

Article title: How Effortful is Boredom? Studying Self-Control Demands Through Pupillometry - Registered Report Stage 1

Note: the revised passages are highlighted in orange.

Decision for round #2 : *Revision needed*

Invitation to respond to reviewer comments

I have now received reviews from all four experts from the first round of reviews. They all appreciate your thoughtful responses and have indicated that the study, as proposed in the current version, would be an interesting and important contribution to the literature.

Thank you for your constructive feedback. We are pleased to learn that the reviewers find our responses thoughtful and consider our study an interesting and important contribution to the literature. We appreciate the guidance provided and have carefully considered the remaining concerns.

A few remaining concerns have been identified, two of which may be relatively easy to address: (1) a more nuanced characterization of relevant theories (anonymous reviewer 1) and (2) clarification of details of the pilot data (Julia Englert).

- 1. Nuanced Characterization of Relevant Theories: We have supplemented our manuscript to provide a more nuanced characterization of relevant theories, including additional perspectives on the experience of effort. Specifically, we have incorporated insights from Kurzban (2016) to ensure our discussion is aligned with the broader theoretical context.**

We now write (p. 3f): *Cognitive effort which can be defined as “intensity of mental [...] work that organisms apply towards some outcome” (Inzlicht et al., 2018), can be measured objectively (e.g., with pupillometry) or experienced subjectively (Bijleveld, 2018; Robinson & Morsella, 2014) which we will refer to as perceived effort¹. A large body of research shows that the objective and the perceived investment of effort tends to feel unpleasant and aversive (David et al., 2022; Kool & Botvinick, 2018; Wolff et al., 2021). **On the other hand, effort is considered to feel unpleasant due to the costs of continuing with the ongoing activity (Kurzban, 2016). Thus, while effort is instrumental for effective self-control, it appears to carry a momentary cost, and the prolonged exertion of effort creates cumulative costs, such as fatigue or tiredness (Ainslie, 2021; Hopstaken et al., 2015; Kurzban, 2016; Kurzban et al., 2013; Westbrook et al., 2013). In addition to efforts’ intrinsic costs (Kool et al., 2013), mobilizing effort towards one goal creates opportunity costs: When we direct our effort towards one activity (e.g., exercising to tackle backpain), we have to forego other, potentially more rewarding activities (e.g., going for a walk with friends). In this case, the aversive sensation of effort is considered to index the costs of continuing with the ongoing activity (Kurzban, 2016). This can create added self-control demands that make pursuing one’s goal and continuing with a task even more self-control demanding and by extension more effortful***

¹ Please note that from now on, we will use the term “effort” to refer to “cognitive effort” for the sake of readability.

(Kurzban et al., 2013). In light of effort's costs (intrinsic and/or opportunity costs), people are selective about when to invest effort in the service of self-control and the sensation of effort has been conceived as an index for the momentary costs of self-control (Bieleke et al., 2023).

2. Clarification of Pilot Data: We have clarified that our pilot data is based on a 100 % congruent and a 50% incongruent Stroop task. We now write (p. 31):

Lastly, while the hard Stroop task in the online study consisted of 50% incongruent trials, the hard Stroop task in the present study will consist of 100% incongruent trials. This change is implemented with the aim to (1) increase the overall difficulty of the hard task and (2) enhance the clarity of our analysis, particularly regarding pupillary reactions, as pupil size is greater in incongruent than in congruent trials. Otherwise, by calculating the mean of the five trials preceding the probe, congruent and incongruent trials would get mixed up.

In addition, there are questions about the reliability and validity of subjective effort due to boredom vs. effort, and regarding the design of the Stroop task. It is not mandatory that you make any changes to the current protocol in response to these two aspects. However, I encourage you to carefully consider these points as potential limitations of the study. Accordingly, if you decide to make no changes to the study protocol, these potential limitations and their impact on the study results should be discussed in your eventual Stage 2 report. I very much look forward to your point-by-point response.

- 1. Reliability and validity of subjective effort due to boredom and due to task difficulty: We acknowledge the significant concerns regarding the reliability and validity of subjective reports of effort due to boredom and effort due to task difficulty. While we have considered implementing additional items for a factor analytical approach, we have not identified other items that we think would capture these specific constructs. We will clearly discuss the potential limitations of these unvalidated items and their impact on our results in our Stage 2 report.**
- 2. Design of the Stroop Task: We have discussed on the proportion of congruent and incongruent trials in our hard Stroop task. We recognize the trade-offs involved and have decided to use the 100% incongruent trial version for the hard task (although before the first round of revision, we planned to use a 50% incongruent version) with the aim to maximize the overall difficulty of our hard task and to enhance clarity of the analysis and interpretation of the pupillary data. By calculating the mean of the five trials preceding the probe, congruent and incongruent trials would get mixed up although the pupil size varies largely between congruent and incongruent trials.**

We will mention in our discussion the relevance of conducting further studies using different difficulty levels and proportions of congruent and incongruent trials to enhance our understanding on the sensation of effort caused by boredom.

We appreciate the reviewers' and your support and constructive suggestions. We believe these revisions and acknowledgments will strengthen our manuscript and look forward to your next feedback.

by [Thomas Meyer](#), 10 Jul 2024 21:20

Manuscript: https://osf.io/dqvm5?view_only=12191f02a5db4689b00b42bab7dbd522

version: 1

Review by Julia Englert, 08 Jul 2024 16:52

The authors have gone to great lengths to improve the theoretical parts of their manuscript and have also collected additional data. They have greatly expanded the theoretical reasoning and discussion, and take a much more comprehensive look at the existing evidence, and potential critical points.

My only remaining substantive concerns are still about the design of the Stroop task, though the current design does in no way invalidate the proposed research and I believe it is up to the authors whether they wish to incorporate any suggestions. The design will be informative about the main aim either way. Your design allows comparing the two tasks globally - there is always a virtually infinite number of possible controls one can add, and they can be left for later studies. As a stand-in for an “easy, boring” versus a “cognitively effortful” task, the manipulation works.

(However, there was also a related point of confusion about the pilot data they collected online.)

In any case, am looking forward to the finalized version of the experiment and the results.

Regarding the Stroop Design:

As mentioned, proceeding as planned should not stand in the way of acceptance.

1. But I still think that sticking with a 100% (in)congruent Stroop task and a total of two conditions limits the interpretation of results. One could even argue that the “easy” condition does not really constitute a Stroop task at all, which is defined by the occurrence of response conflict on some trials – it is “just” a colour or number classification task (with redundant information on the correct response). It also cannot be assumed to activate the same task set as other such tasks, potentially introducing confounds beyond effort and boredom (e.g. “flow” experience, or response times).

We appreciate your thoughtful feedback regarding the limitations of using a 100% incongruent Stroop task and the interpretation of results from only two conditions. To address these concerns, we will carefully interpret our results and explicitly mention in our discussion the necessity of conducting further studies on this topic. Future research should consider implementing Stroop tasks with varying difficulty levels, different proportions of congruent and incongruent trials, or even alternative tasks beyond the Stroop.

2. I am also somewhat skeptical about the choice to make the hard task 100% incongruent (p7 of your response). This essentially turns your experiment into a blocked design, while the hard task was previously mixed. That's not a problem per se; since it's not uncommon for the Stroop paradigm. The fact that comparisons between congruent and incongruent blocks still yield substantial differences shows that it's a valid way of manipulating difficulty. (see, e.g.: https://www.researchgate.net/publication/242663541_Effects_of_Type_of_Design_Blocked_vs_Randomized_on_Stroop_and_Emotional_Stroop_Tasks) . But there are some reasons why this change may be more problematic to your question than your previous mixed version. For one, I see a risk of making the task less easily comparable: The perfect negative contingency between written word and print colour incentivizes task recoding: Participants now have additional reason to ignore the irrelevant feature dimension – this might also have the unintended side effect of reducing difficulty. Second, if the goal is pushing the two task conditions further apart in terms of difficulty and boredom, this might backfire: A mixture of congruent and incongruent trials would add variety, after all. It might not even be perceived as more difficult because the task demands stay the same for the entire session. Even if you do not vary congruency proportion systematically and include a 100% congruent task for the easy condition, consider if your previous mixed difficult condition (or a similar one) might not serve you better.

We understand your concerns regarding the use of 100% incongruent trials for the hard task. After carefully considering the pros and cons, we have decided to proceed with the 100% incongruent task for the following reasons:

- 1. Clarity in analysis and interpretation: As we are only analyzing the five preceding trials before each thought probe, using a 100% incongruent task helps in interpreting the results more clearly. Calculating the mean of the five trials preceding the probe results in the mixing of congruent and incongruent trials despite the substantial differences in pupil size between the trial types.**
- 2. High overall task difficulty: While reaction times in incongruent trials are longer when congruent trials are included in the Stroop task, reaction times are longer in incongruent trials than in congruent trials (irrespective of the proportion of congruent trials). Given that our analysis focuses not only on the incongruent trials of the hard task but on all five trials preceding the thought probe, our hard task will be overall more challenging with 100% incongruent trials compared to a 50% congruent and 50% incongruent version.**

From our point of view, there is no perfect solution for determining the optimal proportion of congruent and incongruent trials. However, we believe that the chosen approach is the best solution for now. We will acknowledge these limitations in our discussion and suggest further studies to explore alternative designs.

3. Regarding congruency proportion in the different tasks and your online data: I am not sure if your Pilot data (Appendix A) includes a hard color Stroop task with 100% or fewer incongruent trials – it seems to be explicit only for the numerical Stroop. If the pilot data is from the color task is for a mixed version with a 50% or at least <100% incongruent proportion, I wonder why you are diverging from a task for which you already have empirical data. If your hard color task was in fact 100% congruent, then you cannot interpret the interaction the way you do, as the numerical task had a different congruency proportion which may account for differences in results: “This interaction suggests that the impact of

difficulty level (easy, hard) on perceived task difficulty depends on the task variant (color, numerical).”, p33.

We appreciate your observation regarding the congruency proportion in our tasks and the interpretation of our pilot data in Appendix A. To clarify, the pilot data in Appendix A does indeed include a hard color Stroop task with a 50% incongruent trial proportion. We understand the advantages of maintaining consistency between our pilot data and the main study design. This is the reason why we first planned to implement a 50% incongruent Stroop task as hard variant. However, after the feedback in the previous review round, we have weighed up the pros and cons of 50% incongruent and 100% incongruent and decided to use the 100% incongruent condition (see our reasons in our response to point 2).

To clarify this point in our manuscript, we now write (p. 31):

Lastly, while the hard Stroop task in the online study consisted of 50% incongruent trials, the hard Stroop task in the present study will consist of 100% incongruent trials. This change is implemented with the aim to (1) increase the overall difficulty of the hard task and (2) enhance the clarity of our analysis, particularly regarding pupillary reactions, as pupil size is greater in incongruent than in congruent trials. Otherwise, by calculating the mean of the five trials preceding the probe, congruent and incongruent trials would get mixed up.

4. If you decide to stick with the blocked (100% congruent vs. 100% incongruent version), I have a suggestion for briefing your participants: There is a possibility that participants who perform the difficult condition first are more vigilant in the beginning, expecting incongruent trials, while those starting with the easy condition might not. You might want to clear this up in the instructions – i.e. prepare participants for the occurrence or non-occurrence of incongruent trials. If the main goal is to maximize the differences between the easy and difficult tasks, then it is not critical for them to be as comparable as possible, or for the instructions to be as similar as possible.

We appreciate your suggestion regarding the briefing of participants. We agree that preparing participants for the occurrence or non-occurrence of incongruent trials will help us maximize the differences in boredom between the easy and difficult tasks. We will incorporate this information into our instructions.

In our manuscript, we now write (p. 13):

The instructions will inform participants that in the easy condition all trials will be congruent (color and color word matching) whereas in the hard condition all trials will be incongruent (color and word not matching each other, see Figure 1 A).

Review by Erik Bijleveld, 03 Jul 2024 09:23

I much appreciate the thoughtful responses to my suggestions. I have no further comments.

Thank you for all your valuable suggestions in the first review round. Your insights have significantly enhanced the quality of our manuscript.

Review by Jonas Dora, 04 Jul 2024 16:53

My apologies for the delayed review. I believe the authors have done a good job addressing the points I raised in my previous review. I am still somewhat unsure to what extent I buy into the self-reports of 'effort due to difficulty' and 'effort due to boredom' but I believe the authors have thought about this sufficiently that we will learn something from this study and I expect the authors to discuss their eventual findings in light of some uncertainty surrounding these self-reports.

Jonas Dora, UW

Thank you very much for your feedback. We will carefully interpret our results and ensure that the discussion addresses the potential validity issues surrounding the self-reports of effort due to difficulty and effort due to boredom.

Review by anonymous reviewer 1, 31 May 2024 20:57

Review of 'How effortful is boredom? Studying Self-control Demands Through Pupillometry' Ver. 2

I remain enthusiastic about this manuscript. I think the authors have largely addressed my concerns in the revised version of the manuscript. I have two remaining points for the authors consideration.

1)The authors have claimed the cost of effort is the fact that it is associated with an aversive experience. This is a mischaracterization of the theories cited and, I think, would require a bit more elaboration (e.g., consider Kurzban, 2016 for a discussion of this point. The aversive experience of effort is generally thought to index the cost of continuing with the task at hand (for varying reasons)). I suggest the authors provide a bit more nuance here.

We appreciate your close look at the topic of effort. We have supplemented our text to incorporate other perspectives on the topic. In our revised version, we now write (p. 3f):

*Cognitive effort which can be defined as “intensity of mental [...] work that organisms apply towards some outcome” (Inzlicht et al., 2018), can be measured objectively (e.g., with pupillometry) or experienced subjectively (Bijleveld, 2018; Robinson & Morsella, 2014) which we will refer to as perceived effort². A large body of research shows that the objective and the perceived investment of effort tends to feel unpleasant and aversive (David et al., 2022; Kool & Botvinick, 2018; Wolff et al., 2021). *On the other hand, effort is considered to feel unpleasant due to the costs of continuing with the ongoing activity (Kurzban, 2016). Thus, while effort is instrumental for effective self-control, it appears to carry a momentary cost, and the prolonged exertion of effort creates cumulative costs, such as fatigue or tiredness (Ainslie, 2021; Hopstaken et al., 2015; Kurzban, 2016; Kurzban et al., 2013; Westbrook et al., 2013). In addition to efforts’ intrinsic costs (Kool et al., 2013), mobilizing effort towards one goal creates opportunity costs: When we direct our effort towards one activity (e.g., exercising to tackle backpain), we have to forego other, potentially more**

² Please note that from now on, we will use the term “effort” to refer to “cognitive effort” for the sake of readability.

rewarding activities (e.g., going for a walk with friends). In this case, the aversive sensation of effort is considered to index the costs of continuing with the ongoing activity (Kurzban, 2016). This can create added self-control demands that make pursuing one's goal and continuing with a task even more self-control demanding and by extension more effortful (Kurzban et al., 2013). In light of effort's costs (intrinsic and/or opportunity costs), people are selective about when to invest effort in the service of self-control and the sensation of effort has been conceived as an index for the momentary costs of self-control (Bieleke et al., 2023).

The question of whether participants can reliably and validly report effort due to boredom vs effort due to task difficulty is centrally important. I think it needs further consideration at both a conceptual and psychometric level. I don't think asking participants at the end of the study if they felt able to assess the difference is sufficient. Moreover, I am not persuaded that the plan to examine the BIC of different models will somehow address this concern. In my mind multiple items of effort due to boredom and multiple items of effort due to task difficulty are required as this would support an SEM measurement model, or some sort of factor analytic approach. The author's responses to the reviews suggests they are under appreciating the significance of this problem. Having said that, given the novelty and importance of the work, I would still be interested in the results even with this significant limitation.

We appreciate your feedback on the reliability and validity of self-reported effort due to boredom and effort due to task difficulty. We understand the importance of this issue. While we really like the idea of implementing additional items to improve the robustness of our measures, we were not able to identify other items that we expect to accurately capture the specific constructs of effort due to task difficulty and effort due to boredom beyond directly asking for a self-report on effort, boredom, and task difficulty separately, which is already included in our study.

We acknowledge the significance of this limitation and will clearly mention in our discussion that our items are not validated, urging caution in interpreting the results. We appreciate your interest in the results despite this limitation.