Response to Reviewers (ArticleID #591 (version: 1), Rasmussen et al.)

Dear Dr Maxine Sherman and reviewers,

Thank you for allowing us to resubmit our revised manuscript; *On the neural substrates of mind wandering and dynamic thought: A drug and brain stimulation study*. We have provided a detailed response to each of the reviewer comments in this document. We have also uploaded the revised version of the manuscript to our OSF folder, and we have attached a version with the changes tracked to this submission.

Reviewer #1 (Anonymous):

R1.1: *If the primary focus of the research is the interaction between tDCS and dopamine on mind wandering, it might be worth considering omitting the manipulation of tDCS dosage and instead increasing the sample size to 60 per group. A recognized issue in tDCS research is the lack of replicability. Improving statistical power through a simpler design and a larger sample size could potentially enhance result stability.*

> First, replication is an issue for all science, not especially tDCS. Second, we agree that increasing the sample size is more valuable than including the dosage manipulation. Thus, we have omitted the 1mA groups (levodopa and placebo) and increased the remaining 4 groups to a maximum of 60 participants per group. This is now changed throughout the manuscript (for example see p. 10 and 13 of the tracked manuscript):
>
> "This study will employ a between subjects' design, whereby participants will be pseudo-randomly allocated to one group according to the following variables: drug (Levodopa, Placebo) x stimulation condition (anodal HD-tDCS, sham HD-tDCS). Thus, the conditions will be: (1) sham HD-tDCS and placebo; (2) sham HD-tDCS and levodopa; (3) 2mA anodal HD-tDCS and placebo; (4) 2mA anodal HD-tDCS and levodopa."
>
> "Participants will continue to be recruited until a Bayes Factor $(BF)_{10} > 6$ or $BF_{01} > 6$ has been reached for the selected hypothesis tests (see above), or until the maximum sample size of 240 complete datasets (60 participants per group; the maximum number dictated by resource constraints) is reached."

R1.2: *The interpretation of null results for the hypothesis "Are relevant individual traits balanced between conditions?" should refrain from concluding the absence of an impact of personal traits in the sample. The sample size per group might be insufficient to explore relationships between traits and mind wandering adequately. Research on clinical samples has demonstrated the prevalence of ADHD in mind wandering among children (Frick et al., 2019, doi: 10.1111/bjc.12241). Previous studies on ADHD-like traits in community samples have involved >100 subjects to yield relevant results (Vatansever et al., 2018, doi:*

10.1017/S0033291718003598; Franklin et al., 2017, doi: 10.1177/1087054714543494). I recommend revising the implications of the null results in this context.

> The hypothesis regarding individual traits is not designed to directly explore the impact of personal traits on the results, it is only for the purpose of ensuring that there are no differences in the presentation of these traits between the four groups. We have revised the manuscript to be clearer that the findings relate to the comparison between groups and not the impact of any individual differences on the results (p. 8):
>
> "A null effect would suggest that there are no meaningful differences in these traits between the four conditions."

Reviewer #2 (Jonathan Smallwood):

R2.1: *My main concern with this study is that it focuses on a single task. Multiple studies have shown that patterns of ongoing thought that are unrelated to the external envioronment vary in terms of their occurence in terms of how hard the external task was. In work from my lab, we have found that in a simple task the occurence of off task states are much more prevalent than in a more demanding task (e.g. Konisih et al., 2017, Cognition; Turnbull et al., 2018, Neuroimage; Turnbull et al., 2019, Nature Communications) and there are countless other examples in the literature. In this context, cognitive control plays an important role in regulating the tendency for people their thoughts in line with the task (a process we call context regulation of thought). In the brain this is linked to the ventral attention system (As defined by Yeo and colleagues, 2011) and in particular the left dorso lateral prefrontal cortex. In our work we found that dLPFC suppress off task thought when tasks are hard but facilitates it when they are easier (Turnbull et al., 2019).*

My concern with the proposed study is that with a task with only one level of difficulty, it will be very hard to properly characterise the role that dopamine plays in regulating thought, especially given the u-shaped relationships that the authors argues underpin the neurpharmaological effects. A simple remedy would be to vary the demands of the task in which spontaneous thought happens, in particular using one task in which the level of dynamic thought is equivalent to what is seen under no task condition, and a second where the experience is reduced. If the study used this design then it would mean that the possibility that dopamine can facilitate both dynamic thought and task focused thought can be seperated (i.e. it should facilitate dynamic thought in the simple task and suppress it in the harder task). If this is not possible, then it would need to be the case that the authors can explicitly caveat the effect, however, it would be a shame to not take advantage of this simple design change to address such a cool and important question.

> We agree task difficulty is a very interesting manipulation to include, and it is something that we would really like to explore in future studies. However, for the purpose of this study it is unfortunately not feasible to include an additional task. Specifically, to include an additional task we would have to do one of the following:

1. Half the current number of trails or participants per conditions, which would substantially weaken our design and power to detect any effects. Indeed, the other two reviewers have requested an increase in subject numbers per group;
2. Double the duration of the task and stimulation, which is outside our ethical approval and would have additional fatigue effects;
3. Double the sample size to incorporate four additional groups who complete an alternative task with low cognitive demands and the same stimulation x drug manipulations. This is unfortunately not feasible at this stage, as this would result in increasing the sample size to 480 participants (60 for each group) and over 1200 hours of testing.

We will ensure that at Stage 2, in the discussion, appropriate caveats regarding any effects of dopamine on mind wandering will be presented. In addition, we will detail that mind wandering varies depending on the demands of a task, thus any results will be most applicable in contexts which require cognitive control. Furthermore, we have ensured that we do not overstep with the claims we make at Stage 1, regarding the conclusions we can draw from the current study design. For example, see p. 5 of the manuscript where we have clarified the effects are in the context of a cognitively demanding task.

"To understand the neurochemical mechanisms underlying mind wandering and the dynamic thought types, this research also aims to investigate whether the changes in mind wandering and dynamic thought, while completing a cognitively demanding task, are being driven by changes in dopamine availability."

Reviewer #3 (Chris Chambers):

R3.1: *In the study design table you note: "BF10 > 6 or BF01 > 6 for stopping rule is enough evidence to establish a meaningful result. For all tests BF10 > 3 or BF01 > 3 is supported by the literature as enough evidence to establish a meaningful result. In addition, if the credible intervals (CIs) do not cross 0 for the probit modelling, this will be interpreted as a meaningful effect." This needs some clarification and possibly simplification. What criteria exactly will determine a definitive outcome? From the 1st sentence it appears to BF>6. But the 2nd sentence seems to downgrade that to BF>3, and the 3rd sentence adds a whole other decision criterion. Is there a difference between these for critical and non-critical hypotheses? This needs to be crystal clear.*

We apologise for the confusion regarding the rationale for the test sensitivity. We have now clarified this in the design table, whereby we will employ $BF_{10} > 6$ or $BF_{01} > 6$ for the interpretation of meaningful effects for all t-tests and for all analyses where we employ the hierarchical order probit modelling, we will use credible intervals to interpret these findings, as this is the appropriate interpretation for the Bayesian modelling analyses. Please refer to the design table in the tracked manuscript to review these changes (p. 6-9):

"**For all hierarchical order probit modelling analyses:** If the credible intervals (CIs) do not cross 0, this will be interpreted as a meaningful effect.
For all other tests:
$BF_{10} > 6$ or $BF_{01} > 6$ will be interpreted as enough evidence to establish a meaningful result."

R3.2: *You discuss "critical" hypotheses and in the design table mention "non-crucial" tests. Does "critical" mean hypotheses that form the basis of the stopping rule and "non-crucial" mean those that are not? This would benefit from some clarification. Perhaps use comparable terminology between them (e.g. critical and non-critical, rather than critical and non-crucial) and make clear for each hypothesis which type it is and what this means. In doing so, make explicit the precise conditions under which testing will stop.*

We have now provided further clarification regarding the role of each analysis in the design table (i.e., if they are hypothesis testing or control analyses). Furthermore, both in the table and in-text we have made it clear what tests are being used for the stopping rule, however we have removed the term "critical" in-text to avoid any confusion regarding the value of these tests for our analysis interpretation. These changes are now clear in the design table in the manuscript (p. 6-9) and throughout the Methods and Proposed Analyses (for example see p. 13 and p. 23):

"Participants will continue to be recruited until a Bayes Factor $(BF)_{10} > 6$ or $BF_{01} > 6$ has been reached for the selected hypothesis tests (see above), or until the maximum sample size of 240 complete datasets (60 participants per group; the maximum number dictated by resource constraints) is reached. This is larger than the sample size which has been used previously to find meaningful results (Rasmussen et al., 2023) and we believe inconclusive results in the chosen tests at this sample size will still offer an important contribution to the literature." (p. 13)

"While tests will be conducted for both thought probes, the results from the freely moving thought t-test is one of the two tests used for the stopping rule in this study, as there was evidence for 2mA anodal stimulation reducing freely moving thought (Rasmussen et al., 2023)." (p. 23)

R3.3: *In the design table you state "BF > 6 or BF > 6 is supported by the literature as enough evidence..." Is the double mention of "BF>6" a typo?*

This was an error on our part, however the rationale for deciding test sensitivity column has now been written more concisely and this statement was removed in the process (see our response to R3.1).

R3.4: *The hypotheses concerning tDCS amplitude are non-directional, which I assume is a deliberate decision due to the mixed results of previous studies. Given the (apparent?) lack of a*

clear rationale for why 1mA tDCS should produce a greater effect than 2mA tDCS, I found myself wondering whether the 1mA condition is necessary at all. If you removed the 1mA condition, you could increase the sample size for a comparison of 2mA vs sham and perform a more sensitive test of the interaction between tDCS x drug. Then, in the event of a positive result, a later study (in a whole new RR) could then hone in on the dosage necessary to cause that effect. Even though this is how I would run the study, I offer it only as a suggestion to consider rather than a strong recommendation. However, if you do keep the tDCS dosage manipulation I would suggest strengthening the rationale for it in the introduction. For me, it really only makes sense to include it if there is some reason for supposing that 1mA tDCS might be more effective than 2mA tDCS.

See our response to R1.1.

R3.5: *There are good reasons for adopting a between-subjects design rather than a within-subjects design (including the fact that it helps to better preserve blinding of tDCS intensity, active vs placebo, and participant demand characteristics) but given the relative high cost in statistical sensitivity associated with between-subjects designs (compared to within), I would recommend including a justification of this particular design choice in the method.*

We have now revised the manuscript to incorporate a justification for the between-subjects approach in the methods section (p. 10).

"A between-subjects approach is most appropriate for this study as it will help preserve the integrity of the stimulation and dopamine blinding. Further, it reduces the likelihood of practice effects in the task or any inter-session changes, which are associated with within-subjects designs."

"This method is designed to reduce the likelihood of group-related confounds in the between-groups design."

R3.6: *You set a maximum sample size of 40 per group across the 6 groups due to resource constraints. It would ideal to include some Bayes Factor Design Analyses (using the exact analysis methods for each hypothesis) to determine just how sensitive this sample size will be able to detect effects of various sizes given the chosen prior (see https://link.springer.com/article/10.3758/s13428-018-01189-8 and https://link.springer.com/article/10.3758/s13423-017-1230-y). I suspect that the design, as currently proposed, would be sufficient to detect only quite large effects, in which case this limitation should be noted. A BFDA would help make this clear.*

Please see our response to R1.1. We have now substantially increased our sample. One of the strengths of a Bayesian approach as that one accumulates evidence towards a hypothesis and doesn't have to rely on probability of a result as is the case with null hypothesis significance testing. We have chosen a standard prior as it would be

speculative to select a different value given the interaction of dopamine, tDCS and mind wandering has not been assessed previously.

R3.7: *On p13 you note: "Participants will also be excluded from the study and replaced during the testing phase if their responses to the end of session questionnaire suggest that the participant did not understand how to correctly generate random number sequences. An example which would suggest the task has not been completed correctly would be if the participant cites a specific pattern that they used to approach the task (e.g., they repetitively used z,z,z,m,m,m,z,z,z,m,m,m to generate the sequences)." This strikes me as a sensible general rule but for a Stage 1 RR needs to be defined comprehensively and precisely, making clear the exact parameters under which a participant's response will be judged to be sufficiently non-random to warrant exclusion. This must fully pre-specified and reproducible and eliminate all possible researcher degrees of freedom in both the definition and implementation.*

We have now made the exclusion regarding participants randomness clearer in the manuscript and we have also emphasised that this will be done before the experimenter is unblinded to avoid any potential biases when removing these participants (p. 14).

"Specifically, if participants cite that they used the same pattern throughout which repeated more than twice at a time (e.g. *z,z,z,m,m,z,z,z,m,m*) or if they state that they only alternated from one key to the other in the same order throughout, with one or more taps on each key at a time (e.g. *z,z,z,m,m,m,z,z,z,m,m,m*, or *z,m,z,m*), they will be excluded. … Participants who meet any of the above exclusion criteria will be removed before the experimenter is unblinded to the data and they will be replaced during the data collection phase."

R3.8: *p16: "At the end of the session, participants and the experimenter will be asked to select which dopamine drug manipulation group they were in (levodopa, placebo), to assess the efficacy of the drug manipulation blinding". Please specify exactly how the outcome of this test will be taken into account in the analysis and interpretation.*

The blinding analyses are conducted separately to the hypothesis driven tests, as they are specifically designed to ensure the study's integrity. We have added clarification on how this data will be analysed into the manuscript (p. 17). We also have a description of the interpretation for this test in the design table (p. 9).

"This data will be used to compare the proportion of correct guesses across both conditions, whereby a lower proportion of correct guesses across the two groups would indicate that the blinding was effective for these conditions."

R3.9: *p17: "The stimulation will be immediately terminated if participants report experiencing any discomfort, or if there are any technical difficulties, including if Nurostym device identifies that the electrode impedances are too high, and self-terminates the stimulation." Presumably*

participants who are excluded due to discomfort will be replaced? This should be stated explicitly as an exclusion criterion.

> Participants who are excluded due to these criteria will be replaced and this has now been moved into the exclusion criteria section of the manuscript (p. 14).
>
> "This will also include if participants report any discomfort from the stimulation or if the Nurostym device identifies that the electrode impedances are too high and self-terminates the stimulation. Participants who meet any of the above exclusion criteria will be removed before the experimenter is unblinded to the data and they will be replaced during the data collection phase."

R3.10: *p17: "Given the FT-RSGT requires participants to respond accurately in time to a metronome tone, it is important to account for any influence of video game or musical training on participant's response variability". How exactly will this be accounted for in the confirmatory analyses?*

> We control for the influence of these factors by including this data in the randomisation script when allocating all participants to their groups. We have now specified that in this section of the manuscript (p. 19).
>
> "Thus, at the beginning of the session, participants will be asked how many hours they spend playing video games and musical instruments each week, and this information will be entered into the randomisation script to ensure these two variables are balanced across the four groups (see supplementary materials)."

R3.11: *Would it make sense to survey participants for the level of tDCS discomfort at the end of the session to account for potential disruptive effects of higher intensity stimulation on task performance? This seems particularly salient given that the tDCS is administered during the FT-RSGT, so (hypothetically) if 2mA stimulation happened to be significantly more uncomfortable than 1mA stimulation (yet insufficiently painful to lead to exclusion), the additional distraction could potentially explain differences in performance between 1mA vs 2mA over and above any cortical effects. One reason I mention this is that we have previously found in our own tDCS experiments that prefrontal 2mA stimulation can be painful in some participants (e.g. see footnote 1 on pp12-13 here: https://ore.exeter.ac.uk/repository/bitstream/handle/10871/124886/Sedgmond_BehaviouralNeuroscience_2020.pdf; in that study we actually had to turn it down to 1.5mA after Stage 1 IPA).*

> As an oversight, we had missed including this questionnaire in the supplementary materials, however we do include an adverse effects questionnaire in all our lab's tDCS studies, which participants complete at the beginning and the end of the session to assess any changes due to the stimulation (please refer to the supplementary materials). In our previous Registered Report study which applied 2mA stimulation to the PFC, alongside

an additional two brain regions, we only had one participant out of 250 who requested that we terminate the stimulation due to it being uncomfortable. Furthermore, very few participants reported experiencing any uncomfortable sensations at the end of the session, and participants were relatively poor at indicating whether they received active or sham stimulation (~ 50% accuracy, with 115 out of 228 participants making the correct judgement). Given that now all participants will receive the same dosage, we don't anticipate this will have an effect on the results, thus we will not directly analyse this data for the purpose of this study.

R3.12: *pp19-20, section "Post-study data exclusion": will participants who are excluded due to being outliers within their group be replaced? I'm assuming so, but please state this explicitly. In addition, re: "Finally, to ensure that extreme outliers during the task do not skew any time on task effects, individual trials which are greater than 3 standard deviations above or below the mean for each group's approximate entropy and behavioural variability scores will also be removed from the analyses." What % of individual trials would need to be removed before an entire participant would be excluded and replaced outright? Are there any other data-based exclusion criteria of any kind, either at the level of trials within participants or participants within the sample? I recommend reviewing these very carefully as these criteria cannot (in general) be changed after IPA.*

The post-study data exclusions will only be employed once data collection is completed, to ensure that we do not continue to remove participants who may be considered outliers during the experiment, but in fact, lie within the appropriate range of the complete sample. In addition, with the individual trial exclusions, the nature of the standard deviation cut offs mean that only a very small proportion of trials (less than 2% of the data overall) will be removed, thus it will not be a great enough percentage of data to justify a cut off for removing entire participants. These additions have now been stated in the manuscript (p. 21).

"The post-study exclusion criteria for participants will replicate the criteria used by Rasmussen et al. (2023) and will be employed once data collection is completed, thus these participants will not be replaced in the final sample."

R3.13: *p20: concerning the probit modelling, the authors note that "there will be several predictor variables, alongside their interactions, however participants stimulation condition (2mA active, 1mA, or sham) or dopamine condition (levodopa vs. placebo) will be entered in as the key predictor for the respective analyses, which are explained in detail below." Please specify the full range of predictor variables and parameters. To nail this down precisely, I recommend including an analysis script as part of the Stage 1 revision based on simulated data.*

We have now added some further clarification to this explanation of the probit modelling (see p. 21), however the section being referred to here is designed to be an overview of the modelling approach and the specific predictors which are used to investigate each

research question are explained in the respective analysis descriptions below the modelling introduction. We have also made this clearer in the manuscript.

"Participants thought probe responses will be entered as the dependent variable into the models and there will be several predictor variables, relating to the measures of task performance (behavioural variability and approximate entropy), block and trial data, alongside their interactions, however participants stimulation condition (2mA active vs. sham) or dopamine condition (levodopa vs. placebo) will be entered in as the key predictor for the respective analyses. The specific predictors employed in each probit model are explained in detail below."

We will be replicating the 23-model structure employed by Rasmussen et al. (2023), with the only change to the listed model definitions being to the "condition" predictor which will alternate between "stimulation", "dopamine" and "stimulation*dopamine". Thus, we have included the code for Rasmussen et al. (2023) here: https://doi.org/10.48610/74fcc20

Minor

R3.14: *Typo: check spelling for Nerostym / Nurostym as both are used in different places*

The typo has now been corrected to be Nurostym throughout (see p. 18).

R3.15: *P19: alterative > alternative (multiple instances)*

This error has now been corrected in the manuscript to be alternative throughout (see p. 20).