



Exploring determinants of test-retest reliability in fMRI: a study with the Monetary Incentive Delay Task

A recommendation by [Dorothy Bishop](#)  based on peer reviews by [Xiangzhen Kong](#) and 1 anonymous reviewer of the STAGE 2 REPORT:

Michael I. Demidenko, Jeanette A. Mumford, Russell A. Poldrack (2024) Impact of analytic decisions on test-retest reliability of individual and group estimates in functional magnetic resonance imaging: a multiverse analysis using the monetary incentive delay task. bioRxiv, ver. 5, peer-reviewed and recommended by Peer Community in Registered Reports. <https://doi.org/10.1101/2024.03.19.585755>

Submitted: 25 March 2024, Recommended: 09 July 2024

Cite this recommendation as:

Bishop, D. (2024) Exploring determinants of test-retest reliability in fMRI: a study with the Monetary Incentive Delay Task. *Peer Community in Registered Reports*, 100747. [10.24072/pci.rr.100747](https://doi.org/10.24072/pci.rr.100747)

Published: 09 July 2024

Copyright: This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Functional magnetic resonance imaging has been used to explore brain-behaviour relationships for many years, with proliferation of a wide range of sophisticated analytic procedures. However, rather scant attention has been paid to the reliability of findings. Concerns have been growing failures to replicate findings in some fields, but it is hard to know how far this is a consequence of underpowered studies, flexible analytic pipelines, or variability within and between participants.

Demidenko et al. (2024) took advantage of the availability of three existing datasets, including the Adolescent Brain Cognitive Development (ABCD) study, the Michigan Longitudinal Study, and the Adolescent Risk Behavior Study, which all included a version of the Monetary Incentive Delay task measured in two sessions. These were entered into a multiverse analysis, which considered how within-subject and between-subject variance varies according to four analytic factors: smoothing (5 levels), motion correction (6 levels), task modelling (3 levels) and task contrasts (4 levels). They also considered how sample size affects estimates of reliability.

The results have important implications for the those using fMRI with the Monetary Incentive Delay Task, and also raise questions more broadly about use of fMRI indices to study individual differences. Motion correction had relatively little impact on the ICC, and the effect size of the smoothing kernel was modest. Larger impacts on reliability were associated with choice of contrast (implicit baseline giving larger effects) and task parameterization. But perhaps the most sobering message from this analysis is that although activation maps

from group data were reasonably reliable, the ICC, used as an index of reliability for individual levels of activation, was consistently low. This raises questions about the suitability of the Monetary Incentive Delay Task for studying individual differences. Another point is that reliability estimates become more stable as sample size increases; researchers may want to consider whether the trade-off between cost and gain in precision is justified for sample sizes above 250.

I did a quick literature search on Web of Science: at the time of writing the search term ("Monetary Delay Task" AND fMRI) yielded 410 returns, indicating that this is a popular method in cognitive neuroscience. The detailed analyses reported here will repay study for those who are planning further research using this task.

The Stage 2 manuscript was evaluated over one round of in-depth review. Based on detailed responses to the reviewer's and recommender's comments, the recommender judged that the manuscript met the Stage 2 criteria and awarded a positive recommendation. **URL to the preregistered Stage 1 protocol:** <https://osf.io/nqgeh> **Level of bias control achieved: Level 2.** *At least some data/evidence that was used to answer the research question had been accessed and partially observed by the authors prior to IPA, but the authors certify that they had not yet sufficiently observed the key variables within the data to be able to answer the research questions and they took additional steps to maximise bias control and rigour.*

List of eligible PCI RR-friendly journals:

- [Brain and Neuroscience Advances](#)
- [Cortex](#)
- [Imaging Neuroscience](#)
- [In&Vertebrates](#)
- [NeuroImage: Reports](#)
- [Peer Community Journal](#)
- [PeerJ](#)
- [Royal Society Open Science](#)

References

1. Demidenko, M. I., Mumford, J. A., & Poldrack, R. A. (2024). Impact of analytic decisions on test-retest reliability of individual and group estimates in functional magnetic resonance imaging: a multiverse analysis using the monetary incentive delay task [Stage 2]. Acceptance of Version 5 by Peer Community in Registered Reports. <https://www.biorxiv.org/content/10.1101/2024.03.19.585755v4>

Reviews

Evaluation round #1

DOI or URL of the preprint: <https://www.biorxiv.org/content/10.1101/2024.03.19.585755v3>
Version of the preprint: 1

Authors' reply, 29 June 2024

The responses to the editor's and reviewer's comments, and the tracked changes (highlighted yellow) document from Stage 2 to Stage 2 resubmission are in the attached PDFs.

[Download author's reply](#)

[Download tracked changes file](#)

Decision by [Dorothy Bishop](#) , posted 17 June 2024, validated 17 June 2024

Exploring determinants of test-retest reliability in fMRI

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.

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.

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?

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.

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.

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.

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).

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.

minor typo issues

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;

L52: semicolon before 'however' (starts a new sentence)

L179: would help if the list of analytic decisions was in the same order as in Table 1

L204 "will stabilize at a sample size between"

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.

L211 and elsewhere, Spearman with capital S

L290, should this be Table S3

L351: Github with capital G

L406: delete 'the'

last para p 21: space before units 'mm'

L741: 'is the opposite of the Fixation model'

L 769: the variability will decrease as a function of square root of N

L844; semicolon before 'however'.

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.

Supplementary material

L 24 "some do not have the necessary..." - is this an explanation for why some subjects were excluded? It's a bit unclear

Figure S2 - did you mean to retain this in the paper? It says IGNORE THIS RESULT in big red writing

L2

====

Note added by Managing Board: To accommodate reviewer and recommender holiday schedules, we will be closed to submissions from 1st July — 1st September. During this time, reviewers will be able to submit reviews and recommenders will issue decisions, but no new or revised submissions can be made by authors. **If you wish your revised manuscript to be evaluated before 1st September, please be sure to submit no later than 30th June.**

Reviewed by [Xiangzhen Kong](#), 06 April 2024

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.

Reviewed by anonymous reviewer 1, 14 June 2024

See attached file. Thank you!

[Download the review](#)