Editor

Many thanks for your submission of the stage 2 report for this article. Two of the original reviewers have submitted reports: one is happy with the paper as it is, and the other has some specific questions that should not be too difficult to address. I thought you did a good job in leading the reader through a very complex set of analyses, but I have a few minor suggestions to make it clearer.

1. Lines 189-193: mention of a sub-hypothesis about sample differences. I don't think this was ever picked up on in the analysis? Just a sentence about it would suffice.
    a. Outside of plotting the variability across samples, this made us realize we did not directly compare AHRB/MLS > ABCD ICC estimates. These have been directly compared using t-tests and reported on line 679-682:
        i. *We had hypothesized that the ICC estimates in the older samples (AHRB/MLS) would meaningfully differ from the younger sample (ABCD). Overall, ICC estimates were higher in the older than younger sample for between-run, t(497.2) = 5.53, p < .001, d = .43, and between-session, t(669.9) = 9.57, p < .001, d = .66.*
2. The Within-run vs Within-session analysis : I wasn't sure whether the lower ICC for within-run might just be a consequence of N trials being smaller?
    a. While we could not directly compare this in the analytic pipeline as we only had two runs of data, this was part of our speculation as we highlight in the discussion section on line 956-959:
        i. "*Higher between-session reliability may be related to decreasing activity from early to later runs (Demidenko, Mumford, et al., 2024) or based on the sessions being an average of two runs/increased trials (Han et al., 2022; Ooi et al., 2024)*."
3. I'm not used to looking at Specification curves and struggled to understand the jump in the curve in Figure 3A. Could you add a sentence to just explain that to the reader. Likewise, it would help to just explain that in Figure 3B one is looking for a cluster of values on the right hand side of the plot as indicative of a variable that is associated with higher ICC.
    a. We had specified in the Stage 1 Registered Report that in-text we would report the upper and lower quartiles. However, as the editor and one of the co-authors noted, this is not a *common* reporting strategy. We replaced the upper/lower quartile spec curve with the full distribution to avoid this type of misunderstanding.
4. p 34, para 2, aim 3. I think you could make a bit more of this. There has been a tendency, I think, for people who are critical of small sample sizes in fMRI studies to think the more the merrier, and go for very large samples. Your analysis (and indeed sampling theory) suggests it is more appropriate to recognise that beyond a sample size of around 250 there may be little additional benefit to increasing sample size - particularly when one considers that this is a cost-benefit decision where the costs are

substantial - not just in terms of paying for scans, but also the time costs of processing additional data. I think if you wanted to be provocative you could argue that for N greater than 250 the researcher would need to justify what gain there would be to justify the additional cost.

   a. The sample size at which the ICC estimate for a brain measure stabilizes is not to be confused with the sample size where a between-subject analysis with that brain measure will reach maximal power. The ICC measure indicates how well a measure correlation will match the true correlation: *r*(A[observed], B[observed]) = *r*(A[true], B[true])*sqrt(Reliability(A[observed]) x Reliability(B[observed]). As such, a low ICC for a brain measure tells us our observed correlation between a brain measure and behavior measure will be much lower than the truth, which implies we will need more subjects to power the detection of the correlation of interest than if the ICC was higher. The stability of the ICC we have estimated simply indicates that we can then more confidently use the ICC measure to inform on potential power scenarios. An example of such power curves linking low ICC to power can be found in Figure 1 of Elliott et al (2020). The bigger question is whether these small correlations have a meaningful impact on our understanding of the brain and are worth pursuing and this is an issue whether or not the ICC is low.

5. p 36. I think these should be described as Exploratory analyses, as they were not preregistered. I won't insist on this, but my inclination would be to put these into Supplementary material. The reader has a huge amount to process in this paper, and by the time I got to this point I was running out of steam. I think it would be reasonable for you to just explain in a couple of sentences that you did conduct these additional analyses, and that readers who are interested can find them in Supplementary materials.
   a. In the previous version of the manuscript, the section was labeled as "**Post Hoc Analyses**". We believe this captures the exploratory nature. We have moved subsequent subsections to the supplemental materials and replaced the text on line 835 with:
      i. "*An exploratory set of analyses were performed to evaluate 1) the effect of analytic decisions on ICC for the Left and Right Nucleus Accumbens and 2) the association between voxelwise Cohen's d estimates at the group-level and the voxelwise ICC maps. These are reported in supplemental **section 2.6**.*"
6. p 29: end of Aim 1a - v briefly compare obtained results with predictions from Table 1. And say something about the prediction re age effect (from lines 189-192).
   a. As noted in the above comment, we have performed these analyses and included this comparison in the results section. See response 7a. below.

7. In fact, at end of each Results subsection, I think it would be useful to have a little section with subheading such as "Summary of results on Aim 1a", where you specifically contrast predictions you made in Table 1 and the results that were obtained.

   a. Given the breadth of results for Aim 1a and Aim 1b, we have included brief subsections with summaries:
      i. **Aim 1a**. *Overall, between-run ICCs are slightly lower than between-session ICCs. Across the three samples, the highest ICCs, on average, are within visual and motor areas and the lowest ICCs are within the ventricles and white matter. In Table 1, it was hypothesized that the optimal analytic decisions would be: FWHM Smoothing 2.5x the voxel size, Motion correction that includes translation/rotation, their derivatives, the first 8 aCompCor components and exclusion of > .90 mFD subjects, the anticipation Model Parameterization, and Contrast Large Gain > Implicit Baseline. Contrary to registered hypotheses: (1) smoothing had a small but linear effect on ICC estimates, whereby the largest median ICC was for the largest FWHM smoothing kernel (3.5x voxel size); (2) Motion correction had minimal and negative impact on median ICCs in case of more rigorous corrections; and (3) the Cue and Fixation Models had higher estimated median ICCs than the Anticipation model. Post hoc analyses illustrated Model Parameterization is largely driven by the Implicit Baseline contrast, as Model Parameterization has a negligible impact on between condition contrasts. Consistent with registered hypotheses, the Large Gain versus Implicit Baseline had the highest estimated median ICC. Contrary to registered hypotheses, there was little evidence to suggest that analytic decisions differentially impacted estimated median ICCs between developmental samples (e.g., oldest MLS/AHRB versus younger ABCD data). Finally, the older samples (AHRB/MLS) had higher between- and between-session estimated ICCs than the younger sample (ABCD).*
      ii. **Aim 1b**: *Similar to Aim 1a, on average, the supra-threshold Session 1 between-run Spearman and Jaccard similarity is slightly lower between-session similarity. Spearman similarity meaningfully differed across Contrast, Model Parametrization and Smoothing, and it is near the ceiling for the upper tail of the Spearman similarity estimates. Like Aim 1a, Model Parametrization is driven by the Implicit Baseline. Finally, mean-based group activity maps illustrate that the Cue and Fixation models are opposite of each other when the contrast is a between condition and implicit baseline comparison.*

minor typo issues

8. General - in several places the text is still in future tense; please address this in the following places: p 12 paras 1-2; line 344; lines 453-545; last para p 18; line 507;
   a. Updated.
9. L52: semicolon before 'however' (starts a new sentence)

    a. updated

10. L179: would help if the list of analytic decisions was in the same order as in Table 1
    a. Updated order in paragraph to Smoothing, Motion Correction, Task Model and Task Contrast.
11. L204 "will stabilize at a sample size between"
    a. Corrected.
12. L205: this seemed a bit disjunctive. I wasn't sure whether it should go in methods. Alternatively, a subhead would help the reader with the transition in topic.
    a. So it is not as abrupt, the packgraph has been moved to the methods section, line 400-409
        i. *Several packages exist to calculate ICC and Jaccard/Dice coefficients. For example, ICC_rep_anova & Similarity in Python (Gorgolewski et al., 2011), fmreli in MATLAB (Fröhner et al., 2019) and 3dICC in AFNI (Chen et al., 2017). However, these packages are either a) limited to a specific ICC calculation (e.g., ICC[3,1]), b) not easy to integrate into reproducible python code (e.g., fmreli), c) do not include similarity calculations (e.g., 3dICC), or do not return information about between-subject, within-subject and between-measure and variance components. Thus, to have the flexibility to estimate ICC(1), ICC(2,1) and ICC(3,1), Dice and Jaccard similarity coefficients and spearman correlations simultaneously, we wrote and released an open-source Python package with reliability and similarity functions that works on 3D NifTi fMRI images (PyReliMRI, Demidenko & Poldrack, 2023).*
13. L211 and elsewhere, Spearman with capital S
    a. Have updated instances of lowercase
14. L290, should this be Table S3
    a. It seems the supplemental ordering of tables is off by one, as Table S3 should be Table S2. This has been revised.
15. L351: Github with capital G
    a. Updated.
16. L406: delete 'the'
    a. Updated
17. last para p 21: space before units 'mm'
    a. Updated
18. L741: 'is the opposite of the Fixation model'
    a. Updated
19. L 769: the variability will decrease as a function of square root of N
    a. Updated
20. L844; semicolon before 'however'.
    a. Updated
21. References: a bit pedantic, but I like to see consistency in whether sentence case or title case is used; there's a bit of a mixture here.
    a. These have been updated within the manuscript.

Supplementary material

1. L 24 "some do not have the necessary..." - is this an explanation for why some subjects were excluded? It's a bit unclear
    a. This exclusion decision was only used for the neuRosim option, where we were computing numerous contrasts (some not used in analyses). Since the feedback conditions were models but not contrasted in the paper, this was not part of the exclusion criteria. In supplemental updated with "*For this demonstration, some do not have the necessary outcome events which prevent the use of data in this case*"
2. Figure S2 - did you mean to retain this in the paper? It says IGNORE THIS RESULT in big red writing
    a. We retained this to be transparent with the reader the method we used, informing them of the package error and why our results diverged. We have modified the figure to say:
        i. "**Deprecated result:** We identified an error in neuRosim with how convolution is estimated. This does not impact other efficiency estimates as Nilearn is used in Stage 2 analyses"

Reviewer #1

The authors adhered closely to the analysis plans outlined in their preregistered report and presented the results accordingly. They provided thorough discussions on the findings. The current manuscript reads well. I have no additional comments to make.

Reviewer #2

I enjoyed reading this new version of Demidenko and colleagues' paper (Stage 2 Registered Report). The current version provides empirical data to test previous aims and hypotheses. I also appreciate the authors' efforts in making this version readable despite the massive number of analyses and findings in this project.

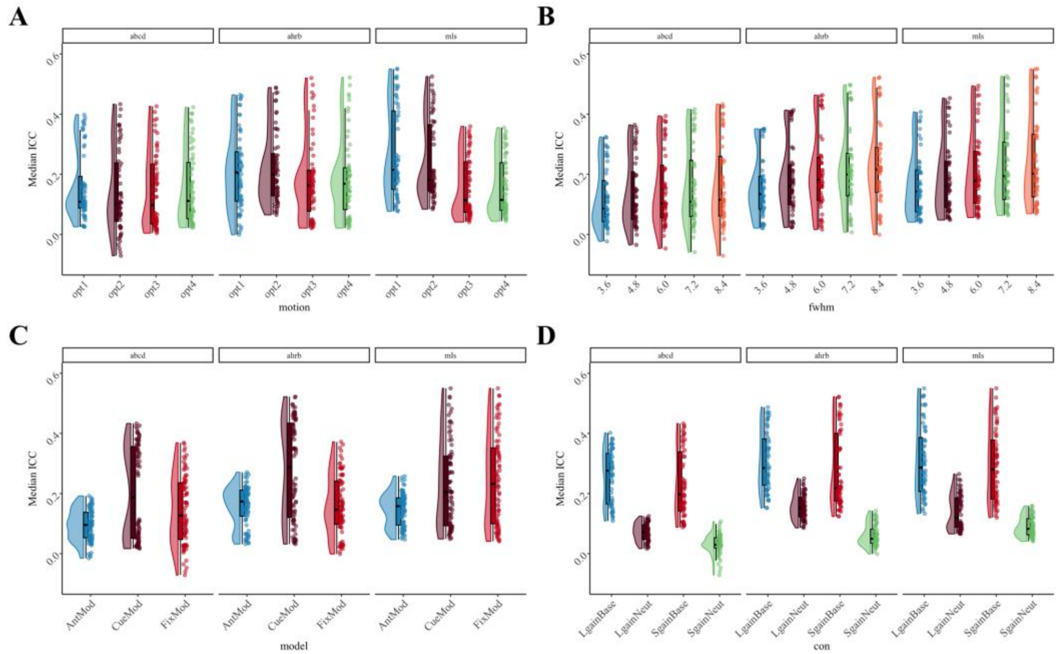
1. The authors explicitly acknowledged the main changes to the previous protocol (Stage 1) and hypotheses, as detailed on Page 22 (Section "Deviations from Stage 1 Registered Report"). I'm okay with the rationale provided in this section. I have one comment about the ABCD study, in particular regarding the need to reduce the complexity of the analyses, with a first analysis performed on N=525 but with a possible extension to 1000 subjects (though, according to Figure 6, this was not needed). The criterion "Ni & Ni -1 > .15"" (not part of Stage 1 as far as I can tell) is not clear to me. In the previous protocol (for Aim 3 in Stage 1), the analysis was planned to be repeated in intervals of 10 subjects. Also, per Liljequist et al (2019), I understand that Equation 1 (Stage 2) and the equation of Figure 1 – Part 2c (Stage 1) are equivalent but still it would be nice for the reader to explain why MSWS and MSE are equal in that particular context of Aim 1 & 2.
    a. We appreciate the opportunity to clarify. Given the two part question, we will respond as such.
        i. The change in the sampling procedure was made in response to the following request by the handling editor, "*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*". Given this comment, we made the modification during Stage 1 and implemented it in the Stage 2 version.
        ii. In Stage 1 Figure 1, we had included an illustration of the analytic workflow. In Stage 2, we trimmed the example for ICC(1) from Figure 1 and instead included the workflow for Aim 3. The difference between ICC(1), what was used in the workflow in Fig 1 in Stage 1, and ICC(3,1), what was proposed in the equation Stage 1 and used in Stage 2, differs in MSWS and MSE. Whereas MSWS is composed of the additive bias in measurement and the error in measurement, MSE is the measurement error w/o the additive bias in measurement, as the sessions are assumed to be fixed. However, in hindsight we realize that the resulting MSWS and MSBS estimates are not easily interpretable. To reduce the burden on the reader and, in part, address the reviewers comment, we did the following:

        1. First, we updated the Equation 1 to: $ICC(3,1) = \frac{MSBS - MSError}{MSBS + MSError} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_v^2}$ and refer to between-subject (σ_r^2) and within-subject variances (σ_v^2) (as described in Liljequist et al [2019; DOI: 10.1371/journal.pone.0219854, Fig 1, pg 9]). Unfortunately, the previous version used the mean squared between-subject error (MSBS) and mean squared within-subject error (MSWS) which a) does not adequately capture the components in the numerator & denominator of the ICC(3,1) formula and b) is not easily translatable. As a result, the PyReliMRI package had to be refactored to return the ICC-specific components (e.g., between-subject variance, within-subject variance/noise, between-measure variance/additive bias). Once refactored, the ICC estimates across all three estimates and studies were reran, extracted and reported for the a) between-subject variance (BS) and b) within-subject variance (WS).

2. Some of the results are difficult to explain, in particular regarding the motion correction options. It seems that no correction (Option 1) had a slightly better ICC (on average) than other options. Although stringent motion correction decreases MSWS, it also decreases MSBS, thus yielding a lower ICC on average. Do head motion artifacts increase MSBS and thus increase ICC overall? This again illustrates the difficulty of interpreting ICC (with stringent motion correction, why a decrease in MSBS is necessarily a bad thing). I already highlighted this issue in my previous feedback about Stage 1 (there is more to reliability than what one can get with the reductionist measure of ICC).
    a. As highlighted above, we have refactored our package, and revised the figures/tables and results in the manuscript. We now use between subject and within subject variance as directly associated with the ICC(3,1). With respect to motion and the ICC, as indicated in **Table 3** from the manuscript and supplemental **Figure S9** (also provided below), the results don't necessarily imply that the motion correction should *not* be used. The ICC is defined as BS/(WS + BS), which can be equivalently written as (1/(1+(WS/BS)) and from this it is clear that if one ICC is larger than another, the ratio, WS/BS, is smaller for the larger ICC. As follows intuition, more stringent motion correction improves both within- and between-subject variance (e.g., decreases for opt 3 compared to baseline) but this decrease is smaller for the within-subject variance compared to the decrease in between-subject variance. Specifically, using coefficients from **Table 3**, WS[opt3]/BS[opt3] = .75(WS[baseline])/(.6(BS[baseline]) = 1.25 WS[baseline]/BS[baseline], so the decrease in the within-subject variance was not large enough compared to the decrease in the between-subject variance to increase the ICC when using the more stringent motion correction of opt3. (note the .75 comes from (.91 - .23)/.91 and .6 =(.27 - .1)/.27, values are coefficients from Table 3). Finally**,** the reviewer is correct, the ICC is a single metric for a particular use case. Decisions about tasks and analytic decisions should be

made holistically. This is why we ended on a more cautious point in the discussions (951-978), e.g.:

i. *In the context of test-retest reliability of estimated BOLD activity, it is important to consider alternative methods to improve reliability, estimation procedures and considerations of what a 'reliable' BOLD estimate implies. In general, the evidence here illustrates that the test-retest reliability for the modified version of the MID task is consistently low using the intraclass correlation (ICC[3,1]), even at its maximum. The analytic decisions at the GLM modeling phase demonstrated improvements in reliability from between-run to between-session. Higher between-session reliability may be related to decreasing activity from early to later runs (Demidenko, Mumford, et al., 2024) or based on the sessions being an average of two runs/increased trials (Han et al., 2022; Ooi et al., 2024). In the current analyses, we focused on univariate maps and the parametric, voxelwise ICC estimation procedures (ICC[3,1]). Parametric and non-parametric multivariate methods are reported to improve reliability estimates over univariate estimates using multi-dimensional BOLD data (Gell et al., 2023; Noble et al., 2021). For example, I2C2 is a parametric method that pools variance across images to estimate a global estimate of reliability using a comparable ratio as ICC (Shou et al., 2013) and the discriminability statistic is a non-parametric statistic that is a global index of reliability testing whether the between-subject distance between voxels is greater than the within-subject voxels (Bridgeford et al., 2021). Each of these metrics uniquely summarizes the within- and between-subject variability of the estimated BOLD data and so a consensus and definition of reliability in task-fMRI remains a challenge (Bennett & Miller, 2010). In our analyses we used the ICC as it estimated the reliability for each voxel in an easy-to-interpret coefficient that is useful in common brain-behavior studies. Cut-offs from the self-report literature (Cicchetti & Sparrow, 1981) are often leveraged in fMRI research (Elliott et al., 2020; Noble et al., 2019); however, these cut-offs should depend on the optimal level of precision necessary for the question and reasonable for the methods (Bennett & Miller, 2010; Lance et al., 2006). Some recommendations have been made to use bias-corrections in developmental samples to adjust for suboptimal levels of reliability (Herting et al., 2017), but these corrections should be used cautiously as they do not account for the underlying problems of the measure or the complexities in the data that prevent accurate measurement of the latent process (Nunnally, 1978).*



A. HLM Estimates for Supra-threshold Mask

	Median ICC(3,1)			Median BS			Median WS		
Predictors	*b*	*CI*	*p*	*b*	*CI*	*p*	*b*	*CI*	*p*
(Intercept)	.23	.20 – .26	<.001	.27	.18 – .35	<.001	.91	.72 – 1.10	<.001
Reference [3.6]									
fwhm [4.8]	.02	.01 – .04	.003	-.03	-.06 – .00	.09	-.23	-.28 – -.18	<.001
fwhm [6.0]	.04	.03 – .06	<.001	-.04	-.07 – -.01	.003	-.36	-.41 – -.31	<.001
fwhm [7.2]	.06	.04 – .07	<.001	-.06	-.09 – -.03	<.001	-.44	-.49 – -.39	<.001
fwhm [8.4]	.07	.05 – .08	<.001	-.07	-.10 – -.04	<.001	-.49	-.54 – -.44	<.001
Reference [opt1]									
motion [opt2]	-.01	-.03 – .00	.07	-.04	-.06 – -.01	.01	-.14	-.18 – -.09	<.001
motion [opt3]	-.05	-.06 – -.04	<.001	-.10	-.13 – -.08	<.001	-.23	-.28 – -.19	<.001
motion [opt4]	-.05	-.06 – -.03	<.001	-.10	-.13 – -.08	<.001	-.24	-.28 – -.20	<.001
Reference [AntMod]									
model [CueMod]	.10	.09 – .11	<.001	.15	.13 – .17	<.001	.26	.23 – .30	<.001
model [FixMod]	.05	.04 – .06	<.001	.12	.10 – .14	<.001	.27	.23 – .31	<.001
Reference [LgainBase]									
con [LgainNeut]	-.17	-.18 – -.16	<.001	-.22	-.25 – -.19	<.001	-.28	-.32 – -.23	<.001
con [SgainBase]	-.02	-.04 – -.01	<.001	-.02	-.05 – .00	.09	.00	-.04 – .05	.93
con [SgainNeut]	-.23	-.24 – -.22	<.001	-.24	-.27 – -.21	<.001	-.31	-.35 – -.26	<.001

B. Analytic Category Model Impact

			New				New				New	
Comparison	χ2	Orig R2	R2	ΔR2	χ2	Orig R2	R2	ΔR2	χ2	Orig R2	R2	ΔR2
[Full] vs [New - fwhm]	95	.72	.69	.03	25	.47	.45	.02	384	.52	.31	.21
[Full] vs [New - motion]	81	.72	.69	.03	81	.47	.42	.05	138	.52	.46	.06
[Full] vs [New - model]	263	.72	.62	.10	162	.47	.37	.10	221	.52	.42	.10
[Full] vs [New - con]	864	.72	.17	.55	397	.47	.17	.30	285	.52	.38	.14

3. The authors put too much emphasis on the impact of model parametrization on reliability. While this makes sense from the current findings, it is worth mentioning that Post Hoc Analyses on Page 61 showed that model parameterization had zero impact on the ICC estimates for both left and right key brain regions (the NAc in the context of the MID task). Higher ICC values were observed for visual and motor regions (Page 38) but ICC for NAc was poor. This begs the question of how to boost reliability for key regions like NAc. Overall, ICC showed low values, indicating poor reliability for the MID task, and regardless of the analysis pipeline, the reliability remained poor even for larger sample sizes (Figure 6A). What recommendations can the authors offer to researchers interested in reward processing (e.g. they should not rely on the MID task to characterize individual differences in reward processing?)
    a. This is a good point. While model parameterization had a meaningful effect across the whole brain it did not influence the activity in the ventral striatal ROIs, which were impacted most by Contrast selection and Motion correction. We now include this *post hoc* caveat in the discussion “*Notably, ICCs in this post hoc region were not meaningfully impacted by Model Parameterization but were impacted by Contrast and Motion correction, suggesting that test-retest reliability may be uniquely impacted by analytic strategy depending on the voxels under consideration. These findings illustrate that the test-retest reliability of the MID task is relatively low, even in the most common ROI such as the Left and Right NAc.”* in the discussion on line 873-876 and conclusion on line 1008-1012, “*While Model Parameterization and Contrast selection had the largest impact on voxelwise ICCs, further work is needed to expand on these findings by evaluating alternative brain regions and analytic decisions that may result in improved test-retest reliability that may be meaningful in individual differences research.*” Based on the analyses, the broad conclusion is that test-retest reliability in the ventral striatal regions is in the poor range. Given that analytic decisions may vary across regions, we do encourage future researchers to evaluate this question further on line 882-884, “*To understand how analytic strategies differentially impact ICCs in different brain regions, we encourage future researchers to use the publicly available estimated maps to probe this question further*”

4. I feel that the results of the subthreshold task voxels (voxels with z < 3.1) are not well reported or exploited (they read like a distraction from the main conclusions). Even the authors mentioned for Aim 2 (in Page 34) that “We avoid interpreting the sub-threshold mask as it includes regions that are high-noise”. If one (obviously) expects high MSBS and MSWS for the subthreshold maps, then the rationale for including these maps in the first place becomes weak given that MID has poor reliability in general. I would suggest (if this is doable) that the authors add another post hoc analysis to assess the reliability of the DMN regions (these regions are expected to be consistently deactivated across subjects).
    a. The motivation to not interpret sub-threshold maps for MSBS/MSWS was due to the wide range of values that would be obtained from the masked voxels. The supplemental Figure S26 demonstrates the outlier values in dropout regions (mOFC), ventricles and the ventral cerebellum, which are likely due to high noise/variance. As stated above, we include in the discussion “*To understand how analytic strategies differentially impact ICCs in different brain regions, we encourage future researchers to use the publicly available estimated maps to probe this question further.*” One comment by the editor (above) was that, “*The reader has a huge amount to process in this paper, and by the time I got to this point I was running out of steam.*”. Hence, running more analyses may further burden the reader.
5. The authors hypothesized that the reliability within sessions would be greater than between sessions. However, the data showed the reverse: between-session estimates were consistently higher than between-run estimates of reliability. The authors proposed an explanation in the discussion section that within-session effects might be decreasing across runs. The reader might get the impression that splitting a session into multiple runs is a bad strategy (I hope I’m reading correctly all these supplementary figures). It would be nice to hear the author's opinion on the use of multiple runs for the MID task.
    a. We highlight in the discussion that it may be the result of the change in activity from run 1 to run 2 but also the effect of power/N trials in the discussion section on line 956-959 “*Higher between-session reliability may be related to decreasing activity from early to later runs (Demidenko, Mumford, et al., 2024) or based on the sessions being an average of two runs/increased trials (Han et al., 2022; Ooi et al., 2024)*.” We do not recommend that splitting sessions into runs is bad, rather, that the change in activity between runs and the number of trials may reduce the ICC estimates in this context.
6. For spatial smoothing, higher fwhm (8.4 mm) yielded better ICC values. As this kernel size was the largest, it seems that the trend would still hold for higher fwhm values. But maybe there is a range of fwhm where the ICC would start decreasing (very large fwhm might result in lower MSBS). I would like to know the authors' opinion on optimal fwhm values (e.g., we typically read in the SPM community that a fwhm of around 2 or 3 times the voxel size should be used).
    a. The reviewer is correct, there is likely a plateau in benefits in ICC as the FWHM smoothing kernels increase. As we noted in the discussion on line 922-925, smoothing should be protocol/region specific: “*Decisions to smooth in the MID*

task are especially important given that larger smoothing kernels have been reported to spatially bias reward-related activity in the MID task (Sacchet & Knutson, 2013).” The cost-to-benefit in smoothing suffers from comparable drawbacks as from the radius used to define an ROI. For instance, in case of the NAc, defining a larger ROI sphere has pros and cons. The pros, it would result in a greater number of voxels and more stable average. The con, it may come at the cost of validity and lower functional precision. Larger spheres, as with larger smoothing kernels, will blend signal/noise from adjacent non-interest regions, such as the WM, CSF and other regions. As described in Ward (2020; DOI: 10.1093/bjps/axz027), decisions should ultimately come down to the research question. Using standard practices is a reasonable approach but protocols are including smaller and smaller voxel sizes. This may modify the inherent smoothness when resampling a 1.4 mm voxel versus 3 mm voxel into 2 mm standard space and so researchers should confirm how smooth their data are after applying a smoothing kernel 2-3x voxel size. Using blanket standards is a common pitfall even in the case of the ICC. Rather than using the cut-off that is appropriate for the data/research question, researchers traditionally use cut-offs as if though they are golden rules (Lance et al., 2006, DOI: 10.1177/1094428105284919), which may not make sense for all use cases.

7. What is the take-home message for the fMRI community? If one has a task with poor reliability (low ICC) using standard analysis pipelines, then no preprocessing or modeling strategy can substantially improve its reliability.
    a. The take home message: If researchers are focused on whole brain analyses, they should invest time in understanding their model parameterization and contrast selection. To date, there has been an overemphasis on stating that “Reliability in task fMRI is bad." but we’re only beginning to scratch the surface on how we may improve reliability and optimize our tasks for individual differences research. We hope that our multiverse analyses and publicly shared/data maps encourage others to explore this question further. We include in the conclusion on line 1008-1012
        i. “*While Model Parameterization and Contrast selection had the largest impact on voxelwise ICCs, further work is needed to expand on these findings by evaluating alternative brain regions and analytic decisions that may result in improved test-retest reliability that may be meaningful in individual differences research*.”
8. The label “Figure 2” is mistakenly used twice for different illustrations (Page 22 and Page 25).
    a. Updated.