp the point about between-subjects and within-subjects sources of variance, as most readers will be familiar with the idea that r is affected by restriction of range
    a. We appreciate this note as out intent is to ensure the reader is on the same page.
    b. First, to alleviate confusion in-text we have revised
        i. *'In the context of multi-session data, there are several ways to estimate an ICCa, but for typical univariate fMRI studies, two specific types (ICC[2,1] and ICC[3,1]) are recommended(For a discussion, see Noble et al., 2021)*
    c. Second, we added a brief paragraph,
        i. The ICC is a statistic adopted from behavioral research to estimate reliability of observed scores across measurement occasions (Fisher, 1934; Bartko, 1966; Shrout & Fleiss, 1979). In the context of multi-session data, there are several ways to estimate an ICC, but for typical univariate fMRI studies, two specific types (ICC[2,1] and ICC[3,1]) are recommended (For a discussion, see Noble et al., 2021). As described in (Bennett & Miller, 2013; Fisher, 1934), the ICC is similar to the product moment correlation. Unlike the product moment correlation, which estimates separate means and variances between distinct classes (e.g., age and height), the ICC estimates the mean and variances within a single class (e.g., measure). For two or more variables from a single class, test-retest reliability estimates the consistency (or agreement) of the observed

scores across the measurement occasions. Using the correlation coefficient as an example, if there are no differences in subjects' scores across two measurement occasions, the correlation coefficient would be 1.0. However, if the measure is affected by systematic and/or unsystematic error across measurement occasions, this would impact the covariance between observed scores across subjects and decrease the linear association between measures across the two occasions. Unlike the product moment correlation, however, the ICC factors out measurement bias which reflects the reproducibility of observed scores across measurement occasions (Liu et al., 2016). While the correlation between two occasions (A = [1, 3, 6, 9, 12] & B = 3xA = [3, 9, 18, 27, 36]) may be perfect (rAB = 1.0), the consistency in observed scores between the two measurement occasions would be lower (ICC[3,1] = .60). In fMRI, the reliability of the blood oxygen-level dependent (BOLD) signal may be impacted by biological (e.g., differences in BOLD across brain region), analytic (e.g., task design and analytic decisions), and participant-level factors (e.g., practice effects, habituation and/or development). These fluctuations, whether typical or atypical, may contribute to observed differences and the reduced consistency in scores across measurement occasions, leading to decreased estimates of reliability.

2. Reviewer 2 makes the point that behavioural variability is also important. I would also add that, if we see between-subjects variation as important, then it is important to have a bit more information on the participants – do we have any information on their demographics? Are they very uniform in terms of social background, ethnicity, scholastic attainment? A few key details, including gender, and method of recruitment of participants, should be reported, even if they are not used in analysis, just to make it possible to compare with other samples. Again, with the two adult samples, some demographic information is needed. The "Neuropsychological Risk" cohort of MLS is to be used: how was that risk defined and how were they recruited?
    a. This is a good point. We should have made it explicit in the registered report that the demographic characteristics would be reported. Specifically, in-text, we now include:
        i. *The mean, standard deviation, count and frequencies will be reported for demographic variables from the ABCD, AHRB and MLS datasets. For ABCD, AHRB and MLS, participants self-reported on Age, Sex and Race/Ethnicity. ABCD: Sex is reported as sex at birth (Male, Female, Other, or Not Reported); Race/Ethnicity is reported on a 5-item scale: White, Black, Hispanic, Asian, Other. AHRB: Sex is reported as sex at birth (Male or Female); Race/Ethnicity is available on a 4-item scale: White, Non-Hispanc, Black, Non-Hispanic, Hispanic/Latinx, Other. MLS: Sex is reported as Sex at Birth; Race is available on an 8-item scale:*

Caucasian, African American, Native American, Asian American, Filipino or Pacific Islander, Bi-Racial, Hispanic-caucasian, and Other.

ii. *Behavioral data from the MID task will be reported as supplemental information, such as the mean and distribution of probe hit rate and mean response times (RT) across subjects. The task design is programmed to achieve a probe hit rate of approximately 60% for each subject. It should be noted that the RT for the probe is not consistently collected across the ABCD, AHRB, and MLS datasets.*

3. Given that the brain is changing rapidly during adolescence, one might expect greater between-subject variability in task performance for the ABCC sample than the young adult samples of MLS and AHRB. This is briefly alluded to on line 155, but it's not clear exactly what the prediction is here. (i.e. what is 'more the case'?).
    a. This is a very good question. Given the lack of evidence on these metrics across development, it is difficult to make precise predictions. As we responded to one review below (#19), we may have insufficient evidence to make explicit predictions about the relative changes in the variance components in *this* task and *these* samples. As a result, we have revised the Aim 2 hypothesis.

Thresholding

4. I can see that there is a limit to the number of factors that you can consider in this kind of analysis, but it did seem a bit surprising that you did not consider thresholding, which is likely to have a key effect. Reviewer 1 queried your choice of threshold. And like Reviewer 2, I wondered what we would learn from analysis of 'task irrelevant voxels'. If you do want to stick with the proposed analysis, it needs to be better justified. (I would also question whether being below threshold makes a voxel "task-irrelevant" – what about voxels that show significant task-related *deactivation*? – I'm not suggesting you do analyses of that, but am flagging up need for caution in use of language here).
    a. This is a very good point. There are, indeed, potential brain regions outside of the masks that may be task-relevant as they relate to deactivation. So labeling them as 'irrelevant voxels' is a poor choice of words. To address the misuse of "task-relevant" v "task-irrelevant", we revise in-text as follows:
        i. *For the continuous estimates of reliability described below, the analyses will be performed separately on task voxels that exceed and do not exceed an a priori specified threshold applied on the NeuroVault (Gorgolewski et al., 2015) meta-analysis collection that comprises the anticipatory win phase across 15 whole brain maps for the MID task (Wilson et al., 2018; Collection: 4258, ID: 68843). The suprathreshold voxels are those that exceed the threshold ($z > 3.1$) and the subthreshold voxels are those that do not exceed ($z < 3.1$) in the map. We acknowledge that the threshold of $z = 3.1$ is arbitrary (uncorrected, p-value = .001) and that the voxels that fall below and above this threshold may not be significantly different (Gelman & Stern, 2006). However, we made this decision to constrain the*

problem space in these analyses (Gelman & Loken, 2014; Simmons et al., 2011).

Aim 2

5. Like reviewer 1 I found the account of the MSBS and MSWS analysis hard to follow. If the ICC is positive, then MSBS will always be greater than MSWS. Won't the ratio between these be predictable from the ICC? I tried a little simulation, and for the range I looked at, they were monotonically related, though in a nonlinear fashion. I think if one reviewer and I find this confusing, so will other readers, so it is worth double checking this aspect. And, given that the ICC will be lower if there are differences in mean between sessions 1 and 2, it may be worth considering how far low ICCs are driven by practice effects.
    a. We recognize that the motivation for this ratio was less clear. Considering the comments from the reviewers, we have elected to remove the analysis that will report the ratio of the MSBS/MSWS from Aim 2.

Adaptive design?

6. It's great to see large datasets being used to address the question of reliability, but your plan involves a huge amount of analysis, and I wondered if it would make sense to adopt a more adaptive approach. For instance, suppose you did an initial analysis of data from 500 of the ABCC participants, and found that some of the analysis factors had no material effect on reliability. Rather than slogging on through the next 1500 samples, you might then decide to drop that variable – and perhaps substitute another (e.g. threshold).
    a. This is a great recommendation. It is worth to note, Aim 1 & Aim 2 would <u>not</u> rely on the full ABCD sample. As specified in-test, *For Aims 1 and 2, we use a* full sample from AHRB & MLS, but a "*subsample of ABCD participants at the University of Michigan site (site = 13)*". The sample size for Site=13 is < 500 and the likely number of test-retest subjects that pass behavioral/fMRIPrep QC are likely much fewer. However, Aim 3 proposed using N = 2000. Since this is a registered report, we had decided to use the maximum value of N = 2000 based on the stability information provided by Marek et al. (2022). However, we appreciate your input and have made the necessary adaptations. We have now included the methods and analyses in our report the following:
        i. Methods, Participants:
            1. For Aim 3, we use a subsample of N = 2,000 of the maximum clean data available from the ABCC sample and use an adaptive design to answer at which N ICC stabilizes. To reduce the use of unnecessary computational resources, the analyses are first performed in N = 500. If the difference between average ICC estimate for interval N_i & N_{i-1} is > .15, the sample will be extended to N = 1000, adding N = 500, until the plotted estimates are stable.
        ii. *Analyses, ICC, Aim 3:*

iii. *Aim 3: evaluates the sample size at which the ICC stabilizes. The chosen pipeline will be based on the highest median ICC for the suprathreshold task-positive mask from Aim 2 for the ABCD study. Based on this pipeline, the first- and second-level analysis steps are repeated for N = 500 from the N = 2000 subsample for only the ABCD data. Then, voxelwise_icc within the brain_icc.py script is used to derive estimates of the median ICC, MSBS and MSWS for the between session reliability across randomly sampled subjects for 25 to 525 subjects in intervals of 50. Similar to the methods in Liu et al. (2023), 100 iterations are performed at each N (with replacement) and the median ICC, the associated MSBS and MSWS estimates are retained from voxelwise_icc. The average and standard error across the 100 iterations are plotted for each interval of N with the y-axis representing the ICC and x-axis representing N. Stability is inferred based on the calculated* difference between the average ICC estimate for interval N_i & N_{i-1}. Here, stability in the estimate is defined as the difference between intervals that does not exceed .15. The value of .15 is used as this, based on recommendations of 'poor', 'moderate' and 'good', would change the inference that a measure is 'moderate' or 'poor' (Cicchetti & Sparrow, 1981). *The plotted values will be used to infer change and stability in the estimated median ICCs and variance components across the sample size. If stability is not achieved by N = 500, the sample is extended to N = 1,000 and the analyses are repeated.*

7. It is possible to incorporate an adaptive design in a Registered Report, though it does add complexity. I suspect, though, that you have already put some thought into which factors to vary in your analysis, and may have discarded some promising candidates. So you could have a prioritised list to work through. E.g. with 1st 500 participants, do study as planned. Then for next 500 continue with optimal level of factors you have tested and vary other steps in the analytic pipeline from your prioritised list.
    b. See above, 6a & 6i.

*Please regard my comments in this section as **suggestions rather than requirements**. You are free to disregard if you prefer.*

8. I was also thinking about the logic of specification curve analysis in this context. I think the value of doing this is greatest when you have a number of predictors that may interact, and so you need to look at all combinations. You are including predictors of two different types: it seems feasible that the smoothing and motion correction processes could interact and so it makes sense to look at all 5 x 6 combinations of those. But the task modelling and contrasts seem a different type of predictor, where it is likely that there may be just a main effect, with one level of each predictor being optimal. Task modeling and contrasts may interact with one another, but it doesn't seem likely that their impact will depend on the smoothing/movement correction settings. So I wondered

if a more economical approach would be to start with specification curve analysis for the 12 task model x task contrast combinations, using the smoothing and motion correction settings you anticipate as optimal. This may allow you to **identify clear winners for task analysis settings**, and these could then be set as **constant for exploring the smoothing/motion settings**. I don't insist you do this – I have little experience with this method, and am happy to be persuaded that the full specification curve analysis with 360 combinations should be done. But if you can save time and energy by reducing the amount of analysis without compromising the value of the study, that would be good.

a. We appreciate this suggestion, and we see the appeal of this method. It does parse the space into something more manageable with respect to the specification curve analyses. After some thought, we are inclined to pursue the proposed 360 model permutations for two main reasons. First, the effect of task parametrization on the resulting ICC is likely related to the effect of motion and smoothing. So modeling it together would be important. However, there are many ways that these analyses can be followed up. Second, related to the latter, we note that the derived first level and group level maps that are used in the analyses will be shared on NeuroVault so researchers may apply other methods (such as those proposed by the editor) to assess their specific effects on reliability:
    i. *Individual & Group Level Maps: The first level contrast maps, as well as the group level maps, that are used for these analyses will be openly shared on NeuroVault.*

9. A related point concerns Aim 3, i.e. finding the sample size at which the ICC stabilizes. Since the denominator for variance is square root of N, won't this just follow the trajectory of a plot of 1/sqrt(n) by n? (see below and https://osf.io/exvz8 for illustrative figure).If so, then N = 2000 is overkill, as the marginal benefits of additional subjects get pretty small after about 500. This might be another justification for adopting a more adaptive approach to allow more flexible use of this large sample, and possibly to explore other factors affecting reliability.
    b. We agree, the variance is scaled by the sample size in many formulas and so the values, such as ICC, MSBS and MSWS, will plateau with increased sample size. Many of our analyses are sensitive to these types of statistical properties, but as Schonbrodt & Perugini (2013; DOI: 10.1016/j.jrp.2013.05.009) previously reported, and Marek et al (2022; DOI: 10.1038/s41586-022-04492-9) and most recently Liu et al. (2023; DOI: 10.1038/s41562-023-01642-5) reminded us in the context of fMRI, it is worth repeating and demonstrating this property in real data. So we appreciate this recommendation and have incorporated this into our analyses as described above.

The three samples

10. I may have missed this, but I could not find a clear rationale for running the analysis on 3 samples. I assume this is to see if there is generalisability of findings across samples/scanners etc?
    a. The purpose of three samples is to evaluate the generalizability and consistency of the findings. In the introduction we include:
        i. *However, differences in modeling decisions across these studies leaves an important question unanswered: Are there certain analytic decisions that consistently improve reliability (e.g., ICC) of neural activity for an fMRI task across samples?*
    b. In the methods we include:
        i. *To answer the questions proposed in Aim 1 and Aim 2, this study will require multiple samples and tasks to obtain a comprehensive view of how analytic decisions impact group and individual reliability metrics (Aim 1) and how within- and between-subject variance is impacted (Aim 2) across multiple samples and the same MID tasktask.*
    c. But to be more explicit, in current study section we now include:
        i. *The purpose of multiple samples with the same task design is to evaluate the consistency in findings across studies that vary in their sample populations and task design as little evidence exists on the consistency of reliability estimates for the same task across independent samples*

Generalisability

11. Reviewers 2 and 3 have particular concerns about generalisability, noting that any recommendations for optimising reliability that come from this analysis may not generalise to other tasks/samples. While this is indubitably the case, my sense is that you have to start somewhere, and just doing this for one task across different samples involves a substantial amount of work. So I think what you have proposed here is worthwhile, and could set a standard for others working with different tasks to follow. However, it is clear that this will be a criticism of the study that you should anticipate, and as noted by reviewer 3, the limitations on generalisability do need to be emphasised.
    a. We appreciate this concern and certainly agree about the issues of generalizability to non-MID tasks. Like most studies, the reliability coefficients will be limited to the data they are derived from. Given the vast nature of the datasets and analytic decisions, we were forced to make a number of judgment calls in specifying our analyses. We agree with the reviewers and plan to include this consideration in the discussion at stage 2.

Evidence for accepting/rejecting hypotheses

12. We could do with a bit more clarity in how you propose to interpret the output from hierarchical linear models. What would be the criterion for deciding a particular processing choice (or combination of choices) was optimal? What would be the criterion for deciding that it made no difference? Would these interpretations be influenced by the overall level of reliability?

a. We appreciate the editor's comment. The HLM models are used to determine which modeling decisions are significantly related to the estimated median ICC. In this case, the interpretation is whether a model parameter, such as motion, is significantly associated with the change (increase/decreased) in the median ICC when all other analytic decisions are zero. Estimates from the HLM would imply which decisions have a non-zero effect on the median ICC. While unlikely, there is a chance that there will be no significant effect of modeling decisions on the resulting median ICC. Hence, the optimal model for Aim 3 is independent of the HLM. As noted in the manuscript, the modeling decisions that will be used in Aim 3 to determine the change in ICC/MSBS/MSWS will be based on the combination of modeling decisions that reflect the highest median ICC in Aim 2 that is reported in the specification curve. Nevertheless, the hypothesized 'optimal' decisions will be contrasted to the HLM results and the specification curve to determine whether a) different nominal/continuous decisions were significantly related to a higher ICC and/or b) different modeling decisions produced the highest median ICC. To be more explicit about this, we have added to Aim 1 and Aim 2 sections in the manuscript:
    i. "[...], *the interpretation focuses on the significant, non-zero effect of an independent variable (e.g., smoothing) on the dependent variable (e.g., median ICC) while the remaining independent variables are assumed to be zero*."

Open scripts and data

13. I see problems with the statements about accessing the MLS and AHRB datasets, as these depend on the permission of named individuals who are not immortal and could become unavailable. I appreciate that when using data provided by others one cannot dictate the terms of data access, but it is problematic if the analyses planned here are not reproducible. I'm not sure what the solution is, but hope that there is one. Ideally, the owners of these datasets might be persuaded of the value of depositing the BIDS compliant data on a repository such as Open Neuro or Neurovault.
    a. As mentioned above, we omitted to explicitly state that we will share the derived statistical maps from the participant and the group GLMs. While this does not make the analyses fully reproducible from raw data, it does provide other researchers with the basic elements (i.e. subject-level contrast maps) needed to reproduce our reliability analyses or apply other methods to assess reliability. We have added in the Data & Code availability statement:
        i. *Individual & Group Level Maps: The first level contrast maps, as well as the group level maps, that are used for these analyses will be openly shared on NeuroVault.*
    b. The BIDS Input data for MLS and AHRB were shared under an agreement that did not grant us the authorization to share the data via an open source platform, such as OpenNeuro. The ABCC data is available through a DUA via the NDA. We followed up with the PIs of the data and were informed that:

1. **AHRB**: After checking with the PI, Dr. Daniel Keating, they are reviewing their consent forms and checking with the Other Reportable Information or Occurrence to determine whether the sharing of the fMRIPrep derivatives would be a major deviation. If we learn that we are authorized to share the fMRIPrep derivatives, we will include it on OpenNeuro.
2. **MLS**: After checking with the PI, Dr. Mary Heitzeg, we were informed that we will be able to share the fmriprep derivatives on OpenNeuro.

ii. Given this new information, we have revised the Data & Code availability statement:

1. *fMRIPrep Derivatives: The fMRIPrep v21.1.0 derivatives and the associated events files for the MLS [**??** and AHRB study] data will be shared on OpenNeuro.org*.

14. In terms of the scripts, it is good to see these openly available; before publication, we will need a version with DOI, but I gather this is possible with Github integration with Zenodo.
    c. We appreciate the recommendation! Yes, we will generate a Zenodo for the github scripts for these analyses as well as the PyReliMRI library.

Minor points

15. Line 25 – 'scripts for'
    a. revised
16. Line 100- I would reword for clarity: '…relative to individual differences, and to assess….'
    b. revised

REVIEWER #1
Signed, Dr. Xiangzhen Kong

17. For estimating reliability at the group level, the authors plan to use "significance thresholded (p< 0.001, uncorrected) group [binary] estimates." This is not a common threshold strategy reported in publishable studies. To enhance the applicability of the results, the authors may consider using a more commonly used group analysis threshold strategy e.g., one that combines an uncorrected p threshold (e.g. 0.001) and a cluster size threshold.
    a. We appreciate the note and recommendation. Since we are not looking to draw group-level inferences but rather evaluate how analytic strategies impact the similarity coefficients and be within comparable threshold used for the supra/subthreshold distinction for the brain masks gathered from neurovault (i.e., collection that comprises the anticipatory win phase across 15 whole brain maps for the MID task [Wilson et al., 2018; Collection: 4258, ID: 68843]). In text, the nature of jaccard similarity has also been revised as follows:
        i. *Since the Jaccard similarity coefficient is sensitive to thresholding and sample size (Bennett & Miller, 2010), in Aim 1 an equal sample size (e.g., N ~ 60) is chosen for each study to compare how the similarity between sessions varies across studies.*

18. The authors may clarify why two metrics (Jaccard's similarity coefficient and tetrachoric correlation) are planned to use and whether they provide complementary information on reliability results.
    a. This is a good question. The purpose of including the tetrachoric correlation is to supplement the group level jaccard similarity metric. This estimates a correlation that would have similar bounds as the individual level estimates. However, in hindsight this appears that the tetrachoric correlation may be a bit redundant when referring to the binary images. Instead, we have added to the PyReliMRI package the option to calculate the 'Spearman Ranked' correlation between two unthresholded images to complement the thresholded binary to demonstrate how the activity across the whole brain correlates between sessions.

19. "the Aim 2 Hypothesis is that, on average, there will be higher between-subject than within-subject variance across the three samples within task-relevant regions." The wording of the Aim 2 hypothesis is a bit confusing. E.g., we may never expect between-subject variance to be lower than within-subject. The authors may consider rephrasing it to make it clearer. Additionally, it would be better to clarify how the "task-relevant regions" will be defined.
    a. This is a good point. This was a challenging hypothesis to settle on due to different evidence on reliability in task fMRI and the lack of information on changes in between- and within-subject variance in the MID task. In this case, we may have insufficient evidence to make explicit predictions about the relative

changes in the variance components in *this* task and *these* samples. As a result, we have revised the Aim 2 hypothesis:

   i. *Due to the poor reliability of individual estimates in task fMRI (Elliott et al., 2020), reported evidence of high between-subject variability in neural activity (Turner et al., 2018), and limited evidence on changes in between- and within-subject variance components in the MID task, we do not have a specific Aim 2 Hypothesis.*

20. "Aim 3 evaluates at what sample the ICC stabilizes using the most optimal pipeline (e.g., highest ICC) from Aim 2." Is it correct that all samples will be used for Aim 2, and only smaller subsets of the samples will be used for Aim 3?
    a. As specified in the manuscript, the stability of ICC (aim 3) will be calculated in the ABCD study sample only, given that is the only sample large enough to perform these analyses. We have emphasized *only* in-text to make that more explicit.

21. Year 1 and year 2 data were used for each of the three datasets. As reliability estimates may decrease with the time lag between two scans, I would suggest the authors report the exactly time lag in each dataset.
    a. This is a good recommendation. We will attempt to calculate these for subjects from the data provided and include this information in the stage 2 manuscript. In cases this is not directly available in the provided data, we will work with the data sharing group to obtain this information.

22. "Structural and functional MRI preprocessing is performed using (?) 23.0.0rc0". The package name is missing. Is it fMRIPrep?
    a. We appreciate the reviewer catching this. It has been corrected.

23. The meaning of A/B/C/D in Figure 1c and Equation 3 is unclear. In addition, there seems to be some confusion regarding the two "Ses-1" in the first row. The authors may consider clarifying these points to enhance the readability of the manuscript.
    a. Given the comment above and anothers reviewer's comment, we have expanded the PyReliMRI package and instead propose to use spearman rho. We have also elaborated on each equation in-text.

REVIEWER #2

Major suggestions

24. My biggest concern is that the current work is based on only a very narrow set of data: anticipatory cue periods from the monetary incentive delay task. This will strongly limit the generalizability of the current results – it is unclear if the findings from this particular work would have any bearing on many other potential tasks that could be run. Given the

premise of the article, it's not clear why the authors chose to limit themselves in this way. This work would have a much bigger impact if the authors were to test the pipelines across a range of tasks (e.g., each of the tasks in the ABCD, HCP). It may not be possible to apply all of the combinations of analysis steps in the new tasks (e.g., the different versions of task parametrization), but even if only a subset of analysis options were repeated, it would be very helpful for understanding the degree of generalizability of the current results to other contexts. If the authors choose to keep their focus narrowly on the MID, this should be reflected more clearly in the title and abstract, and in discussing the types of conclusions that can be drawn from the current work.

a. The reviewer certainly brings up an important point about generalizability. We settled on the task and the phase of the task for a couple of reasons, some of which are discussed in-text but elaborated below. To ground the work specifically in this task, we add it explicitly in the abstract and current study section.
    i. First, the goal was to determine which/how analytic decisions *consistently* impact the reliability coefficients within and across studies. To achieve this, it would be prudent to limit the amount of parameters that differ across task designs. Of course, HCP, ABCD, IMAGEN or UKBiobank provide sufficiently large datasets but they also incur limitations with regards to the types of tasks administered. There is, generally, little overlap between the cognitive tasks to determine which task parameterization(s) alter the results. In the case of the current study, while we are unable to generalize to other tasks, we are able to evaluate to the degree (if any) some conclusions are replicable across studies. We believe this is an important starting point; assessing the generalizability to other tasks or task features, while clearly important, is simply beyond the scope of the present analysis.
    ii. Second, finding additional tasks that are comparable across studies where there are >2 waves of data with N > 50 is surprisingly challenging (as evidenced by a search of the OpenNeuro archive).
    iii. Third, we focus on the anticipatory contrast for two reasons. First, as described in Demidenko et al. (2021, DOI: 10.1002/brb3.2093), the anticipation phase plays a prominent role in motivation. Second, across studies since 2003, the part of the task that was modeled as the onset and window of the duration in the GLM has varied significantly (this is also discussed in supplemental and in-text).

25. The authors discussed reliability issues at length, but did not touch on the topic of validity. These usually go hand in hand, as methods to increase reliability at times reduce the validity or utility of a dataset – for example, consistency of datasets is likely to increase monotonically with smoothing, although this will come at the cost of being able to distinguish meaningful/valid results from distinct areas of the cortex. Similarly, different task contrasts will have different levels of validity in terms of the underlying cognitive constructs that they can be connected with. At a minimum, I think this bears additional discussion in the final manuscript. If the authors could also propose a

confirmatory test of validity (e.g., using a motor task to find distinct components of the motor system), that would also help with the interpretation of the results.

a. We appreciate what the reviewer is bringing up and we certainly agree. Validity is an essential complement to reliability. One can increase the reliability of a measure but could be way off target in measuring the underlying construct. While we did not discuss the issue at length in the introduction, we did note that '*Poor reliability can hamper validity in cognitive neuroscience research, reducing the ability to uncover brain-behavior effects*'. However, we intend to allocate space in the discussion in the stage 2 report of the manuscript. We agree that these two cannot always be treated independently when conducting individual differences research. Construct validity is essential to both replication (Flake et al., 2022, DOI: 10.1037/amp0001006) and generalizability issues (Flake et al., 2022, DOI: 10.1017/S0140525X21000376). As discussed in Clifton (2020, DOI: 10.1037/met0000236) there are tradeoffs between identifying the latent variable and the correct identification of the latent variable. Specifically, systematic error that is favorable to reliability may be disadvantageous when it comes to validity. We plan to elaborate on the importance of this distinction more in subsequent versions of the manuscript.

26. I was surprised that the authors did not consider the amount of per subject data (rather than the total number of subjects) in their evaluation of the reliability, given that recent reports have consistently found this to be a primary way of improving reliability of the findings. This might make sense if the focus of the work is only on analytic means to improve reliability, but given the proposed analyses in Aim 3, it seems reasonable to ask whether the amount of data per subject or number of subjects makes a bigger difference for reliability (I would guess the first).

    a. This is a valid point by the reviewer. The amount of data per subject would improve reliability, such is the case recommending longer resting state scans or the point in Elliott et al. (2021, DOI: 10.1016/j.tics.2021.05.008) that longer assessments, or trials, may reduce the error score (pg. 780). In the case of these MID task fMRI data, while there are some subtle time differences between task length (as a function of longer trials, on average) the total number of trials for a given run (i.e., 50) or condition (i.e., 10) is constant per run across studies. Of course, this is an important feature to discuss for future research whereby similar task designs have a greater number of trials. The number of trials for a fMRI task is especially important when picking a protocol that fits all the necessary acquisitions (e.g., task, resting, structural) within a prespecified time. For example, the IMAGEN study lengthened the ISI/ITI in their MID task to improve the ability to differentiate stimuli and trials but this was at the cost of eliminating all loss conditions. This relates to an important discussion in fMRI research, what are the acquisition protocols and preprocessing decisions optimized for? The trade-offs become quite important as it relates to individual differences research.

27. In a similar vein, the current work is based on a very small amount of data from each participant (comparison of single runs), and for a set of contrasts that likely have relatively small effects (e.g., relative to sensory/motor stimulation). I am concerned that the authors will be at floor in terms of reliability, preventing them from finding meaningful differences across analysis strategies. Is there any evidence that this will not be the case? Including a positive control dataset with more data per participant would help to alleviate this issue. The hypothesis that between subject variance is larger than within subject variance may not be true in the context of this very small amount of data.
    a. This is a good note by the reviewer. With respect to the MID task, reliability estimates in [small] adult samples have reported ranges of .45 - .80 across the NAcc, Insula and mPFC (n = 14 test-retest sample; Wu et al., 2014, DOI: 10.1016/j.neuroimage.2013.08.055). This was in a task design where there were 90 trials compared to our 100 trials of data. Conversely, Kennedy et al (2021, DOI: 10.1016/j.neuroimage.2022.119046) reported ICC scores < .20 in the MID task across early adolescence. One would prefer the larger ABCD sample in Kennedy et al., however, this study used the DEAP derived ROI data for the MID tasks that uses a modeling technique of the cue onset as the impulse response (see Hagler et al., 2019, DOI: 10.1016/j.neuroimage.2019.116091). This approach is different from how the task has been modeled in the literature since 2003.
    b. This MID task and the current design are widely used in the literature, which led its selection for large-scale data collection in the ABCD study. As such, it would be valuable to assess more thoroughly how reliable the task is across runs and sessions using multiple studies with similar preprocessing strategies. Furthermore, it will be valuable to understand how the analytic decisions impact the estimated values of reliability. Whether our analytic decisions will result in significantly different estimates of reliability is an empirical question. The evidence above suggests that the ICC may not be at the floor and there may be sufficient variability to ask this question. Our analysis will provide evidence regarding how much reliability estimates vary and whether these are as a function of limited between subject variability and/or increased across session variability. As a result, this variance may be captured in the HLM for which the power is derived across N = 360 model comparisons (nested within study). This dependency, of course, will be an important study consideration.

Minor suggestions

28. Effects of younger vs. older samples may be influenced by study design (e.g., fMRI imaging parameters, block lengths, etc.) that are not matched across datasets. How will this be addressed? It seems like it may be challenging to interpret differences across datasets.
    a. The reviewer is correct - there are meaningful differences between the three samples in the acquisition parameters and/or age that likely interact. When using open data, there is an inherent challenge in finding the perfect data scenario to answer this question. While we do not think we have a perfect scenario here, we

believe by sharing the derived participant and group level maps and summarizing the resulting metrics across each study separately will provide a reasonable amount of information that future researchers can build off of and confirm in these data that they are specifically interested in. Furthermore, an implicit assumption in contrast findings between studies is that some of reported effects across studies are interchangeable if the methods are similar. These results will provide sufficient information to indicate how troublesome this assumption is.

29. Table 1: #5 says “censor high motion frames” but not what the cut off for these will be? This should be specified in the report.
    a. Thank you for the clarification. In hindsight we realized we had not included this information and did not update this from fMRIPrep’s default setting of FD = .50 (tailored for rsfMRI). We have revised this in-text to indicate that the threshold used is FD = .90, a more appropriate threshold for task fMRI (Siegel et al., 2014, DOI: 10.1002/hbm.22307; Zhao et al., 2023, DOI: 10.1016/j.neuroimage.2023.119946). This has been updated in table 1 and updated in supplemental materials:
        i. *Frames that exceeded a threshold of 0.9 mm FD or 1.5 standardized DVARS were annotated as motion outliers.*

REVIEWER #3

30. First, while the topic is important, the novelty of this work is not clear. There is already a massive literature about test-retest reliability in fMRI in different domains (visual, motor, language, memory, decision making…etc.). Many previous test-retest studies are not cited in the current version. I believe it would be useful to discuss reliability with respect to its origin/source and its implications in fMRI: e.g. differences due to situational factors (e.g. Dubois and Adolphs 2016 TICS) or ‘meaningful’ differences that reflect inherent cognitive and behavioral differences (Seghier and Price 2018 TICS). For instance, if task performance would vary across sessions, should fMRI activations vary as well to reflect such differences in task performance? My point here is that test-retest reliability in task performance (behavioral data) should not be overlooked when explaining test-retest reliability in fMRI (as long as there is variability in behavioral data there will be variability in fMRI data).
    a. We appreciate the reviewers point. Yes, there is a long list of studies evaluating test-retest reliability in task fMRI. While we do not cover the literature at breadth, a number of the citations we point the authors to (e.g., Elliott et al., 2020; Noble et al.., 2019, 2021; Dubois & Adolphs, 2016, Bennett & Miller, 2010, 2013; FrohNer et al., 2019; Herting et al., 2017) have covered a broad range of studies in their meta-analyses and/or systematic reviews. The work we specifically focused on here are how analytic decisions impact the resulting ICCs. The novel

aspects of our work are the analysis of the effect of analytic decisions on reliability estimates, the consistency of these across multiple studies, and the large sample size. With respect to the sample size, in-text we specifically note:

i. *The median reported sample size in fMRI is <30 subjects (Poldrack et al., 2017; Szucs & Ioannidis, 2020). From the review of task fMRI reliability by Bennet and Miller (2010), the median sample for individual (continuous) reliability is 10 subjects (mean = 10.5 [range = 1 to 26]) and for group (binary) reliability is 9.5 subjects (mean = 11.2 [range = 4 to 45]). A recent review and analysis of task fMRI reliability suggests sample sizes are increasing but remain lower than the median sample size in task fMRI, whereby the median sample size for individual reliability in the meta-analysis are 18 subjects (mean = 26.4 [range = 5 to 467]) and the analyses are 45 & 20 subjects (Elliott et al., 2020)*

b. We appreciate the reviewer for reminding us of the Seighier & Price (2018), which we are familiar with but did not cite here. The reviewer raises a number of points that we will address in the discussion. Reliability is a single parameter that represents the data in a particular way that remains very common in individual differences research. In fact, here we evaluate two different forms that have traditionally been used (group maps and individual maps). With respect to individual continuous estimates of test-retest reliability, the reliability coefficient dates back as far back as Fisher (1934) and Harris (1913, DOI: 10.1093/biomet/9.3-4.446) but Shrout & Fleiss (1979; DOI:10.1037//0033-2909.86.2.420) is commonly cited within the framework of classical test theory (Algina et al., 2015, https://www.sciencedirect.com/science/article/pii/B9780080970868420702). It is common that researchers cite cut-offs that refer to Excellent, Good, Fair or Poor reliability (Cicchetti & Sparrow, 1981). However, the premise of the work here is not to deduce whether the reliability is poor, good or excellent, but instead to add to our understanding of how the analytic decisions impact the estimates of reliability. A promising follow-up study would be to expand on any prospective differences to gain clarity on how different phenotypes or behaviors may impact the derived estimates of reliability. These will be important discussions in the stage 2 manuscript. In the meantime, we include in-text the following:

i. Furthermore, behavioral data from the MID task will also be reported as supplemental information, such as the mean and distribution of probe hit rate and mean response times (RT) across subjects. The task design is programmed to achieve a probe hit rate of approximately 60% for each subject. It should be noted that the RT for the probe are not consistently collected across the ABCD, AHRB, and MLS datasets.

31. In the same way, performance during the selected MID task varied across runs and subjects. The targeted accuracy was set to 60%, which is relatively low (is the task challenging?). How the first-level GLM models will be defined? The authors mentioned in

Page 12 that the design matrix included 15 task-relevant regressors (5 cue and 10 feedback types), but it is not clear if all trials will be included or not? For example, it is likely that correct trials might show different activations (and different reliability) than incorrect trials. Again, task performance is a huge confounder in such test-retest type of analysis and thus the authors should clarify how correct versus incorrect trials will be modelled.

a. The MID task is calibrated to the individual subject's performance. As noted in supplemental Figure S1, *The probe window increases/decreases as the participants probe hit rate increases/decreases below a target of ~60%*. Thus, the task is designed to achieve, on average, 60% *probe hit rate* for each subject. We have revised in supplemental section 1.2:
    i. *For example, during the MID task each trial starts with a cue type and consists of three phases: anticipation, probe and outcome (that is, feedback). The task regressors include different cue (five) and feedback types (ten), totaling 15-task regressors* ***that are included in the GLM***.
b. It is important to note that we model the task in a manner that is consistent with the prior literature. With the exception of modulated task regressors, the prior literature has not modeled differences in performance (i.e., probe hit versus miss) across trials. The current paper is not focused on modeling performance in this design but given the variability in how participants engage with the task, this would be a valuable future study.

32. As the authors know, test-retest reliability may vary with task. For instance, it is likely that tasks involving primary sensory cortices might show higher reliability than tasks activating high-level processing regions. Even for the same domain, reliability might vary with activation condition (e.g. see for the language domain: Otzenberger et al. 2005). In this context, I'm not sure what is the exact impact of the expected findings on current practices; i.e. the best parameterization (or combination of methodological choices) might not generalize to other contexts that (1) use different tasks and cognitive domains, (2) different populations, (3) different time points between repeated sessions, (4) different magnetic fields, (4) different paradigms (block, event-related, mixed…etc), (5) different acquisition protocols (high temporal or spatial resolution EPI), (6) variable run/session length (e.g. Genovese et al. 1997 MRM) …etc. There are so many factors that can impact upon the BOLD signal and its reliability. Also, there are already many guidelines (recommendations) about optimal pipelines and analysis protocols to improve reliability and thus it would be nice that the authors could spell out the expected implications to current practices.
    a. The reviewer makes great points and we agree with the concerns, but we do not think that these points devalue the importance of understanding how analytic flexibility impacts the results. In fact, these points are arguments in favor of building on the prospective results from this study. The strength in sample size, consistency in task and range of analytic decisions may produce results to motivate a number of the questions the reviewer notes. Specifically, we use three distinct samples (N >= 60) in this work with a task design that is fairly

comparable. Prior work, as noted in-text, calculated reliability coefficients in mostly small (median n < 40) samples. Given the error surrounding these estimates, it is unclear whether a reliability of .30 in study A (N = 15) and reliability of .55 in Study B (N = 12) is simply reflecting the variability at low Ns or the extreme fluctuations in ICC that would be expected at larger sample sizes. An assumption in the field is that fMRI tasks evoke some reliable brain state that may be useful in understanding behavior (e.g., Green et al., 2023, DOI: 10.1016/j.tins.2023.04.001). …

33. I find the inclusion of task-irrelevant voxels slightly confusing. Basically, task-irrelevant voxels are just the complement set of task-relevant voxels (assuming relevant voxels + irrelevant voxels = the whole brain as a constant). So, I'm not sure what would be gained by that inclusion, in particular when using binary maps to calculate the similarity metrics like Jaccard. Note also that the definition of task-irrelevant voxels is contingent to the selected threshold and the used baseline/control condition.
    a. In response to the reviewers' comment and the highlight by the handling editor, we have revised the below in-text:
        i. *For the continuous estimates of reliability described below, the analyses will be performed separately on task voxels that exceed and do not exceed an a priori specified threshold based on the NeuroVault (Gorgolewski et al., 2015) meta-analysis collection that comprises the anticipatory win phase across 15 whole brain maps for the MID task (Wilson et al., 2018; Collection: 4258, ID: 68843). The suprathreshold task-positive voxels are those that exceed the threshold ($z > 3.1$) and the subthreshold task voxels are those that do not exceed ($z < 3.1$) in the map. We acknowledge that the threshold of $z = 3.1$ is arbitrary (uncorrected, p-value = .001) and that the voxels that fall below and above this threshold may not be significantly different (Gelman & Stern, 2006). However, to constrain the problem space this is a researcher's decision that is made in these analyses (Gelman & Loken, 2014; Simmons et al., 2011).*

34. Some statements need to be integrated in a more coherent way; e.g. Line 161 "there will be higher between-subject than within-subject variance across the three samples within task-relevant regions" and Line 329 "the reliability within sessions would be hypothesized to be greater than between sessions" versus Line 395 "To test whether the between-subject variance is lower than within-subject variance consistently across the three samples".
    a. We appreciate the reviewers' comments. This is related to Aim 2 and the associated hypothesis. As discussed above, we have insufficient evidence to make explicit predictions about the relative changes in the variance components in *this* task and *these* samples. As a result we have removed the hypotheses and these explicit statements in the manuscript.

35. In Line 735, I didn't get what the authors meant by this statement: "so it could be the case that a less efficient model captures the data variability better and the estimated variance of contrast estimates is reduced." How this statement should be understood with respect to modelling efficiency of Figure S2?
    a. We have added an example, referring to Figure S2, to illustrate our point:
        i. *The efficiency of a model's design matrix only reflects part of the first level model's variance, which is the product of the inverse of the efficiency and the residual variance. The most efficient design matrix may not fit the data well, increasing the residual variance and the overall variance of the estimated contrast. For example, consider CueMod and AntMod for the LGain v BL contrast. CueMod has higher efficiency due to lower overlap between the anticipation regressor (only modeled during Cue Onset+Cue Duration) and the Feedback regressor, but if the anticipation-based brain activation continues throughout the fixation period, CueMod will not capture this variability as well as AntMod. Whether CueMod outperforms AntMod for this contrast depends on whether the increased efficiency of CueMod is overshadowed by an increase in residual variance due to poor model fit.*

36. In the same way, Line 134 "However, there is little empirical research on whether the culprit in the reportedly low reliability of fMRI signal across measurement occasions is a decreased between-subject and/or an increased within-subject variability." If there is a decrease in the between-subject variability, it would then make sense to expect a subsequent increase in reliability. I feel there is a confusion in the current version (at least to me) about the exact impact of low/high variability toward low/high reliability that should explained beyond the issue of reliability quantification with ICC. Put another way, ICC is just one type of objective measures used for the assessment of reliability, but the relationships between intra/inter-subject variability and reliability in general is much broader than that.
    a. Because reliability via the ICC is directly defined as the ratio of between subject variability to total variability, the reliability can be increased by *increasing* differences between subjects or *decreasing* the differences within-sessions. As illustrated in a recent paper from Xu et al. (2023, Fig2), the most efficient direction to improve reliability in individual differences will depend on what the within and between subject variance are. To alleviate confusion, we will revisit in-text to make necessary tweaks to the language to improve coherence.

37. In the study of Kennedy et al. (2022) with the ABCD dataset, the authors assessed reliability and stability for short term (within-session) and long-term (between-session) and reported overall poor reliability and stability. Please add a statement about how the current study can help make sense of Kenedy et al.'s findings.
    a. We have added this in-text:

i. *We expand on these findings by evaluating how consistent these results are across studies and which analytic decisions impact the reliability estimates.*

38. Jaccard similarity index might show low values for small activated volumes, which means that Jaccard values can decrease with more conservative thresholds. The issue of thresholding is thus critical when calculating Jaccard.
    a. The reviewer is correct and we had noted this in the initial submission as a follow up analysis. Upon reviewers' comments and reflection, since this fact is well delineated in the literature we instead focus on Aim 1 for the constrained subject. Specifically, we revised in text:
        i. *Since the Jaccard similarity coefficient is sensitive to thresholding and sample size (Bennett & Miller, 2010), in Aim 1 an equal sample size (e.g., N ~ 60) is chosen for each study to compare how the similarity between sessions varies across studies.*

39. Equations (1), (2) and (3): explain what these abbreviations/parameters stand for? (In addition to what is already mentioned in Figure 1)
    a. These details have been added in-text
        i. *The parameters in Equation 1 are: MSBS is the Mean Squared Between Subject Error and MSError is the Mean Squared Error.*
        ii. *In Equation 2, J(A, B) is the similarity coefficient between A (session 1) and B (session 2). This is derived from intersection, $|A \cap B|$, which represents the elements that are common to both A and B divided by the union, $|A \cup B|$, or the elements that are both in A and/or B. In Equation 3, the Spearman Rank Coefficient, as implemented in Scipy stats using spearmanr (Virtanen et al., 2020), is ranked correlation between unthresholded images A and B, whereby Σd^2 is the sum of squared differences between ranked values in session A and B, normalized by $(n * (n^2 - 1))$.*

40. Line 456: how the sample size N=60^5 was defined?
    a. Apologies for the confusion. This is the expected minimum of the test-retest data that is expected to be available once the data is preprocessed and analyzed. Specifically, note #5 specifies:
        i. *The sample size will depend on the N after quality control and preprocessing of the smallest test-retest data from MLS or AHRB. As noted in footnote two, this information is not available at the Stage 1 submission.*

41. Explain what the tetrachoric correlation will add on top of what one can get with Jaccard (both indices use binary data).
    a. The purpose of including the tetrachoric correlation is to supplement the group level similarity metric, jaccard, that is proximal to the correlation coefficient at the individual level. However, this appears that the value may be a bit redundant when referring to the binary images. Instead, we have added to the PyReliMRI package the option to calculate the Spearman rank correlation coefficient between two images which would complement the thresholded binary to 'account for information about the variability and magnitude of brain activity' in each group map across sessions (Churchill et al., 2010 pg 856, DOI: 10.1002/hbm.21036). By addressing this and the below, we have refined the figure and the formula in the manuscript.

42. Figure 1b: Jaccard index (i.e. Intersection over Union) can be simplified as Red divided by the sum of Red, Blue and Green (no need for the two extra Red squares in the denominator of 1b).
    a. The figure has been simplified to reflect this recommended change.

43. Table S3: TE of the BOLD run for MLS/GE Sigma looks odd (3.6ms !).
    a. Thank you for bringing this up – this definitely is a typo that has been corrected. We had completed the information from previous work and recently got a spreadsheet with these values from the MLS study coordinators.