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.

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.

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).

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?)

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).

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.

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).

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.

The label "Figure 2" is mistakenly used twice for different illustrations (Page 22 and Page 25).