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

Note: the revised passages are highlighted in orange.

Reply to the recommender

I have now received the detailed and constructive evaluations of four experts. They all welcome the proposed study as a relevant and timely contribution of high theoretical and practical interest. I share their overall positive evaluation and believe that this submission is a promising candidate for eventual Stage 1 in-principle acceptance. However, all reviewers have also raised several critical concerns. I will not attempt to reiterate all of them, especially as the reviewers clearly elaborate specific ways in which these concerns can be addressed. I would only like to highlight a few recurrent and particularly important issues.

Thank you for your feedback and for summarizing the evaluations provided by the expert reviewers. We appreciate the thoroughness with which both you and they assessed our submission and the constructive suggestions offered for improvement. As you will see, we have clarified several parts of the manuscript, and have also adapted the design to incorporate your suggestions and those of the reviewers. To ensure everything is working properly, we collected data from two additional participants (see Appendix B). We appreciate your patience and are grateful for the comments that have improved our registered report.

As pointed out by all reviewers, the theoretical framing of the study and the definitions of its core constructs can be made more precise. Most importantly, the term "effort" appears to be used interchangeably for "investment of effort" and "perception of effort", and its operationalizations (pupillometry vs. subjective report) also seem to fit onto such a distinction. Similar concerns have been raised concerning the definition of the second key construct, boredom. I believe that the reviewers offer helpful and very specific suggestions as to how the theoretical framing can be improved.

We have carefully considered the reviewers' comments and have made significant efforts to enhance the precision and clarity of our theoretical framework. Specifically, we have expanded our literature review to include more relevant papers on effort and pupillometry to strengthen the theoretical foundation of our study. Moreover, we have refined our conceptualization of effort to distinguish between the physiological objective effort we measure with pupillometry and the perception of effort we assess with self-reports. Regarding effort, in the introduction section we now write:

As self-control is defined as the "efforts people exert to stimulate desirable responses and inhibit undesirable responses" (de Ridder et al., 2012, p. 77), self-control is by definition linked to effort. Notably, self-control is also influenced by motivational aspects which would not only facilitate the regulation of behavior towards ones goal, but also reduce the sensation of effort (Wennerhold & Friese, 2023) and consequently its costs.

Cognitive effort which can be defined as "intensity of mental […] 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 effort[1]*. 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). Thus, while effort is instrumental for effective self-control, it appears to carry a momentary cost by being unpleasant, 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).* (p.3)

[…]

The idea of selectively investing effort is consistent with influential theories like the motivational intensity theory (Brehm & Self, 1989) which states that individuals do only decide to invest effort into a task of known difficulty if the required amount of effort can be justified by the individuals' importance of success. Other theories like Ego Depletion or mental fatigue theories posit that prior engagement in demanding tasks or the exertion of self-control can result in impaired performance on subsequent tasks (Baumeister et al., 1998; Kurzban et al., 2013; Marcora et al., 2009). (p.4)

[…]

Thereby, we understand task difficulty related effort as the perception of effort that individuals feel like having to expend into the task in response to the experienced task difficulty, and boredom related effort as the perception of effort that individuals have to expend into the task in response to the experience of boredom. (p.10)

Regarding pupillometry we now write:

Moreover, pupil size measurements, as an objective physiological measure of effort show consistency and correlate with the self-reported perception of effort (e.g., Koelewijn et al., 2015; Wals & Wichary, 2023; Zénon et al., 2014). (p.8)

[…]

This decline in pupil dilation aligns with the reduced demands and required effort of cognitive tasks over time. However, pupil size has not only been found to decrease over time but also to increase (Bijleveld, 2018; Timme et al., 2022). Bijleveld (2018) reported that both the perceived feeling of effort and physiological effort, as measured by phasic peak pupil dilation, increased over time in easy and difficult trials of a cognitively demanding task. This increase aligns well with the idea of a rise in boredom over time and the proposal that staying engaged with a boring task might enhance the effort that has to be invested to complete the task (Wolff & Martarelli, 2020). Although Hopstaken et al. (2015) did not explicitly mention this observation, Figure 5 in their paper suggests a tendency of an increase in phasic peak pupil dilation in the first block of the task (which had a comparable duration to the whole task in Bijleveld, 2018) before the peak pupil dilation started to decrease over time. This could indicate a progressive increase in boredom, starting early in the task and intensifying over time and related to this, an increase in effort needed to keep a good performance while the

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

individual's willingness to perform is still present. Over time other mechanisms might become more relevant in explaining total pupil dilation, such as the individual's decision to stop investing effort into the task all together, which could result from too high levels of boredom or fatigue. This idea aligns well with motivational theories like the motivational intensity theory (Brehm & Self, 1989) which suggests that effort is only mobilized to the degree that is justified by a tasks potential reward value. Within this framework, it is conceivable that fluctuations in boredom alter how much effort should be mobilized toward the task because boredom has been theorized to reduce the value people ascribe to a boredom-inducing activity (Wolff & Martarelli, 2020). However, it should be noted that Timme et al. (2022) found the averaged pupil dilation over a period of ten minutes to first decrease before showing a tendency to increase which is opposite to the findings of other two studies described above. Although pupil dilation was calculated differently in this study, these differences demonstrate the persistent uncertainty regarding how and why the perception of effort and physiological indicators of effort change during the performance of cognitive tasks over longer periods of time. Moreover, none of these studies included the assessment of boredom. Although it is highly speculative why these differences in results emerge between studies, it highlights the importance of employing a combination of self-report and pupillometry to allow for a deeper comprehension of the temporal dynamics of task difficulty and boredom dependent cognitive effort during self-controlled behavior. (p.9ff.)

Other issues that seem particularly important relate to (a) the validity of the self-report methods for the assessment of the key constructs, (b) design and labelling of the LCT and HCT conditions, (c) concerns that the 10 min break is insufficient to minimize control carry over effects, and (d) questions about the data analytic approach (e.g., more detailed analysis of task performance, use of linear mixed models).

Thank you for highlighting these issues, and we appreciate the opportunity to address them.

(a) Regarding the validity of our self-report methods, we acknowledge the comments of the reviewer and included as suggested an overall effort measure in our study to further validate our constructs. We are open to additional validation or changes in our study design if the reviewer and recommender have any additional concerns.

(b) In response to concerns about the design and labeling of the LCT and HCT conditions, we have changed the labels to "easy" and "difficult" Stroop tasks to reflect the difference more accurately between the conditions.

(c) Regarding the adequacy of the 10-minute break, while one study suggest this is sufficient (Tyler & Burns, 2008), we agree that opting for a longer delay is more prudent to prevent any potential carry-over effects. Therefore, we have decided to modify our study design to test participants on two separate days to mitigate any potential carry-over effects.

(d) We have enhanced our data analytic approach by incorporating more linear mixed models (LLMs) and comparing models using the Bayesian Information Criterion (BIC).

Finally, there appears to be some agreement among the reviewers that the empirical literature on fatigue is relevant and can inform this study (including "ego-depletion" studies,

amongst others). I am unsure how much elaboration on fatigue is needed or helpful for this RR, but I would encourage the authors to carefully consider and comment on this issue in their response (if not in the report itself). Among the reasons why this may be important is that boredom and fatigue have been argued to overlap in phenomenology and function. Moreover, the authors include a measure labelled "energy" ("How high is your energy level right now?"), which may be construed as an inverse of fatigue (at least in subjective terms, since energy depletion may not be more than a metaphor, e.g. see

https://doi.org/10.1017/CBO9781139015394.002) . If fatigue is indeed what the authors intend to measure, then it is clear that it needs to be defined clearly.

Thank you for bringing up the importance of considering fatigue in our study. We agree that fatigue is a relevant factor that can influence participants' experiences and performance, particularly in tasks requiring effort over longer periods of time. In response to this feedback, we have expanded our study to include fatigue in our thought probes (i.e., "How much fatigue do you feel?"). In addition, we have further incorporated additional literature on fatigue. As we changed our design and decided to test participants on two separate days, we will no longer assess carry-over effects. Consequently, we will not evaluate pre-session energy levels (or pre-session fatigue) as we previously planned. In the revised manuscript, we now write:

Thus, while effort is instrumental for effective self-control, it appears to carry a momentary cost by being unpleasant, 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). (p.3)
[…]

Notably, recent research indicates that both difficult and boring tasks can contribute to fatigue (Pickering et al., 2023). Fatigue is related to a decreased perception of value in exerting effort and less willingness to continue investing effort (Dora et al., 2022; Müller & Apps, 2019). Thus, maintaining focus on a task can become more effortful over time due to fatigue resulting from both task difficulty and boredom. (p.7)
[…]

Research investigating the dynamics of pupil dilation during the performance of cognitive tasks has demonstrated a gradual reduction in phasic (stimulus-evoked) pupil diameter during the execution of both high and low cognitive-demanding tasks (Hopstaken et al., 2015; Timme et al., 2022). This decline in pupil dilation aligns with the reduced demands and required effort of cognitive tasks over time. However, pupil size has not only been found to decrease over time but also to increase (Bijleveld, 2018; Timme et al., 2022). Bijleveld (2018) reported that both the perceived feeling of effort and physiological effort, as measured by phasic peak pupil dilation, increased over time in easy and hard trials of a cognitively demanding task. This increase aligns well with the idea of a rise in boredom over time and the proposal that staying engaged with a boring task might enhance the effort that has to be invested to complete the task (Wolff & Martarelli, 2020). Although Hopstaken et al. (2015) did not explicitly mention this observation, Figure 5 in their paper suggests a tendency of an increase in phasic peak pupil dilation in the first block of the task (which had a comparable duration to the whole task in Bijleveld, 2018) before the peak pupil dilation started to decrease over time. This could indicate a progressive increase in boredom, starting early in the task and intensifying over time and related to this, an increase in effort needed to keep a

good performance while the individual's willingness to perform is still present. Over time other mechanisms might become more relevant in explaining total pupil dilation, such as the individual's decision to stop investing effort into the task all together, which could result from too high levels of boredom or fatigue. (p.9)
[…]

During the Stroop task participants will be prompted eleven times to report how bored they are ("How much boredom do you feel?"), as how difficult they are experiencing the task ("How difficult is the task?"), how much effort they are investing due to task difficulty ("Due to the difficulty of the task, how much effort do you have to invest into the task?"), due to boredom ("Due to boredom, how much effort do you have to invest into the task?") and overall ("Overall, how much effort do you have to invest into the task?"), and how much fatigue they are experiencing ("How much fatigue do you feel?"). They will indicate their answer by pressing a key between 1 (not at all) and 9 (very much) on their keyboard. (p.14)

I encourage you to address the concerns that have been raised in a revised RR and very much look forward to your reply.

We appreciate the opportunity to revise and resubmit our registered report and are committed to enhancing the quality of our manuscript.

*by **Thomas Meyer**, 29 Jan 2024 12:33*
Manuscript: ***https://osf.io/dqvm5?view_only=12191f02a5db4689b00b42bab7dbd522***
version: 1

Reply to the reviewers

Review by Julia Englert, 29 Jan 2024 07:15

This is a strong proposal on a worthwhile topic. It is of clear theoretical interest because it aims to disentangle two meaningfully different concepts, namely challenge and boredom in relation to effort, and it is practically relevant, since those two concepts reflect challenges that almost everyone is facing on a daily basis. The research paradigms are suitable to the question, and the combination of dependent measures – realtime reports on subjective experience, task performance and pupil dilation as a physiological marker seem like they should synergize well in capturing those concepts. The theoretical reasoning in the introduction is also sound but could be made more precise in a small number of places.

Thank you for your thoughtful review and positive feedback on our registered report. We appreciate your acknowledgment of the theoretical significance of our study. We noted your feedback on making the theoretical reasoning more precise and revised the introduction in accordance with your suggestions below.

However, I have a few larger, interrelated concerns about the design and the proposed computation of dependent measures. On the one hand, I believe the Stroop and Flanker paradigms are well-suited to the question of disentangling effort and boredom. The same can be said of the task switching paradigm, which here would result in a single (LCT) vs. dual (HCT) task version (wherein participants are cued as to whether to categorize the word or the colour) of the Stroop test. I believe that, in its current form, the authors would not be

making full use of the information these tasks and measures have to offer. Therefore, I strongly encourage you to refine both their design and analyses to accomplish this.

Major notes on the protocol:

1. In my opinion, an "overall" difficulty level expressed in mean RTs and error rates does not tell us as much as it should and creates unnecessary confounds. Both cognitive challenge and boredom should have specific effects on Stroop interference, switch costs, sequence effects, and speed-accuracy tradeoff. Comparing only a single task 100% congruent Stroop -i.e. a Stroop task without Stroop interference – to a dual 50% congruent task, means that switch costs cannot be separated from interference, and any interference or sequence effects can only be compared within the dual task version, because in the single task, there is no variation in trial type. Comparing only overall means and RTs between two task versions means that all of these differences get lumped together. In the absence of compelling reasons to the contrary, please consider looking at switch costs & Stroop interference specifically, rather than combining all trials in a block. These "difficult" trial conditions are why the tasks are challenging in the first place, while the "easy" trials in the HCT might even provide a brief respite from it (and thus, introduce noise).

We appreciate your insights into the importance of examining specific trial conditions such as switch costs and Stroop interference. We recognize the value in considering these aspects for a more comprehensive understanding of factors contributing to task difficulty and cognitive processing. As our aim is to investigate the dynamics of boredom, boredom related effort, task difficulty, and task difficulty related effort in two tasks with different overall difficulty and boredom levels, we prefer to maintain the 100% congruent version to ensure this Stroop task version is as boring as possible. Moreover, our calculated sample size is predicted based on the difference in boredom experienced between the two tasks proposed for this study. When we modify the tasks in a way that makes them more comparable in their potential for inducing boredom, we will need to recruit more than the initially planned 95 participants to detect a difference in boredom levels between the two tasks. However, in response to your suggestions, we have decided to modify the HCT including 100% incongruent trials instead of the previous 50% incongruent trials. This adjustment will enhance the overall difficulty of the HCT condition, providing a more challenging contrast to the LCT condition and reducing potential "noise" that could be introduced by congruent trials in the HCT. We assume that this adjustment will not significantly impact the level of boredom in the HCT task, allowing us to still base our calculation of the required number of participants on the difference in boredom between the LCT and HCT tasks we tested in our pilot study. The revised methods section reads as follows:

While in the easy condition participants will get the instruction to indicate the color of the presented color word, the task in the hard condition switches according to another stimulus presented on screen which either demands to indicate the color if a "+" is displayed (in 80 % of trials) or the word if a "x" is displayed (in 20% of trials). 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). (p.13)

2. The single (LCT) and dual (HCT) task versions of the Stroop task should be varied independently of the level of congruency proportion. In addition, I recommend using more

than just two levels of congruency proportion (0% and 50%), because that way, its relationship to the two constructs can be investigated more systematically.
Consider performing both the dual and single task versions under a larger number of different levels of congruency proportion – for instance, you could add an intermediate level of 25% incongruent trials, and perhaps even a taxing level (~75% incongruent). If this is not feasible, e.g. because it would prolong the sessions too much, consider replacing the 0% condition with one that has at least *some* incongruent trials – especially if you take up the suggestion to vary congruence proportion independently of task switching. (Since 100% agreement between word and colour would mean that the task switching cues can safely be ignored.)

Thank you for your suggestion regarding the variation of congruency proportion in our Stroop task versions. While we aim to maximize the contrast between the LCT and HCT conditions in terms of overall boredom and task difficulty, we recognize the potential benefits of including additional task versions with various levels of difficulty by altering the percentage of incongruent trials. Unfortunately, testing all participants under multiple task conditions with varying levels of incongruence would not be feasible due to testing time constraints. However, we are convinced that investigating intermediate difficulty levels would add value to this research field. In the manuscript, we added this aspect in the limitation section. We now write:

Future research will need to test whether the results generalize to other tasks, including Stroop tasks of intermediate difficulty without task-switching components and with various levels of congruency. ***(p.25)***

3. Another reason for removing the confound between single vs. dual task Stroop and % congruent is that it allows controlling for the proportion congruency effect which might otherwise muddy effects of difficulty: Stroop interference is stronger when more trials are congruent. While this effect may or may not itself be related to increased cognitive control and therefore "effort", it might undermine overall effects of difficulty because interference gets smaller as the task gets more difficult (i.e. contains more incongruent trials; see for

instance, Rothermund et al., 2022; https://doi.org/10.5334/joc.232) .

Thank you for raising this concern regarding the potential confound between single vs. dual-task Stroop conditions and the proportion of congruent trials, as well as for providing this interesting paper explaining the impact of the proportion congruency effect on the Stroop interference effect under various conditions. In response to your feedback, we modified the HCT condition to include 100% incongruent trials. We have opted to keep the task switch element in the HCT, as it aligns with our aim of increasing task difficulty and making the LCT and HCT tasks as distinct as possible in terms of both boredom and difficulty. However, we understand your concern that task switching trials might lead to a similar effect on Stoop interference as the proportion congruency effect. Since our main focus in this study is on overall boredom, task difficulty, and the associated effort with the aim to understand the experience of effort (measured with thought probes) and potentially related physiological effort (measured with pupil dilation), we prefer to maintain the task switching component. Nevertheless, we will acknowledge in our limitations section that task switching trials could have influenced performance and cognitive control, warranting further investigation in future research. We hope that our current revision

adequately addresses your concerns.
In our limitations section, we now write:

Moreover, a constraint lies in the task switching component of the hard Stroop version, which may result in varying levels of required cognitive control throughout the task. Similar designs incorporating switching components (e.g., changes in the proportion of congruent and incongruent trials) have been demonstrated to impact Stroop interference (Rothermund et al., 2022), potentially affecting not only performance, but also pupillary reaction to the trial's demands as well as other psychological states measured in this study. (p.25)

4. Speed accuracy tradeoff: Effort and boredom should differentially affect the relationship to response speed and error rates. That is, if participants are bored, they might get sloppy. We can then expect a greater proportion of early answers, which are also more often wrong (e.g. a tradeoff) – dividing response times into bins may be one way to accomplish this. (That kind of inattention should also lead to smaller interference effects). On the other hand, being challenged and exerting effort should cause both an increase in response times and errors (or at least an increase in one and no change in the other) – a general cost of difficulty on both variables.

Thank you for highlighting the relevance of speed-accuracy tradeoff when investigating performance. While our focus is on exploring the experience of effort due to task-induced boredom, we agree that investigating the influence of boredom and effort on performance could provide valuable insights. We decided to include an item to measure overall effort ("Overall, how much effort do you have to invest into the task?") which we ask participants to answer eleven times. So far, we plan to report correlations of boredom, boredom related effort, task difficulty, task difficulty related effort, overall effort, and fatigue on error rates, reaction times, and pupil size (tonic, phasic). Please feel free to provide further recommendations on this aspect of our study if you see the need for additional analysis on this topic.

5. Phasic Pupil response: As mentioned, I strongly advise using separate averages for your different trial conditions (switch vs. repetition, congruent vs incongruent). Since you already link recordings to stimulus onset, please also compute separate averages for phasic pupil dilation, as there should be larger effort-based responses for switch trials compared to repetition trials, and for incongruent trials compared to congruent trials. This way, you should get a much clearer moment-to-moment picture. (I also see a chance there is no difference in pupil dilation, or even smaller pupil dilation in HCT when only "easy" trials are compared.)

Thank you for your suggestion regarding separate averages for phasic pupil dilation. We agree that this approach would provide a clearer moment-to-moment picture of pupil responses. To address this concern, we will compute separate averages for phasic pupil dilation and include them in the appendix. As mentioned earlier, we have further decided to include only incongruent trials in the difficult Stroop task and only congruent trials in the easy Stroop task. In the revised manuscript we now write:

Phasic peak pupil dilation during the Stroop task will be reported as an average across all trials for both tasks. Additionally, it will be reported separately for the difficult task distinguishing between trials involving the color naming task and trials involving the word

naming task. Each block includes the five preceding trials before the presentation of the probe. (p.25)

For our main analysis predicting pupil size with the variables task, time, effort due to boredom, and effort due to task difficulty, we plan to include the five preceding trials before the probe. This is aimed at accurately matching perceived effort (due to boredom and task difficulty) and pupil size. We understand your point about analyzing specific trial conditions separately and to address this, we have decided to change the difficult task to comprise 100% incongruent trials. This adjustment allows us to minimize the difference between trials (and their influence on pupil dilation) while keeping the task switching component and thus the high difficulty level and lower boredom level of the task.
However, to include all five preceding trials (and not only trials without switch) is important for accurately capturing participants' experiences at the moment they are asked to report their perceived effort. Analyzing pupil size only during specific trials (e.g., trials without switch) may fail to fully capture the influence of task switch on the ratings of boredom-related effort and task difficulty-related effort, potentially leading to inconsistencies between subjective reports and pupil size data. Therefore, we prefer to include all trials in our main analysis to ensure consistency between participants' ratings and the corresponding pupil size data. Please note that the order of the trials is randomized to ensure that the sequence of five preceding trials will not be the same for each probe and each participant. This approach minimizes the potential confounding effect of task switching trials.

Minor notes on the introduction and theory:

1. I know that the term (and phenomenon) are controversial, but I found it surprising that the term "ego depletion" was not mentioned at all, given that it's probably one of the first associations that come to mind in this context. Other readers might also be curious about this, even if there are good reasons why it isn't applicable or why you prefer other terms.

Thank you for bringing up the term "ego depletion." We agree, this is an important term in the field and readers might expect it to be covered. Indeed, in our initial draft we discussed its inclusion, but opted not to include it, given the inconsistent literature on the phenomenon. However, we recognize its relevance in the context our study. Therefore, we have included it in the revised section where we discuss relevant theories. We now write:

Influential theories like ego depletion or mental fatigue theories posit that prior engagement in demanding tasks or the exertion of self-control can result in impaired performance on subsequent tasks (Baumeister et al., 1998; Kurzban et al., 2013; Marcora et al., 2009). (p.4)

2. The conceptualisation of self-control and effort and its juxtaposition with impulsiveness might also be worth expanding on, a bit, see, e.g. Hofmann et al., 2009 (

https://doi.org/10.1111/j.1745-6924.2009.01116.x9) or Wennerhold & Friese, 2022

(https://doi.org/10.1111/spc3.12726) .

Thank you for suggesting further exploration of the conceptualization of self-control and effort, particularly in relation to impulsiveness. We have incorporated this idea into our introduction, drawing from the literature you mentioned by Hofmann et al. (2009) and Wennerhold & Friese (2022). In the revised registered report, we now write:

This example illustrates that self-control is essential for reaching our goals. Indeed, self-control is understood to be a fundamental aspect of human functioning (Ainslie, 2021; Bieleke & Wolff, 2021) helping us to overcome impulses that offer short-term gratification but are not in line with our long-term goals (Hofmann et al., 2009). Better self-control is linked to a wide range of positive outcomes, such as success, health, and happiness (Hofmann et al., 2014; Moffitt et al., 2011). As self-control is defined as the "efforts people exert to stimulate desirable responses and inhibit undesirable responses" (de Ridder et al., 2012, p. 77), self-control is by definition linked to the investment of effort. Notably, self-control is also influenced by motivational aspects which would not only facilitate the regulation of behavior towards ones goal, but also reduce the sensation of effort (Wennerhold & Friese, 2023) and consequently its costs. (p.3)

a. (Note: Searching for "Hofmann" in the draft, I saw that you cited a 2014 publication by Hofmann et al. that isn't in the reference list).

Thank you for bringing this to our attention. We have added the missing reference by Hofmann et al. (2014) to our reference list and ensured that all other references are listed in the references section.

Minor suggestions for the methods:

1. For the start of each Stroop trial: One of your task cues is a plus sign, which looks a lot like the fixation cross – you should replace at least one of these with something more distinct, or you will get repetition priming and potential task preparation benefits on the colour trials!

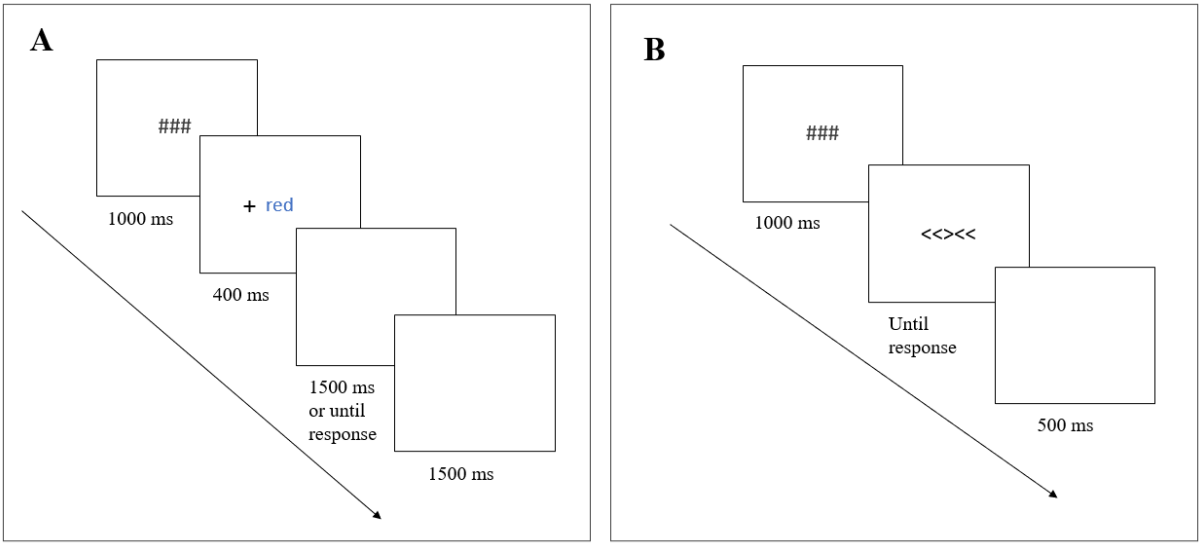
Thank you for your thoughtful suggestion. We have replaced the fixation cross with a row of fixation stimuli ### to prevent potential priming effects and to minimize task preparation benefits on the color trials. In the revised methods section we now write:

Each trial consists of the presentation of fixation stimuli (###, including pixels of all four colors of the color words to ensure a more similar luminance, displayed for 1000 ms on screen), a color word (green, red, blue, yellow) that will be presented for 400 ms either in green (RGB: 0,85,0,255), blue (RGB: 0,0,85,255), red (RGB: 85,0,0,255) or yellow (RGB: 85,85,0,255), a blank screen (displayed until reaction or for 1500 ms), and an intertrial interval (displayed for 1500 ms). (p.13)

P. 14:

Figure 1

Example of an Incongruent Trial of the Stroop Task (A) and of the Flanker Task (B)



Note. Naming the font color ("+" presented next to the word) will be the task in 80 % of the trials of the hard Stroop task (A). In 20 % of the trials a "x" will be presented and thus, the task will be to indicate the word. All trials of the hard Stroop task will be incongruent whereas all trials of the easy Stroop task will be congruent. Flanker task trials (B) will be either incongruent or congruent (50 % of all trials).

2. Consider adding a probe about regarding perceived overall difficulty of a given session to the thought probes for carry over effects

Thank you for your suggestion. Following the concerns of another reviewer, we decided to change our design and test participants on two separate days. Therefore, we will no longer asses carry over effects.

3. Supplementary measures and sample characteristics: If in accordance with ethics: it might also be worth inquiring after potential sleep deprivation, ADHD status, and current stimulant medications. (They might also be worth considering as potential exclusion criteria).

We appreciate your recommendation on sample characteristics that could be relevant in our study. We will incorporate items regarding ADHD diagnosis, stimulant medications, and sleep deprivation into our measures. We do not plan to exclude participants based on these factors, as our study employs a within-subjects design with a large sample size. However, we will include these variables when describing the characteristics of our sample. In the revised manuscript we now write:

In addition to the measures needed to investigate the core research question of this paper we will implement supplementary measures to provide a more comprehensive description of the sample and to enhance our understanding of experiencing boredom and related characteristics. For this purpose, participants will answer three questionnaires before they take part in the experiment, the Short Boredom Proneness Scale (SBPS; Struk et al., 2017); German version by (Martarelli et al., 2021), the Beck Depression Inventory II (BDI-II, Beck et al., 1996); German version by (Kühner et al., 2007), and the Brief Self-control Scale

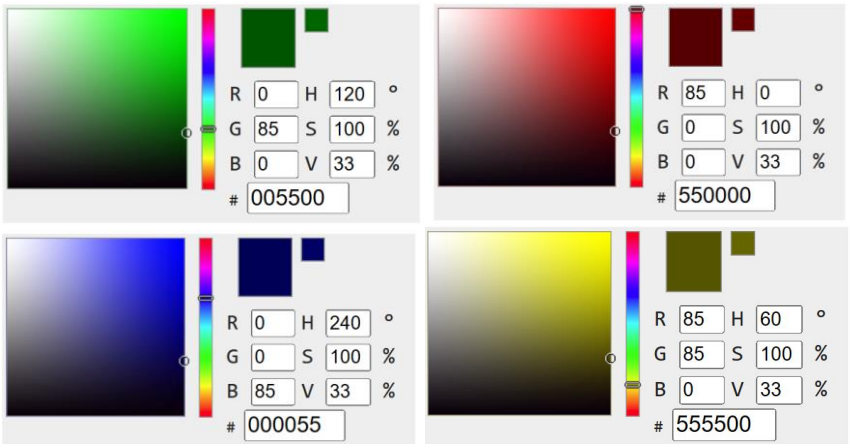
(BSCS; Tangney et al., 2018); German version by (Bertrams & Dickhäuser, 2009). They will also indicate whether they have been diagnosed with ADHD, if they have consumed any stimulants on the day of testing (caffeine, nicotine, amphetamine, others), and whether they feel sleep deprived ("Do you feel sleep deprived?" Answered with "yes" or "no"). We will include responses in the sample description. After the completion of the experiment, participants will get offered to take with them some sweets. The amount chosen by each participant will be measured.

4. Please add some information on the lighting conditions, since these have a large impact on pupil dilation. Is the laboratory cut off from external light sources, and can you ensure the level of brightness? This info will also be needed to replicate your experiment, as lighting conditions could effect pupil sensitivity.

Thank you for raising these concerns. The testing room is cut off from external light sources, and there will be no illumination in the testing room. In the revised manuscript we write:

The testing room will be cut off from external light sources, and it will not be illuminated throughout all testing sessions.

The colors selected for our stimuli are similar in darkness. Please see our colors presented below.



However, we cannot guarantee the brightness of the stimuli during the display of color words is exactly the same. Therefore, we decided to measure peak pupil dilation during the presentation of a blank screen to minimize potential issues with differences in luminance between the trials. In the revised manuscript we write:

Tonic (baseline) pupil size will be averaged separately for each last 500 ms of the fixation stimuli (###) presentation. Phasic (stimulus-evoked) peak pupil dilation for each trial will be obtained by correcting peak pupil dilation during the presentation of the blank screen (as this time period typically corresponds to the occurring of peak pupillary response in the Stroop task, see for example Hershman & Henik, 2020) for baseline pupil size (averaged pupil size during a period of 500 ms before stimulus onset). Both tonic (baseline) pupil size and phasic (stimulus-evoked) pupil size will be averaged in blocks with each block consisting of the last five trials before each probe. (p.15)

Minor questions/ suggestions for the analysis and interpretation:

1. Power analysis: Aiming for a test power of 95% is laudable and I hope it will not prove too ambitious, especially in an elaborate laboratory setting. If the authors' are confident they can reach this target, all the better. If you are not sure this is feasible, I would find it perfectly fair to specify a sample range starting at a slightly lower power (e.g. required n for 80% and 95% power).

Thank you for your feedback regarding our power analysis. We appreciate your acknowledgment of the ambitious aim for a test power of 95%. While we recognize that achieving this level of power will be challenging, we remain committed to maintaining a power of 95% for our study. To make sure that we achieve our goal of testing 95 participants, we decided to increase the compensation for participants.

Participants will give written-informed consent and be free to end the experiment at any time. They will receive 100 CHF or course credits as compensation if they are students at UniDistance Suisse. (p.12)

2. Carry over effects: How will boredom be included in subsequent analyses if you uncover significant differences? Will the alpha level be the same as for hypotheses tests?

Thank you for your question. We have decided to change our study design and test participants on two different days. Consequently, we are no longer planning to assess carry-over effects.

3. In the HCT, are you analyzing only the colour trials, or all trials? Since there is a well-documented asymmetry in reading vs colour categorisation, including all trials would make the HCT and LCT (which has no word trials) less comparable.

For this topic, we would like to reply similar as for your previous point 5.

For our main analysis predicting pupil size with the variables task, time, effort due to boredom, and effort due to task difficulty, we plan to include the five preceding trials before each probe. This way, we can match perceived effort (due to boredom and task difficulty) and pupil size as accurately as possible.

To account for differences between the trials, we decided to report separate averages for different classes of trials (see the section of our manuscript below). We further decided to change the difficult task to include 100% incongruent trials. This adjustment allows us to minimize the difference between trials (and their influence on pupil dilation) while keeping the task switching component and thus the high difficulty level of the task.

If we decide not to include all five trials before the probe to calculate the pupil dilation, but only color task trials, our prediction of pupil size with the subjective experience of participants (which includes all trials independent of their task switch type) would be less accurate. Therefore, we prefer to include all five preceding trials in our main analysis to ensure consistency between ratings and pupil size data. The

order of our trials is randomized to ensure that the sequence of five preceding trials will not be the same for each probe and each participant. This minimizes the potential confounding effect of task switching trials.

To comprehensively understand our variables, we will conduct an analysis of descriptive statistics and present them in a table. We will calculate the means and standard deviations for various measures across eleven different time points. These statistics will be computed separately for our two difficulty levels (easy, hard). Tonic pupil size, phasic peak pupil size, ratings of boredom, fatigue, task difficulty, boredom related effort, task difficulty related effort, overall effort as well as the reaction time and error rates during the Stroop task and the flanker task will be reported in eleven blocks corresponding to the thought probes. Reaction times and error rates in the flanker task will be reported separately for congruent and incongruent trials, as well as combined. Phasic peak pupil dilation during the Stroop task will be reported as an average across all trials for both tasks. Additionally, it will be reported separately for the hard task distinguishing between trials involving the color naming task and trials involving the word naming task. Each block includes the five preceding trials before the presentation of the probe. (p.22f.)

4. Please also distinguish between congruent and incongruent trials in the flanker task (see comments about methods)

Thank you for your advice regarding the distinction between congruent and incongruent trials in the flanker task. We will add reaction times and error rates for congruent and incongruent trials separately in our descriptive analysis part and include the type of flanker trials (congruent, incongruent) in our main analyses:

Difficulty Manipulation's Influence on Performance

To access whether the level of difficulty in the first task (Stroop task; easy, hard) influences participants' performance in the secondary task (flanker), we will conduct two 2 (previous task type: easy, hard) x 2 (flanker type: congruent, incongruent) ANOVAs. One ANOVA will analyze participants' error rate as dependent variable, while the other will focus on the reaction time. Results will be visualized with two bar charts (one showing the error rate, the other one showing the reaction time) providing the difficulty level on the x-axis, the error rate or the reaction time on the y-axis, and four differing bars (one for the easy Stroop version and congruent flanker trials, one for the easy Stroop version and incongruent flanker trials, one for the hard Stroop version and congruent flanker trials and one for the easy Stroop version and incongruent flanker trials). (p.23)

[…]

Tonic pupil size, phasic pupil size, ratings of boredom, task difficulty, boredom related effort and task difficulty related effort as well as the reaction time and error rates during the Stroop task and the flanker task will be reported in eleven blocks corresponding to the thought probes. Reaction times and error rates will be reported separately for congruent and

incongruent trials, as well as combined. (p. 24f.)

In any case, I hope these notes have been helpful to the authors and am looking forward to seeing this promising research progress.

Review by Erik Bijleveld, 24 Jan 2024 10:11

I read the Stage 1 report "How effortful is boredom?" by Radke and colleagues. The authors propose that boredom can act as a "self-control demand", and thus can increase effort on tasks. They aim to distinguish the proposed effect of boredom on effort from the (well-established) effect of task difficulty on effort. Their study proposes to use both subjective measures (of experienced difficulty, boredom, and a few others) and pupillometry.

I think the rationale of the paper is clear and convincing, and the general topic of this paper is timely. In general, I think this is a worthwhile study, and I would be interested in seeing the outcome of this study. My comments are intended constructively, and would still be interested in this paper if the authors would not follow all of my recommendations:

Thank you for taking the time to review our manuscript and for providing your thoughtful feedback. We are happy to hear that you found the rationale of our paper clear and convincing. We value your constructive comments, and we are committed to addressing them as best as possible to enhance the quality of our manuscript.

1) Relevant empirical papers

The idea to study the association between (high-frequency) subjective measures vs. physiological and/or behavioral measures in lab tasks has been done, to my knowledge, a few times before.

- On the topic of effort, I did two experiments where I studied the association between pupillometry and subjective feelings of effort (Bijleveld, 2018, Consciousness and Cognition). Though the designs were within-subjects, I found that effort--both subjective effort and phasic pupil dilation--increases with time. I think Figures 3A-B and 4A-B are very similar to what the authors have in mind. To be sure: (1) This previous paper does not challenge the novelty of the authors' plan; I understand their study will add important insights on boredom; and (2) This comment should not be seen as a request for citation or anything; I just think the paper is quite relevant to the authors' plan.

- On the topic of boredom as a source of effort, a recent study showed that a boring task (Mackworth Clock task) made people very fatigued, to a similar degree as an extremely demanding task (TLoadDBack task; Pickering et al., 2023, Canadian Journal of Experimental Psychology). This finding seems well in line with the authors' main argument.

- If the authors haven't seen these papers, I think it is worth looking Hopstaken et al. (e.g., 2015, Biological Psychology), Müller et al. (2019, Nature Communications), and Dora et al. (2022, JEP:G). All these studies examine how feelings of effort and/or fatigue dynamically fluctuate during task performance, and they may support part of the authors' arguments.

We greatly appreciate your recommendations on relevant empirical papers related to our research. We fully agree that this is every relevant to our plan! We included these papers in our manuscript to enhance our theoretical framework:

Thus, while effort is instrumental for effective self-control, it appears to carry a momentary cost by being unpleasant, 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). (p.4)

[…]

Notably, recent research indicates that both difficult and boring tasks can contribute to fatigue (Pickering et al., 2023). Fatigue is related to a decreased perception of value in exerting effort and less willingness to continue investing effort (Dora et al., 2022; Müller & Apps, 2019). Thus, maintaining focus on a task can become more effortful over time due to both task difficulty and boredom. (p.7)

[…]

Research investigating the dynamics of pupil dilation during the performance of cognitive tasks has demonstrated a gradual reduction in phasic (stimulus-evoked) pupil diameter during the execution of both high and low cognitive-demanding tasks (Hopstaken et al., 2015; Timme et al., 2022). This decline in pupil dilation aligns with the reduced demands and required effort of cognitive tasks over time. However, pupil size has not only been found to decrease over time but also to increase (Bijleveld, 2018; Timme et al., 2022). Bijleveld (2018) reported that both the perceived feeling of effort and physiological effort, as measured by phasic peak pupil dilation, increased over time in easy and difficult trials of a cognitively demanding task. This increase aligns well with the idea of a rise in boredom over time and the proposal that staying engaged with a boring task might enhance the effort that has to be invested to complete the task (Wolff & Martarelli, 2020). Although Hopstaken et al. (2015) did not explicitly mention this observation, Figure 5 in their paper suggests a tendency of an increase in phasic peak pupil dilation in the first block of the task (which had a comparable duration to the whole task in Bijleveld, 2018) before the peak pupil dilation started to decrease over time. This could indicate a progressive increase in boredom, starting early in the task and intensifying over time and related to this, an increase in effort needed to keep a good performance while the individual's willingness to perform is still present. Over time other mechanisms might become more relevant in explaining total pupil dilation, such as the individual's decision to stop investing effort into the task all together, which could result from too high levels of boredom or fatigue. This idea aligns well with motivational theories like the motivational intensity theory (Brehm & Self, 1989) which suggests that which suggests that effort is only mobilized to the degree that is justified by a tasks potential reward value. Within this framework, it is conceivable that fluctuations in boredom alter how much effort should be mobilized toward the task because boredom has been theorized to reduce the value people ascribe to a boredom-inducing activity (Wolff & Martarelli, 2020). However, it should be noted that Timme et al. (2022) found the averaged pupil dilation over a period of ten minutes to first decrease before showing a tendency to increase which is opposite to the findings of other two studies described above. Although pupil dilation was calculated differently in this study, these differences demonstrate the persistent uncertainty regarding how and why the perception of effort and physiological indicators of effort change during the performance of cognitive tasks over longer periods of time. Moreover, none of these studies included the assessment of boredom. Although it is highly speculative why these differences in results emerge between studies, it highlights the importance of employing a combination of self-

report and pupillometry to allow for a deeper comprehension of the temporal dynamics of task difficulty and boredom dependent cognitive effort during self-controlled behavior. (p.9ff)

2) Relevant theory

I was somewhat surprised the authors did not draw from Motivational Intensity Theory, which posits that effort emerges from the (nonlinear) interaction between (a) subjective task demands and (b) the importance of success on the task. Though the authors argue that boredom increases demands, I think one could also argue that boredom may reduce the importance of success. A good introduction is Richter et al. (2016, advances in motivation science). I could also see that this theory may help to make sense of findings post-hoc, i.e., in the discussion.

Thank you for bringing Motivational Intensity Theory and its potential relevance to our study to our attention, particularly in understanding possible findings. Indeed, it is very likely that boredom modulates potential motivation, thereby providing a neat explanation for how much effort is licensed by the task at any given time (and as a function of boredom). The non-reliance on linear dynamics is also very consistent with how boredom might fluctuate over tasks were demands constantly change (of course, this latter point does not really apply to our task here but it might be interesting to touch on in future studies). We therefore included Motivational Intensity Theory into our introduction and consider it for the interpretation of our results. Thank you very much for pointing us in this direction.

The idea of selectively investing effort is consistent with influential theories like the motivational intensity theory (Brehm & Self, 1989) which states that individuals do only decide to invest effort into a task of known difficulty if the required amount of effort can be justified by the individuals' importance of success. (p. 4)

[…]

Over time other mechanisms might become more relevant in explaining total pupil dilation, such as the individuals' decision to stop investing effort into the task all together. This idea aligns well with motivational theories like the motivational intensity theory (Brehm & Self, 1989) which suggests that which suggests that effort is only mobilized to the degree that is justified by a tasks potential reward value. Within this framework, it is conceivable that fluctuations in boredom alter how much effort should be mobilized toward the task because boredom has been theorized to reduce the value people ascribe to a boredom-inducing activity (Wolff & Martarelli, 2020). (p.9)

3) Procedure

A methodological challenge with pupillometry is that (phasic) pupil dilation is not a super fast measure; the pupil response has a delay of several 100s of milliseconds (Hoeks & Levelt, 1993; Hershman & Henik, 2000, Memory & Cognition). Another methodological challenge is that pupil dilation depends strongly on brightness of the screen. In my experience, the effect of stimulus type (e.g., a letter string vs. a fixation cross) is a lot stronger than the effect of task difficulty, which may make it hard to interpret data if brightness is not kept constant and/or if stimuli have varying different durations (e.g., if stimulus duration depends on RT, as is the case in the proposal; I would definitely try to avoid that). It is currently not clear if the authors are aware of these challenges. One way to deal with this issue would be to search

previous papers that used pupillometry and Stroop, and closely model the procedure after these previous papers (such as Hershman & Henik, 2020).

Thank you for highlighting the methodological challenges associated with pupillometry, particularly regarding the delay in pupil response and the impact of screen brightness. We appreciate your insights and have carefully considered them in refining our procedure. To address these challenges, we have made adjustments to our trial structure. While we initially presented stimuli until participants reacted to minimize waiting time between the response and the appearance of the next trial, and thus potential boredom, we recognize the importance of optimizing conditions for pupillometry. After carefully considering previous studies and evaluating various options, we decided to implement the following procedure.

Each trial will consist of a 1000 ms presentation of a fixation cross row that includes pixels of all four colors of the stimuli to ensure a more similar luminance between the fixation screen and the task screen, followed by a 400 ms presentation of the color word, a presentation of a blank screen for 1500 ms or until participants' response, and finally, a 1500 ms interstimulus interval. Peak pupil dilation will be measured during the presentation of the blank screen and baseline corrected using the 500 ms period preceding stimulus onset.

We believe these adjustments allow us to maintain the task design as close as possible to the version used for self-control research and to optimize measurement conditions for pupillometry.

Each trial consists of the presentation of fixation stimuli (### including pixels of all four colors of the color words to ensure a more similar luminance, displayed for 1000 ms on screen), a color word (green, red, blue, yellow) that will be presented for 400 ms either in green (RGB: 0,85,0,255), blue (RGB: 0,0,85,255), red (RGB: 85,0,0,255) or yellow (RGB: 85,85,0,255), a blank screen (displayed until response or for 1500 ms), and an intertrial interval (displayed for 1500 ms). (p.13)

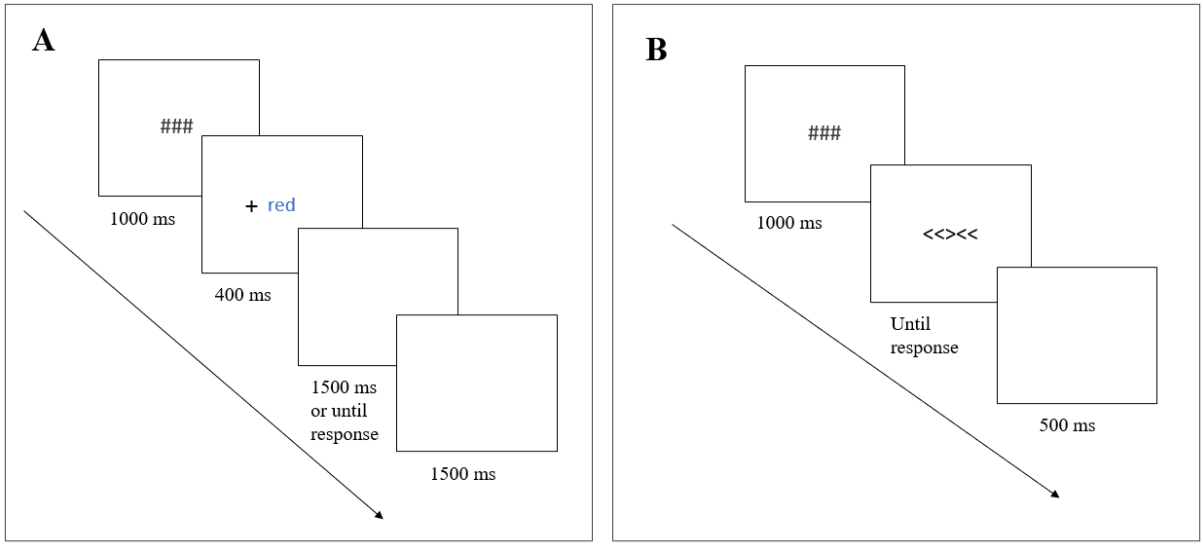
[…]

Tonic (baseline) pupil size will be averaged separately for each last 500 ms of the fixation stimuli (###) presentation. Phasic (stimulus-evoked) peak pupil dilation for each trial will be obtained by correcting peak pupil dilation during the presentation of the blank screen (as this time period typically corresponds to the occurring of peak pupillary response in the Stroop task, see for example Hershman & Henik, 2020) for baseline pupil size (averaged pupil size during a period of 500 ms before stimulus onset). Both tonic (baseline) pupil size and phasic (stimulus-evoked) pupil size will be averaged in blocks with each block consisting of the last five trials before each probe. A larger phasic (stimulus-evoked) peak pupil dilation will serve as indicator for higher stimulus-induced cognitive effort. Conversely, a lager tonic (baseline) pupil size will serve as indicator for higher cognitive effort, irrespective of the effort induced by the task stimulus. Analyzing tonic (baseline) pupil size is essential for our research question as effort related to boredom persists throughout the whole trial, i.e., during the presentation of the fixation stimuli and during the presentation of the task stimulus. Relying solely on tonic pupil size as a means of baseline correcting stimulus-evoked pupil size and not analyzing it independently would neglect the possible effect of boredom on cognitive effort and could lead to an overlook of this effect. (p. 15)

P.14:

Figure 1

Example of an Incongruent Trial of the Stroop Task (A) and of the Flanker Task (B)



Note. Naming the font color ("+" presented next to the word) will be the task in 80 % of the trials of the hard Stroop task (A). In 20 % of the trials a "x" will be presented and thus, the task will be to indicate the word. All trials of the hard Stroop task will be incongruent whereas all trials of the easy Stroop task will be congruent. Flanker task trials (B) will be either incongruent or congruent (50 % of all trials).

4) Procedure

The authors plan to do both sessions on the same day, with a 10-minute break in between. I understand that this is desirable from a practical point of view, especially since the authors plan to recruit 95 participants (which is great: this is a lot more than commonly done with this type of design). In other studies with similar designs, however, sessions are often conducted on different days (e.g., Lin et al., 2020, Psych Science), to prevent carry-over effect. As it stands, the duration of the break is justified solely based on what seems to be a rather weak finding (an ego-depletion paper from 2008, with N=10 per condition, and a rather odd manipulation). Yet, this is a crucial design choice for the present study: although a 10-minutes break may be enough, I recommend justifying it better.

Thank you for your valuable feedback regarding our procedure. We appreciate the suggestion to provide a stronger justification for the duration of the break. We have decided to err on the side of caution and will test on two separate days. While we anticipate some challenges in recruitment, we believe this adjustment will eliminate any potential carry-over effects. As such, we have decided to change the study design and test participants on two separate days.

In this within-subjects design study participants will take part in two sessions separated by at least one day, each consisting of a Stroop task and a subsequent flanker task. (p. 12)

5) Analysis

p. 17-18 'Thought probes': The authors seem to treat time as a categorical variable in their ANOVA. If so, this seems somewhat odd, given that the authors make linear predictions (e.g., "we expect an increase ... over time", p. 18). I recommend doing this in an linear mixed-effects modeling framework, in which the effect of time is included as a fixed effect. This seems more appropriate and consistent with what the authors plan to do under "Effort and pupil size".

We agree with your feedback regarding the consistency of our analyses. We have revised our analyses section and have incorporated linear mixed-effects models instead of ANOVAs for analyzing changes over time in participants self-reports:

We expect changes over time in participants' self-reports for both the easy and the hard variant of the task. Changes over time concerning those variables will be analyzed conducting six Linear Mixed Models including the perception of boredom, boredom related effort, task difficulty, task difficulty related effort, overall effort, and fatigue as outcome variables, task (easy, difficult) and time (probe one to eleven) as fixed effects accounting for differences among participants by including random intercepts and different effects of time on the outcome variables by including random slopes for time. (p.20f.)

6) Analysis

"Effort and pupil size": The authors explain that they will use "random effects among participants allowing for random intercepts". Could they clarify this? I can see the need for a random intercept: this tells the model people may differ, overall, e.g., in their phasic pupil dilation. However, I can also see the need for random slopes: these would tell the model that people may differ in, e.g., how strongly phasic pupil dilation depends on time. In the literature, there is some discussion as to how to navigate random slopes in experimental designs. It may be helpful to look at the (very influential) guidelines by Barr et al. (2013, J. Memory and Language) and the (very reasonable) response by Bates et al. (https://arxiv.org/abs/1506.04967). In any case, I think it would be good to decide a priori what the random-effects structure of their models will look like.

Thank you for your suggestions regarding the random effects structure in our analysis. We agree that participants might not only differ overall in their phasic pupil dilation, but also over time. We have revised our models and integrated random slopes:

Our first model will predict pupil size with the task's difficulty level (easy, hard), time (thought probe one to eleven), and cognitive effort due to task difficulty as fixed predictors, accounting for general differences between participants in tonic and phasic pupil size by including random intercepts and different effects of time on pupil size among participants by including random slopes for the variable time. (p.20)

Review by Jonas Dora, 04 Jan 2024 18:17

Download the review *(downloaded and added below)*

Review Radtke et al. PCI RR

Thank you for giving me the opportunity to review this Stage 1 registered report proposal. To start off this review, I would like to point out that I am not an expert on the details of processing and analyzing pupillometry data, so I hope that one of the other reviewers was able to review that aspect of the proposal. The authors attempt to clarify the role boredom plays in effortful self-control. This is a very interesting question given that many cognitive tasks we use to study effort and self-control are highly repetitive and thus it is plausible that boredom might be a confounder. The thing I am not convinced by (yet) is whether this study will be able to disentangle effort and boredom, partially because I did not fully understand what the authors consider effort and boredom to be. I list my comments for the authors to consider below.

Thank you for dedicating your time to reviewing our proposal, and for your valuable insights. We share the opinion that effort and boredom are constructs that are difficult to disentangle. We have taken note of your feedback regarding the clarity of our definitions for effort and boredom, and we tried to provide a more detailed and explicit explanation of our understanding of effort in our revised manuscript. We are happy to include further clarification of our understanding of effort and boredom in our manuscript if you feel that more clarification is needed despite the changes we already incorporated.

One thing that does not become clear to me is how the authors think conceptually about effort and boredom. On p. 3, the authors talk about effort as both a behavior (investing effort or working hard) and a subjective experience (feeling of effort during mental activity). Later on, we see that they plan to measure effort both via self-report and pupillometry. I am guessing that these two operationalizations fit onto the two different efforts (self-report for subjective experience and pupillometry for investment of effort) but this is not explicitly explained or justified. How does the investment of effort relate to the experience of effort, and how do both relate to boredom? On p. 5, the authors talk about boredom mainly as a subjective experience (being bored). Much of what they write about the experience of boredom on p.5 also applies to the experience of effort (reflect the cost of ongoing actions and to signal that maybe we should do something else). Thus, while effort and boredom feel different and often occur under slightly different circumstances (mostly people experience effort when task difficulty is high and experience boredom when task difficulty is low, though there are exceptions as the authors note), they are though to have a similar purpose. The authors then propose that experiencing boredom can be effortful. This can make sense when it comes to how much effort needs to be invested, though I think we would need to think hard about how to gain confidence in the idea that it is the perception of boredom that leads to the higher effort investment and not something else that co-occurs with the rise of boredom. I

don't think it makes sense to conceptualize that experiencing boredom can be effortful in the subjective sense as these are both experiences that are thought to track a similar thing, so I would not know how to disentangle them. In summary, I think it is important that the authors clarify how they think conceptually about effort and boredom and are clear in what ways they might relate to each other. I believe this paper is a great example for clearly distinguishing the investment of effort from the feeling of effort: https://doi.org/10.1016/j.concog.2018.05.013

We agree that it is crucial to clarify how we conceptualize effort and boredom in our study. Effort, as we see it, can be both invested and experienced. Thus, we distinguish between perceived effort, which pertains to the individual's perception or feeling of effort, and objective effort, which refers to the measurable intensity of mental work applied towards an outcome. In our study, we plan to assess perceived effort through thought probes and objective effort using pupillometry. We have incorporated these definitions into our manuscript to provide clarity on our conceptual framework. Additionally, we have highlighted the relationship between pupil size measurements as an objective physiological indicator of effort and the perceived self-reported experience of effort, as demonstrated in previous studies.

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 effort. (p.3)
[…]
Moreover, pupil size measurements, as an objective physiological measure of effort, show consistency and correlate with the self-reported perception of effort (e.g., Koelewijn et al., 2015; Wals & Wichary, 2023; Zénon et al., 2014). (p.8)
[…]
In the present study we will investigate as how effortful task-induced boredom is subjectively perceived and physiologically measurable. To achieve this, we will access the temporal dynamics of task difficulty and boredom-related effort during the performance of an easy and hard version of a cognitive task (Stroop task). We integrate a promising research protocol that complements the subjective assessment of perceived effort with the implementation of pupillometry (tonic and phasic pupil size) as objective measure of effort, resulting in a high temporal resolution of physiological and subjective data. (p.10)

In turn, this is how we define boredom:
Boredom occurs when one's resources feel not adequately utilized (Wolff et al., 2024). Put simply, we get bored when we feel we are wasting our time. Consistent with this conceptualization, boredom tends to occur when tasks are too easy (or too hard; Westgate & Wilson, 2018), feel meaningless (van Tilburg & Igou, 2017), and/or when a person feels they have no agency (Danckert & Eastwood, 2020). (p.7)

In our understanding, effort and boredom might both signal that we should do something else, but they are still different constructs. Investing effort is directly linked to costs, while boredom is linked to reduced value in the first place. By reducing the

value of an ongoing activity and biasing behavior towards alternative activities, boredom enhances the opportunity costs of sticking with the ongoing activity. Thus, continuing with the ongoing activity despite feeling bored leads to increased costs, resulting in a higher amount of effort that is required to complete the task. Furthermore, effort can be experienced without experiencing boredom at the same time.

Here, one primary interest lies in the effort demands that boredom might create. But clearly, effort is not only required when we are bored and have to stay focused. We fully agree, the feeling of effort is therefore also influenced by a plethora of other factors For example, fatigue, motivation, frustration, or other constructs. To address this, we plan to directly ask participants to indicate the effort they invest as a consequence of boredom and analyze whether pupil dilation, as a physiological marker of effort, can be explained by this subjective experience. While we recognize the influence of other variables on effort, our primary focus remains on understanding the effort required to endure boredom in both easy and difficult tasks that are often used in the context of self-control research. We will further ensure to discuss and account for other possibly relevant variables in our discussion when interpreting results.

1. Thought probes: Related to my first point, I am concerned about the self-reports regarding effort due to task difficulty and effort due to boredom. The authors need to clarify how effort could result from boredom. Also, it is important to ensure that these items have validity. Do people have insight into effort that arises from task difficulty and effort that arises from boredom? Do we know how participants answer these items and that they insight into this process so that they can self-report on it? Have previous studies used these items and validated them in some form?

   Thank you for raising important points regarding the validity of the thought probes. We understand your concern about ensuring the validity of self-reported boredom-related effort and task difficulty-related effort. In an ongoing study focused on heart rate and skin conductance, similar measures have been implemented. To address your point, we have included items at the end of this study to assess if participants felt able to assess the amount of effort due to boredom and effort due to task difficulty they experienced during the task. We compared how they felt they could assess this with the self-reported ability to assess other states that occur during the study. The questionnaire we included after completing the tasks, contains six items which were by now answered by 43 participants on a four-point Likert scale (1 = completely disagree, 2 = somewhat disagree, 3 = somewhat agree, 4 = completely agree). Please find the items listed with their mean, standard deviation, and mode below.

1. I was able to assess how much boredom I perceived.

(M = 2.98, SD = 0.64, $Mode$ = 3)

2. I was able to assess how difficult the task was for me.

(M = 3.44, SD = 0.59, $Mode$ = 4)

3. I was able to assess how much effort I had to invest due to boredom.

(M = 2.79, SD = 0.67, $Mode$ = 3)

4. I was able to assess how much effort I had to invest due to the difficulty of the task.

(M = 3.21, SD = 0.71, $Mode$ = 3)

5. Due to boredom, I had to invest effort into the task.

(M = 3.05, SD = 0.84, $Mode$ = 3)

6. Due to the difficulty of the task, I had to invest effort into the task.

(M = 2.30, SD = 0.94, $Mode$ = 3)

These preliminary results indicate that participants generally felt able to assess how much effort they had to invest due to these variables. They further agreed that they had invested effort into the task due to boredom. Therefore, we have decided to retain these variables, as they are core to our hypotheses. Thanks again for highlighting the potential issue that participants might struggle with these ratings, as this has allowed us to incorporate these additional questions into our other study to take this possibility into account.

2. Analyses: Higher perceived task difficulty in HCT, higher perceived boredom in LCT, lower performance in HCT: these are all clear and basic effects to expect. Thought probes: As I mention above, I am not sure what boredom-related effort and task difficulty-related effort are and whether people are able to report on these processes, so I am not sure what this analysis can tell us. Pupil size: Again, I think the authors need to present some evidence that we can have confidence in the self-reports of boredom-related and task-difficulty related effort. Also, the authors cannot rely on explained variance to compare the models. Adding more variables to a model will always increase the explained variance, but it also increases the likelihood of overfitting the data. For that reason, model comparison needs to penalize complexity in some form. This could be done by comparing models with an information criterion (AIC/BIC/WAIC) or via cross-validation.

   Thank you for your feedback regarding the analyses and methods of model

comparison. We hope we have been able to address your concern about our variables in our response to point one.
Concerning your last point, we agree that relying solely on explained variance does not adequately account for model complexity and the risk of overfitting. Therefore, we have decided to report AIC, BIC and WAIC for model comparison in our analyses and base our decisions on the BIC. The BIC penalizes model complexity more heavily than the AIC, providing a more conservative approach to model selection. Here is the revised text:

To evaluate the relative performance of these models in predicting pupil size, the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and WAIC (Widely Applicable Information Criterion) of the models will be reported. Our decision regarding the optimal balance between model fit and complexity will be based on the BIC. ***(p.22)***

3. Interpretation given different outcomes: This makes sense but rests on the assumptions that a) we can disentangle boredom-related and task difficulty-related effort and b) participants are able to report on this. These assumptions are crucial and I believe need to be explored to ensure that this very interesting project will succeed in telling us something new about boredom during effortful self-control.

We agree that disentangling boredom related effort and task difficulty related effort is essential for the success of our study. We have made efforts to address this by conducting further analyses in another study to explore if participants feel able to report on their experience of boredom related effort and task difficulty related effort. We show the results of our analyses in our reply to point 2. We further decided to add one item on the experience of overall effort ("Overall, how much effort do you have to invest into the task?") in our study and will show correlations with boredom related effort and task difficulty related effort.

I hope my review is helpful and good luck moving forward.

Jonas Dora, University of Washington

Review by anonymous reviewer 1, 08 Jan 2024 19:34

1A. The scientific validity of the research questions.

The proposed research questions are scientifically justifiable.

However, I have some concerns about the link between 'theory/conceptualization' and 'operationalization / measurement' of key variables.

1) The concept of 'effort' is not 1) precisely and explicitly defined,
2) sufficiently related to existing scholarship / definitions of 'effort',
3) measured in a manner that has unequivocal validity.

We have taken your feedback into careful consideration and have now integrated a comprehensive definition of effort into our manuscript. By doing so, we aim to enhance the clarity and alignment of our conceptual framework with existing definitions on effort.

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 effort. (p.3)

To strengthen validity of measurement, we have decided to add a thought probe regarding the overall perception of effort. In the revised manuscript we now write:

During the Stroop task participants will be prompted eleven times to report how bored they are ("How much boredom do you feel?"), as how difficult they are experiencing the task ("How difficult is the task?"), how much effort they are investing due to task difficulty ("Due to the difficulty of the task, how much effort do you have to invest into the task?"), due to boredom ("Due to boredom, how much effort do you have to invest into the task?") and overall ("Overall, how much effort do you have to invest into the task?"), and how much fatigue they are experiencing ("How much fatigue do you feel?"). They will indicate their answer by pressing a key between 1 (not at all) and 9 (very much) on their keyboard. (p.14f.)

Moreover, at times the authors use the term 'cognitive effort' and at other times the term 'effort' is used (I suspect the authors do not intend to refer to distinct concepts with these two terms and thus I would recommend using a singular term). I note the authors provide a helpful definition of 'self-control' and I suggest doing the same for 'effort'. When it comes to defining effort, it is my understanding that this term has been used in a variety of different ways by different authors and I think it would be useful to link the authors definition to existing scholarship to show how their work fits with the work of others. I will not articulate

different definitions here beyond pointing out the work of Mulder 1986 because I think Mulder's distinction between 'processing complexity' and 'compensatory control' is particularly apt in the present context (of course the authors are free to disagree with me about the applicability of Mulder's distinctions).

Thank you for drawing our attention to this issue. Additionally, we have provided a clear definition of effort, drawing from Inzlicht et al. (2018), Robinson and Morsella (2014), and Bijleveld et al. (2018) to integrate our understanding of effort and align it with existing literature. We have further integrated a footnote clarifying that we refer to cognitive effort when writing about effort in general, for the sake of readability.

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 effort[1]. (p.3)

[1] *Please note that from now on, we will use the term "effort" to refer to "cognitive effort" for the sake of readability.*

I am enthusiastic about the conceptual distinctions between a) task difficulty, b) boredom, c) boredom related effort, d) task difficulty related effort. However, it would be helpful for the authors to be clearer on how these concepts are being defined and, importantly, it would be useful to show that these concepts can be validly and distinctly reported by participants (and therefore measured by researchers). For example, can participants distinguish between the effort they have to exert because of boredom vs task difficulty? Do participants consider the 'boringness' of a task when they rate task difficulty? Or, to put it differently, will participants interpret these questions in the same way as the researchers intend them to be interpreted. I wonder if there is some creative way to examine this with single item probes. Perhaps showing autoregressive effects over time within concepts vs lagged correlations across concepts would convince a reader of their psychometric distinction? Or perhaps there is a better way of addressing these issues…At the end of the day, I think the proposed work would be publishable without an explicit demonstration of the validity of the probe items, so a note on this point in the limitations section would likely be sufficient. I offer this point in the spirit of trying to make the findings stronger.

Thank you for highlighting the important aspects concerning the validity of our

thought probes. We appreciate your comment and share the same concerns. It is worth noting that we have already used the same probes in an ongoing study. Additionally in this ongoing study, we have included post-study questions aimed at assessing participants' perceptions of boredom and task difficulty as effortful, along with their ability to estimate the effort levels attributed to boredom and task difficulty during the task. Findings from this study suggest that most participants felt able to assess how much effort they had to invest due to these variables (see results below). They further agreed that they had invested effort into the task due to boredom.

The questionnaire we included after completing the tasks, included six items which were answered on a four-point Likert scale (1 = completely disagree, 2 = somewhat disagree, 3 = somewhat agree, 4 = completely agree). It was answered by 43 participants. Please find the items listed along with their mean, standard deviation, and mode below.

1. I was able to assess how much boredom I perceived.
(M = 2.98, SD = 0.64, $Mode$ = 3)
2. I was able to assess how difficult the task was for me.
(M = 3.44, SD = 0.59, $Mode$ = 4)
3. I was able to assess how much effort I had to invest due to boredom.
(M = 2.79, SD = 0.67, $Mode$ = 3)
4. I was able to assess how much effort I had to invest due to the difficulty of the task.
(M = 3.21, SD = 0.71, $Mode$ = 3)
5. Due to boredom, I had to invest effort into the task.
(M = 3.05, SD = 0.84, $Mode$ = 3)
6. Due to the difficulty of the task, I had to invest effort into the task.
(M = 2.30, SD = 0.94, $Mode$ = 3)

However, we do agree that the validity of our items remains to be determined. Despite these limitations, the present study aims to contribute insights into the validity of these items, by combining subjective and objective methods. Following your concern (which aligns with ours) we will include a point addressing this in the limitations section of the registered report:

It is important to acknowledge limitations regarding the measurement of certain constructs in this study. Specifically, the validity of items used to assess perceptions of task difficulty, boredom and related effort need further validation. *(p.25)*

We further acknowledge the possibility that participants may consider boredom when rating task difficulty. This would lead to the outcome that effort invested due to task difficulty includes the effort invested due to boredom. If this is true, we will see this in our analyses as we first include effort due to task difficulty into our first model and add effort due to boredom in our third model. If the BIC is lower in our first model, this would indicate that balance between model fit and complexity is better in the first model. We would discuss this in our discussion section.

We further added more information to our manuscript about how we understand our key concepts.

While perceived task difficulty and task-induced boredom are likely to dynamically vary in hard and easy tasks, it is likely that these dynamics are not identical. More specifically, hard and easy tasks likely differ in their perceived difficulty and boringness from the start, and the temporal dynamics of perceived difficulty and boringness are likely to follow different trajectories too (Bieleke et al., 2021; Wolff & Martarelli, 2020). As a result, it is possible for both easy and hard tasks to exhibit various levels of effort throughout their execution. By exclusively assessing perceived effort in relation to task difficulty (as is traditionally done in self-control research), the additional effort requirements that can arise from task-induced boredom might be overlooked. Furthermore, assessing perceived effort only after task completion neglects the temporal fluctuations in effort demands during a self-control-demanding task. (p.7)

[…]

Subjective experiences will be assessed several times during the experiment with thought probes asking for the perception of boredom, task difficulty, task difficulty related cognitive effort, boredom related cognitive effort, overall cognitive effort, and fatigue. Thereby, we understand task difficulty related effort as the perception of effort that individuals feel like having to expend into the task in response to the experienced task difficulty, and boredom related effort as the perception of effort that individuals have to expend into the task in response to the experience of boredom. (p.10)

Finally, regarding the concept of 'effort', I suggest the authors are explicit about using pupil size as the gold-standard for operationalizing effort and provide some background on the existing work correlating pupil size and self-reported effort.

Thank you for your feedback on our section of effort and pupil size. We have tried to make significant improvements to this section mentioning relevant literature on our research topic as well as research linking pupil size to the subjective experience of effort. In the revised version we now write:

P. 7ff:

To uncouple the temporal dynamics of effort (due to task difficulty and boredom) and boredom during different self-control tasks, triangulation of methods holds promise. First, self-reports can provide a valuable reading of people's state (Cooper-Martin, 1994; Johnson et al., 1995) and researchers have highlighted the need to track the dynamics of people's feelings with higher resolution (Mills & Christoff, 2018; Waugh et al., 2015). Complementing self-reports, pupil size can serve as a physiological indicator of effort with, generally speaking, greater pupil size indicating increased effort: Phasic (stimulus-evoked) changes in pupil diameter have been found during the completion of various tasks that require effort with a greater pupil size being related to a greater extend of effort (e.g., in inhibition, updating, working memory tasks; van der Wel & van Steenbergen, 2018). Moreover, pupil size measurements, as an objective physiological measure of effort, show consistency and correlate with the self-reported perception of effort (e.g., Koelewijn et al., 2015; Wals & Wichary, 2023; Zénon et al., 2014).

While phasic (stimulus-evoked) changes in pupil size tend to occur as reactions to the immediate demands of a task, tonic (baseline) pupil changes tend to reflect the state of the individual more generally (Cohen Hoffing et al., 2020). Considering changes of pupil size from a general psychophysiological perspective, the activation of the sympathetic pathway of the autonomic nervous system leads to the dilation of the pupil whereas the activation of the parasympathetic pathway induces its constriction (Kardon, 2005; Mathôt, 2018; McDougal & Gamlin, 2008). While our understanding of the exact neurological processes of pupil dilation during the exertion of effort is somewhat restricted (Mathôt, 2018), we do know that the locus coeruleus (LC), a brain area suggested to play an important role in behavioral regulation and, consequently, self-control (Aston-Jones & Cohen, 2005), affects changes in pupil size. Research indicates that when the LC is more active, the pupils tend to enlarge (Joshi et al., 2016). Connections from the orbitofrontal cortex (OFC) and the anterior cingulate cortex (ACC) to the LC were found (Aston-Jones et al., 2002; Rajkowski et al.,

2000). As those brain regions are linked to the evaluation of rewards and costs, a responsiveness of the LC to ongoing cost-reward evaluations is suggested, consequently shaping the resulting behavior (Aston-Jones & Cohen, 2005). Given the association of the LC activation to cost-reward evaluations and self-control, the relation to pupil size, and the link between greater effort and larger pupils, these findings imply that the LC likely contributes to the dilation of pupils when self-control and effort (either difficulty- or boredom-related) are exerted during the engagement in cognitive tasks. Research investigating the dynamics of pupil dilation during the performance of cognitive tasks has demonstrated a gradual reduction in phasic (stimulus-evoked) pupil diameter during the execution of both high and low cognitive-demanding tasks (Hopstaken et al., 2015; Timme et al., 2022). This decline in pupil dilation aligns with the reduced demands and required effort of cognitive tasks over time. However, pupil size has not only been found to decrease over time but also to increase (Bijleveld, 2018; Timme et al., 2022). Bijleveld (2018) reported that both the perceived feeling of effort and physiological effort, as measured by phasic peak pupil dilation, increased over time in easy and hard trials of a cognitively demanding task. This increase aligns well with the idea of a rise in boredom over time and the proposal that staying engaged with a boring task might enhance the effort that has to be invested to complete the task (Wolff & Martarelli, 2020). Although Hopstaken et al. (2015) did not explicitly mention this observation, Figure 5 in their paper suggests a tendency of an increase in phasic peak pupil dilation in the first block of the task (which had a comparable duration to the whole task in Bijleveld, 2018) before the peak pupil dilation started to decrease over time. This could indicate a progressive increase in boredom, starting early in the task and intensifying over time and related to this, an increase in effort needed to keep a good performance while the individual's willingness to perform is still present. Over time other mechanisms might become more relevant in explaining total pupil dilation, such as the individual's decision to stop investing effort into the task all together, which could result from too high levels of boredom or fatigue. This idea aligns well with motivational theories like the motivational intensity theory (Brehm & Self, 1989) which suggests that which suggests that effort is only mobilized to the degree that is justified by a tasks potential reward value. Within this framework, it is conceivable that fluctuations in boredom alter how much effort should be mobilized toward the task because boredom has been theorized to reduce the value people ascribe to a boredom-inducing activity (Wolff & Martarelli, 2020). However, it should be noted that Timme et al. (2022) found the averaged pupil dilation over a period of ten minutes to first decrease before showing a tendency to increase which is opposite to the findings of other two

studies described above. Although pupil dilation was calculated differently in this study, these differences demonstrate the persistent uncertainty regarding how and why the perception of effort and physiological indicators of effort change during the performance of cognitive tasks over longer periods of time. Moreover, none of these studies included the assessment of boredom. Although it is highly speculative why these differences in results emerge between studies, it highlights the importance of employing a combination of self-report and pupillometry to allow for a deeper comprehension of the temporal dynamics of task difficulty and boredom dependent effort during self-controlled behavior.

2) I am concerned about the coherence of the concept 'level of (self) control'. First, sometimes the term 'control' is used and at other times the term 'self-control' is used, again, I encourage the authors to be precise and use one term for one concept or articulate the conceptual difference between control and self-control.

Thank you for your feedback regarding the coherence of our terminology. We agree that using one term consistently throughout the paper will enhance clarity and coherence. Therefore, we have revised the term "control" to "self-control" to ensure uniformity in our terminology.

Second, throughout, the authors refer to two levels of the Stroop Task as 'High Control' and 'Low Control' and operationalize these two conditions in terms of Stroop task parameters (i.e. a difficult and an easy version of the Stroop (difficult and easy are appropriately further specified). However, the introduction, and the logic of their hypothesis, suggest that self-control is necessary to persist with easy=boring tasks. Thus, I recommend that the authors more clearly and coherently define or label their two Stroop conditions (using the more concrete 'easy' vs 'difficult' seems like a promising way to go).

Thank you for your thoughtful suggestion regarding the terminology used to describe the two levels of the Stroop Task. After considering your feedback, we agree that using more concrete terms such as "easy" and "hard" will enhance clarity. Therefore, we have revised the terminology throughout the manuscript to refer to the Stroop conditions as "easy" and "hard" instead of "high control" and "low control."

Again, as I mentioned above a more sophisticated discussion and definition of 'effort' along the lines of Mulder's work might help when considers the best way to conceptualize the two versions of the Stroop. This issue of operationalizing level of (self) control in terms of Stroop task parameters is particularly concerning when it comes to the proposed analyses for the second task. That is, if the probe results of task 1 demonstrate that boredom is effortful then the proposed analyses for the second task may need to be altered.

We agree that if the probe results demonstrate that boredom is indeed effortful, it may necessitate adjustments to our planned analyses for the second task. To address this concern, we have incorporated more sophisticated analyses for the second task: P.21f

Task Performance in Secondary Task

For assessing task performance in the secondary task, we will focus on the error rate and reaction time separately. We aim to analyze if participants' performance in the secondary task (flanker) is influenced by the difficulty level of the preceding task (Stroop task; easy, hard). We further intend to test whether perceived task difficulty related cognitive effort exerted in the Stroop task predicts task performance in the flanker task, and whether boredom and task difficulty related cognitive effort together predict task performance more accurately.

Difficulty Manipulation's Influence on Performance

To access whether the level of difficulty in the first task (Stroop task; easy, hard) influences participants' performance in the secondary task (flanker), we will conduct two 2 (previous task type: easy, hard) x 2 (flanker type: congruent, incongruent) ANOVAs. One ANOVA will analyze participants' error rate as dependent variable, while the other will focus on the reaction time. Results will be visualized with two bar charts (one showing the error rate, the other one showing the reaction time) providing the difficulty level on the x-axis, the error rate or the reaction time on the y-axis, and four differing bars (one for the easy Stroop version and congruent flanker trials, one for the easy Stroop version and incongruent flanker trials, one for the hard Stroop version and congruent flanker trials and one for the easy Stroop version and incongruent flanker trials).

Perceived Effort's Influence on Performance

We will explore the extent to which the perceived effort, resulting from perceived task difficulty in the Stroop task, can account for variations in error rates and reaction time during the flanker task. Additionally, we will investigate whether perceived effort due to task difficulty in the Stroop task predicts performance (error rate, reaction time) in the flanker task, and whether considering perceived boredom related and task difficulty related effort in the Stroop task together, provides a more accurate prediction of the error rate and reaction time in the flanker task than task difficulty related effort alone. To investigate this, we will employ several Linear Mixed Models (LMMs), using either the error rates or the reaction time in the flanker task as the outcome variables. Perceived effort due to boredom and effort due to task difficulty will be obtained by calculating the mean of this variables across the probes. Three distinct models will be tested for each performance variable (error rate, reaction time). The first model will assess the impact of the difficulty level (easy, hard; fixed effect) and perceived effort due to task difficulty (fixed effect) in the Stroop task as well as the type of flanker trial (congruent, incongruent) on the reaction time in the flanker task while accounting for random effects among participants with random intercepts and random slopes. The second model will differ from the first by incorporating boredom related effort as a fixed effect instead of task difficulty related effort. The third model will include both perceived task difficulty and boredom related effort as fixed effects. AIC, BIC, and WAIC will be reported. The model with the lowest BIC will be considered as the most optimal one.

Two plots for each performance variable (error rate, reaction time) will illustrate the relationship between performance in the flanker task on the y-axis and perceived effort scores (effort due to boredom and effort due to task difficulty) on the x-axis. Two distinct lines will represent the difficulty levels of the task (easy, hard).

1B. The logic, rationale, and plausibility of the proposed hypotheses.

The proposed logic and rationale are strong. The proposed hypotheses are plausible.

1C. **The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis).**

The proposed methodology and analysis pipeline is sound and feasible.

1D. **Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.**

Sufficient and clear methodological detail is presented, which would allow for a close replication of the proposed work.

1E. **Whether the authors have considered sufficient outcome-neutral conditions for ensuring that the obtained results are able to test the stated hypotheses.**

Yes, the authors have sufficiently considered outcome -neutral conditions – carry over and manipulation checks for example. However, I don't recall seeing an explicit statement about whether or not the probe data would be analyzed if the hypothesized 'task difficulty', 'boredom', and 'performance' hypotheses are not confirmed (p. 17). I dont have an opinion on what is best to do in this regard...so I leave it to the authors.

We appreciate your suggestion and agree that it is important to consider how to handle the analysis of probe data in such a scenario. If task difficulty, boredom and performance will not differ between our easy and difficult task, we would address this in the discussion. However, we would conduct our main analyses regardless, as these analyses would still provide valuable insights into the dynamics of boredom and effort during task performance. We included this information into our study-design template. It's worth noting that this scenario is unlikely, as we previously used these tasks in our online study to assess whether our tasks differ regarding these variables.

P.39:

Research Question	**Hypothesis**	**Sampling Plan**	**Analysis Plan**	**Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis**	**Interpretation given different outcomes**
Manipulation Check: Do hard and easy Stroop task versions differ in terms of their perceived overall task difficulty, the overall level of boredom they induce, and in terms of how	Participants report greater overall levels of boredom and less overall task difficulty and show a better performance in the easy compared to the hard.	Ninety-five non-color-blind participants between 16 and 45 years from the general population	To test this hypothesis, we will perform four paired t-tests comparing either task difficulty, boredom, error rates or reaction times (as dependent variables) in the easy and		Differences: Stroop Tasks classified as difficult (and high self-control demanding tasks) are perceived as more difficult, they induce less boredom and lead to a worse performance in comparison to easy Stroop Tasks

good someone performs in the tasks?		will be recruited[2].	hard Stroop version (as independent variable).		(often classified as low self-control tasks). No differences: The manipulation of the difficulty level of the Stroop Tasks does not impact perceived difficulty levels or performance. This would indicate that these variables are not affected by the type of Stroop task that is used. **We will conduct our main analyses even if no differences between the tasks emerge, as we would still gain valuable insights into the dynamics of boredom and cognitive effort during task performance.**

[2] To determine the required sample size for the present study, we conducted a G*Power Analysis Faul et al. (2007). Given that the effect size observed in the calculated t-tests in our online study was larger for task difficulty (d = 0.92) than for boredom (d = 0.34), our power analysis focuses on the effect size of the difference in boredom between the easy and the hard color Stroop. The power analysis for a one-tailed paired t-test was calculated based on this effect size. The analysis indicated that 95 participants are necessary to detect a difference in boredom between the tasks with a power of 95% at an alpha level of 0.05.

References

Ainslie, G. (2021). Willpower with and without effort. *Behavioral and Brain Sciences*, *44*, e30. https://doi.org/10.1017/S0140525X20000357

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Aston-Jones, G., Rajkowski, J., Lu, W., Zhu, Y., Cohen, J. D., & Morecraft, R. J. (2002). Prominent projections from the orbital prefrontal cortex to the locus coeruleus in monkey. *Society for Neuroscience Abstract*, *28*, 86–89.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265. https://doi.org/10.1037/0022-3514.74.5.1252

Beck, A. T., Steer, R. A., & Brown, G. (1996). *Depression Inventory-II (BDI-II)*. American Psychological Association (APA) PsycTests. https://doi.org/10.1037/t00742-000

Bertrams, A., & Dickhäuser, O. (2009). Messung dispositioneller Selbstkontroll-Kapazität. *Diagnostica*, *55*(1), 2–10. https://doi.org/10.1026/0012-1924.55.1.2

Bieleke, M., Barton, L., & Wolff, W. (2021). Trajectories of boredom in self-control demanding tasks. *Cognition & Emotion*, *35*(5), 1018–1028. https://doi.org/10.1080/02699931.2021.1901656

Bieleke, M., & Wolff, W. (2021). The self-regulation of human performance. *Performance Enhancement & Health*, *9*(2).

Bijleveld, E. (2018). The feeling of effort during mental activity. *Consciousness and Cognition*, *63*, 218–227. https://doi.org/10.1016/j.concog.2018.05.013

Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annual Review of Psychology*, *40*(1), 109–131. https://doi.org/10.1146/annurev.ps.40.020189.000545

Cohen Hoffing, R. A., Lauharatanahirun, N., Forster, D. E., Garcia, J. O., Vettel, J. M., & Thurman, S. M. (2020). Dissociable mappings of tonic and phasic pupillary features onto cognitive processes involved in mental arithmetic. *PLoS ONE*, *15*(3), e0230517. https://doi.org/10.1371/journal.pone.0230517

Cooper-Martin, E. (1994). Measures of cognitive effort. *Marketing Letters*, *5*(1), 43–56. https://doi.org/10.1007/BF00993957

Danckert, J., & Eastwood, J. D. (2020). *Out of My Skull: The Psychology of Boredom*. Harvard University Press.

David, L., Vassena, E., & Bijleveld, E. (2022). *The aversiveness of mental effort: A meta-analytic review of the association between mental effort and negative affect*. https://doi.org/10.31234/osf.io/m8zf6

de Ridder, D. T. D., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., & Baumeister, R. F. (2012). Taking stock of self-control: A meta-analysis of how trait self-control relates to a wide range of behaviors. *Personality and Social Psychology Review*, *16*(1), 76–99. https://doi.org/10.1177/1088868311418749

Dora, J., van Hooff, M. L. M., Geurts, S. A. E., Kompier, M. A. J., & Bijleveld, E. (2022). The effect of opportunity costs on mental fatigue in labor/leisure trade-offs. *Journal of Experimental Psychology: General*, *151*(3), 695–710. https://doi.org/10.1037/xge0001095

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Hershman, R., & Henik, A. (2020). Pupillometric contributions to deciphering Stroop conflicts. *Memory & Cognition*, *48*(2), 325–333. https://doi.org/10.3758/s13421-019-00971-z

Hofmann, W., Friese, M., & Strack, F. (2009). Impulse and Self-Control From a Dual-Systems Perspective. *Perspectives on Psychological Science*, *4*(2), 162–176. https://doi.org/10.1111/j.1745-6924.2009.01116.x

Hofmann, W., Luhmann, M., Fisher, R. R., Vohs, K. D., & Baumeister, R. F. (2014). Yes, but are they happy? Effects of trait self-control on affective well-being and life satisfaction. *Journal of Personality*, *82*(4), 265–277. https://doi.org/10.1111/jopy.12050

Hopstaken, J. F., van der Linden, D., Bakker, A. B., & Kompier, M. A. J. (2015). The window of my eyes: Task disengagement and mental fatigue covary with pupil dynamics. *Biological Psychology*, *110*, 100–106. https://doi.org/10.1016/j.biopsycho.2015.06.013

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The Effort Paradox: Effort Is Both Costly and Valued. *Trends in Cognitive Sciences*, *22*(4), 337–349. https://doi.org/10.1016/j.tics.2018.01.007

Johnson, N. E., Saccuzzo, D. P., & Larson, G. E. (1995). Self-Reported Effort versus Actual Performance in Information Processing Paradigms. *The Journal of General Psychology*, *122*(2), 195–210. https://doi.org/10.1080/00221309.1995.9921232

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron*, *89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Kardon, R. H. Anatomy and physiology of the autonomic nervous system. In Miller, N. R., Newman, N. J., Biousse, V., & Kerrison, J. B. (Ed.), *Wash and Hoyt's Clinical Neuro-Ophthalmology* (6th ed., 649-714).

Koelewijn, T., Kluiver, H. de, Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research*, *323*, 81–90. https://doi.org/10.1016/j.heares.2015.02.004

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, *2*(12), 899–908. https://doi.org/10.1038/s41562-018-0401-9

Kühner, C., Bürger, C., Keller, F., & Hautzinger, M. (2007). Reliabilität und Validität des revidierten Beck-Depressionsinventars (BDI-II). Befunde aus deutschsprachigen Stichproben [Reliability and validity of the Revised Beck Depression Inventory (BDI-II). Results from German samples]. *Der Nervenarzt*, *78*(6), 651–656. https://doi.org/10.1007/s00115-006-2098-7

Kurzban, R. (2016). The sense of effort. *Current Opinion in Psychology*, *7*, 67–70. https://doi.org/10.1016/j.copsyc.2015.08.003

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *The Behavioral and Brain Sciences*, *36*(6), 661–679. https://doi.org/10.1017/S0140525X12003196

Marcora, S. M., Staiano, W., & Manning, V. (2009). Mental fatigue impairs physical performance in humans. *Journal of Applied Physiology*, *106*(3), 857–864. https://doi.org/10.1152/japplphysiol.91324.2008

Martarelli, C. S., Bertrams, A., & Wolff, W. (2021). *Short Boredom Proneness Scale - German Version (SBPS) [Database record].* APA PsycTests. https://doi.org/10.1037/t82956-000

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, *1*(1), 1–23. https://doi.org/10.5334/joc.18

McDougal, D. H., & Gamlin, P. Pupillary Control Pathways. In *The Senses: A Comprehensive Reference* (Vol. 1, pp. 521–536). https://doi.org/10.1016/B978-012370880-9.00282-6

Mills, C., & Christoff, K. (2018). Finding Consistency in Boredom by Appreciating its Instability. *Trends in Cognitive Sciences*, *22*(9), 744–747. https://doi.org/10.1016/j.tics.2018.07.001

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M., & Caspi, A. (2011). A

gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(7), 2693–2698. https://doi.org/10.1073/pnas.1010076108

Müller, T., & Apps, M. A. J. (2019). Motivational fatigue: A neurocognitive framework for the impact of effortful exertion on subsequent motivation. *Neuropsychologia*, *123*, 141–151. https://doi.org/10.1016/j.neuropsychologia.2018.04.030

Pickering, T., Wright, B., Schücker, L., & MacMahon, C. (2023). Active or passive? Investigating different types of cognitive fatigue. *Canadian Journal of Experimental Psychology = Revue Canadienne De Psychologie Experimentale*, *78*(1), 50–65. https://doi.org/10.1037/cep0000312

Rajkowski, J., Lu, W., Zhu, Y., Cohen, J., & Aston-Jones, G [G.] (2000). Prominent projections from the anterior cingulate cortex to the locus coeruleus in Rhesus monkey. *Society for Neuroscience Abstract*, *26*, 838–15.

Robinson, M. M., & Morsella, E. (2014). The subjective effort of everyday mental tasks: Attending, assessing, and choosing. *Motivation and Emotion*, *38*(6), 832–843. https://doi.org/10.1007/s11031-014-9441-2

Rothermund, K., Gollnick, N., & Giesen, C. G. (2022). Accounting for Proportion Congruency Effects in the Stroop Task in a Confounded Setup: Retrieval of Stimulus-Response Episodes Explains it All. *Journal of Cognition*, *5*(1), 39. https://doi.org/10.5334/joc.232

Struk, A. A., Carriere, J. S. A., Cheyne, J. A., & Danckert, J. (2017). A Short Boredom Proneness Scale. *Assessment*, *24*(3), 346–359. https://doi.org/10.1177/1073191115609996

Tangney, J. P., Boone, A. L., & Baumeister, R. F. (2018). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. In R. F. Baumeister (Ed.), *World library of psychologists. Self-regulation and self-control: Selected works of Roy Baumeister* (pp. 173–212). Routledge. https://doi.org/10.4324/9781315175775-5

Timme, S., Wolff, W., Englert, C., & Brand, R. (2022). Tracking Self-Control - Task Performance and Pupil Size in a Go/No-Go Inhibition Task. *Frontiers in Psychology*, *13*, 915016. https://doi.org/10.3389/fpsyg.2022.915016

Tyler, J. M., & Burns, K. C. (2008). After Depletion: The Replenishment of the Self's Regulatory Resources, *7*(3), 305–321. https://doi.org/10.1080/15298860701799997

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

van Tilburg, W. A. P., & Igou, E. R. (2017). Boredom begs to differ: Differentiation from other negative emotions. *Emotion*, *17*(2), 309–322. https://doi.org/10.1037/emo0000233

Wals, S. F., & Wichary, S. (2023). Under Pressure: Cognitive Effort During Website-Based Task Performance is Associated with Pupil Size, Visual Exploration, and Users' Intention to Recommend. *International Journal of Human-Computer Interaction*, *39*(18), 3504–3515. https://doi.org/10.1080/10447318.2022.2098576

Waugh, C. E., Shing, E. Z., & Avery, B. M. (2015). Temporal Dynamics of Emotional Processing in the Brain. *Emotion Review*, *7*(4), 323–329. https://doi.org/10.1177/1754073915590615

Wennerhold, L., & Friese, M. (2023). Challenges in the conceptualization of trait self-control as a psychological construct. *Social and Personality Psychology Compass*, *17*(3), Article e12726. https://doi.org/10.1111/spc3.12726

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PLoS ONE*, *8*(7), e68210. https://doi.org/10.1371/journal.pone.0068210

Westgate, E. C., & Wilson, T. D. (2018). Boring Thoughts and Bored Minds: The MAC Model of Boredom and Cognitive Engagement. *Psychological Review*, *125*(5), 689–713. https://doi.org/10.31234/osf.io/9j86p

Wolff, W., & Martarelli, C. S. (2020). Bored Into Depletion? Toward a Tentative Integration of Perceived Self-Control Exertion and Boredom as Guiding Signals for Goal-Directed Behavior. *Perspectives on Psychological Science*, *15*(5), 1272–1283. https://doi.org/10.1177/1745691620921394

Wolff, W., Radtke, V. C., & Martarelli, C. S. (2024). Same same but different. In M. Bieleke, W. Wolff, & C. Martarelli (Eds.), *The Routledge International Handbook of Boredom* (pp. 5–29). Routledge. https://doi.org/10.4324/9781003271536-3

Wolff, W., Sieber, V., Bieleke, M., & Englert, C. (2021). Task duration and task order do not matter: No effect on self-control performance. *Psychological Research*, *85*(1), 397–407. https://doi.org/10.1007/s00426-019-01230-1

Zénon, A., Sidibé, M., & Olivier, E. (2014). Pupil size variations correlate with physical effort perception. *Frontiers in Behavioral Neuroscience*, *8*, 286. https://doi.org/10.3389/fnbeh.2014.00286