How Effortful is Boredom? Studying Self-Control Demands Through Pupillometry

Registered Report Stage 2 (under review)

Vanessa C. Radtke[1,2], Wanja Wolff[2,3], and Corinna S. Martarelli[1]

[1] Faculty of Psychology, UniDistance Suisse, Brig, Switzerland

[2] Faculty of Sports Science, University of Konstanz, Konstanz, Germany

[3] Faculty of Educational Psychology, University of Bern, Bern, Switzerland

Author Note

We have no known conflict of interest to disclose. Study materials, data, and scripts are shared on Open Science Framework (OSF) at https://osf.io/jn54q/.

Correspondence concerning this article should be addressed to Vanessa C. Radtke, UniDistance Suisse, Schinerstrasse 18, 3900 Brig, Switzerland.

Email: vanessa.radtke@fernuni.ch

CRediT authorship contribution statement

Vanessa C. Radtke: Writing – original draft, Methodology, Investigation, Data curation, Visualization, Formal analysis, Project administration, Conceptualization. **Wanja Wolff:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization, Funding acquisition. **Corinna S. Martarelli:** Writing – review & editing, Supervision, Methodology, Project administration, Conceptualization, Funding acquisition.

Abstract

Self-control is essential for managing actions, yet its exertion is perceived as effortful. Performing a task may require effort not only because of its inherent difficulty but also due to its potential for inducing boredom, as boredom has been shown to be self-control demanding itself. So far, the extent of self-control demands during boredom and its temporal dynamics remain elusive. We employed a multimethod approach to address this knowledge gap. Ninety-five participants took part in an easy and hard version of the Stroop task. During both tasks, they indicated several times their sensation of task difficulty, boredom, boredom-related effort, difficulty-related effort, overall effort, and fatigue. We tested if pupil size, as a physiological indicator of cognitive effort, was predicted more accurately by difficulty- and boredom-related effort together than by task-difficulty-related effort alone. The best model fit included boredom-, difficulty-related effort, and their interactions with task type (easy, hard Stroop). Tonic pupil size increased during the easy Stroop, while phasic pupil size decreased with greater boredom-related effort in both tasks. Greater difficulty-related effort was linked to increases in tonic and phasic pupil size in the easy, but not in the hard Stroop. Finally, boredom-related effort in the Stroop predicted performance in a subsequent flanker task. Our results provide preliminary support that enduring boredom may not only be perceived as effortful but also be reflected in psychophysiological changes. Moreover, it may influence subsequent behavior. This underscores the importance of considering boredom as a potential confound in self-control research and broader study designs.

Preregistered Stage 1 protocol: https://osf.io/tv5df (date of in-principle acceptance: 09/09/2024).

Keywords: boredom, cognitive effort, self-control, pupillometry, Stroop

Public Significance Statement

This study seeks to uncover the relationship between boredom and cognitive effort during task completion. To disentangle the dynamic interplay between boredom and cognitive effort during both easy and difficult tasks, we employed a combination of objective measures (pupil size) and subjective measures (self-reports). Understanding how boredom influences cognitive effort required during the fulfillment of different self-control demanding tasks sheds light on a crucial aspect of human behavior, specifically how enduring boredom impacts cognitive effort during task completion. Our findings carry important implications for society and well-being, as they offer insights into strategies for enhancing task performance and maintaining motivation in various areas of life.

Registered Report: How Effortful is Boredom? Studying Self-Control Demands Through Pupillometry

To effectively navigate everyday life, people have to control their behavior on a regular basis (Hofmann et al., 2012). For example, in order to make progress towards the goal of alleviating their back pain, a person has to control the impulse to linger on their couch and has to engage in repeated stretching and muscle strengthening exercises instead. 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, 2021b) 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 effort. Notably, self-control is also influenced by motivational aspects which would not only facilitate the regulation of behavior towards ones goals, but also reduce the sensation of effort (Wennerhold & Friese, 2023) and consequently its costs.

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]. A large body of research shows that the objective and the perceived investment of effort tends to feel unpleasant and aversive (David et al., 2022; Kool & Botvinick, 2018; Wolff et al., 2021). On the other hand, effort is considered to feel unpleasant due to the costs of continuing with the ongoing activity (Kurzban, 2016). Thus, while effort is instrumental for effective self-control, it appears to carry a momentary cost, and the prolonged exertion of effort creates cumulative costs, such as fatigue or tiredness (Ainslie, 2021; Hopstaken et al., 2015; Kurzban, 2016; Kurzban et al., 2013; Westbrook et al., 2013). In addition to efforts' intrinsic costs (Kool et al., 2013), mobilizing effort towards one goal creates opportunity costs: When we direct our effort towards one activity (e.g., exercising to tackle backpain), we have to forego other, potentially more rewarding activities (e.g., going for a walk with friends). In this case, the aversive sensation of effort is considered to index the costs of continuing with the ongoing activity (Kurzban, 2016). This can create added self-control demands that make pursuing one's goal and continuing with a task even more self-control demanding and by extension more effortful (Kurzban et al., 2013). In light of effort's costs (intrinsic and/or opportunity costs), people are selective about when to invest effort in the service of self-control and the sensation of effort has been conceived as an index for the momentary costs of self-control (Bieleke et al., 2023).

Momentary Self-Control Costs and Performance in Self-Control Demanding Tasks

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

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

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). To test this, a plethora of studies has investigated the effects exerting self-control has on accumulated self-control costs (e.g., exertion, fatigue, tiredness; Wolff et al., 2021) as well as on performance in subsequent self-control demanding tasks (see Dang, 2018; Giboin & Wolff, 2019; Hagger et al., 2010). This has most frequently been studied with a sequential task design (Dang et al., 2021; Englert & Bertrams, 2021; Solomon et al., 1980; Vassena et al., 2014). Here, participants are asked to complete either a more difficult or less difficult version of a task before performance on a secondary task is measured. While a large body of research has found that prior self-control exertion has detrimental effects on performance in subsequent self-control demanding tasks (Giboin & Wolff, 2019; Hagger et al., 2016; Hagger et al., 2010), high-powered pre-registered replication failures and evidence for a large amount of grey literature (Wolff et al., 2018), as well as for publication bias has called the evidential value of this proposition into question (Carter & McCullough, 2014; Holgado & Mesquida, 2023). One potential reason for these mixed findings might be that the hard and easy tasks that are typically used in sequential task design studies, differ not only with respect to their structural self-control demands caused by task difficulty, but on other properties too. One such difference could be that easy variants might not only be less challenging but also more boring (Wolff & Martarelli, 2020). Take, for instance, the Stroop task (Stroop, 1935), which is frequently used in self-control research (Dang, 2018; Wolff et al., 2018). Here, participants are asked to categorize a stream of color words (e.g., red, blue) according to the font color they are presented in. In the easy version, the font color and the semantic meaning of the word are the same (i.e., congruent). In contrast, font color and semantic meaning differ (i.e., incongruent) in the hard version[2]. Consistent with the hypothesis that hard and easy Stroop versions might differ not only with respect to their structural self-control demands caused by task difficulty, participants tend to find the easy version of the Stroop to be more boring (Bieleke et al., 2021; Hunte et al., 2022; Mangin et al., 2021). This difference matters because continuing to work on a boring task is understood to be self-control demanding in its own right (Bieleke & Wolff, 2021a).

How Could Boredom Act as a Self-Control Demand?

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). Crucially, boredom is consequential because it is understood to signal that one should do something else (Bench & Lench, 2013; Elpidorou, 2014). With respect to the mechanisms by which boredom operates, research shows that boredom leads to a devaluation of the current activity, and increased reward sensitivity (Milyavskaya et al., 2019), thereby prompting exploration behavior (Agrawal et al., 2022; Bench & Lench, 2013; Danckert, 2019; Geana et al., 2016). By biasing behavior towards alternative activities, boredom increases the opportunity costs of sticking with what one is currently doing, thereby increasing the self-control demands (Bieleke & Wolff, 2021a; Bieleke et al., 2023) and the perceived effort (Eastwood et al., 2012) needed to keep engaged with the boring task. Take for example a lengthy zoom call. During such a meeting, many readers have probably experienced moments of boredom and have felt the urge to look at their phone and browse social media. Not giving in to this boredom-induced urge has likely made it harder to stay focused on what is discussed in the call. Consequently, self-control and

[2] In more demanding versions, participants even have to respond to instantaneous rule changes in the categorization task, thereby adding task switching as a further self-control demand.

effort are not only required for performing challenging activities, but also for underchallenging ones, making boredom a possible confound in self-control research but also in other research fields (Meier et al., 2023) that has been largely overlooked for so far (Wolff & Martarelli, 2020).

The Stroop task can provide an intuitive example of this. At the very beginning of the task, a period to become accustomed to it is required. During this phase, participants need to familiarize themselves with several key aspects, such as the colors presented, the location of the corresponding keys, and the task instructions. These additional challenges can result in a higher task difficulty, potentially leading to a greater demand for cognitive effort to successfully complete the task. In addition, the task would still feel novel, thereby likely being less boring for the participant. However, as the task progresses, participants become increasingly familiar with its demands. They encountered all the words and colors in different combinations, repeatedly pressed the keys, and have a clear understanding of the instructions. Consequently, the task can get easier, leading to a reduction in the self-control demands of the task. However, as the same stimuli appear and the same keys are pressed over a longer period, the task might also get more and more boring. Thus, while potentially less effort is required to deal with the structural difficulties of the task, more effort might be needed to ward of the boredom-induced urge to get the task over with (Bieleke et al., 2021). 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. 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.

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 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.

The Present Study

In the present study, we investigated how effortful task-induced boredom is subjectively perceived and if it can be physiologically tracked. To achieve this, we assessed 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 integrated a promising research protocol that complemented 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. Subjective experiences were assessed several times during the experiment with thought probes asking for the perception of boredom, task difficulty, task difficulty related effort, boredom related effort, overall 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. A secondary task (flanker task) was included in our study to further assess how the processing of the first task influences performance in the second task. However, note that our primary focus was on examining the temporal dynamics of task difficulty and boredom related effort in the easy and hard version of the first task (Stroop task). To the best of our knowledge, this is the first study directly assessing the perception of effort due to boredom and disentangling task difficulty and boredom related effort. Moreover, implementing a within-subjects design, this study contributes valuable insights into how boredom and effort interact across varying difficulty levels of cognitive tasks within the same participants. By uncoupling effort from its direction (i.e., effort to deal with an intrinsically hard task vs. effort to keep working on a boring task), this study addresses the research gap on how enduring boredom impacts effort during the performance of cognitive tasks. As a result, this study enhances our understanding of boredom and its role in self-control. We predicted boredom and boredom related effort to increase over time while we assumed task difficulty and task difficulty related effort to decrease. Furthermore, we predicted that pupil size (tonic and phasic) is predicted more accurately by overall cognitive effort (difficulty- and boredom-related) than by effort related solely to task difficulty.

Method

Sample Characteristics

To ensure a significant difference in the difficulty level and the potential for boredom induction by our tasks, we conducted an online study that included four Stroop task versions (easy and hard versions of color Stroop and numerical Stroop, see Appendix A). Based on the results, we concluded that the color Stroop task was the most suitable variant for our study. 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 was larger for task difficulty (d = 0.92) than for boredom (d = 0.34), our power analysis focused on the effect size of the difference in boredom between the easy and the hard color Stroop. Thus, we adopted a conservative approach in calculating the required sample size. The power analysis for a one-tailed paired t-test was calculated based on this effect size. The analysis indicated that 95 participants were necessary to detect a difference in boredom between the tasks with a power of 95% at an alpha level of 0.05. Replacement participants were recruited to achieve the calculated sample size, when participants drop out, z-transformed error rate in the Stroop task was more than three standard deviations away from the mean of all participants or errors in the recording of oculomotor data occurred (i.e., participants for whom the calibration procedure did not work correctly, indicated by an average error in measuring horizontal and vertical eye positions exceeding an angular accuracy of 0.8°). A total of 110 participants took part in at least one session. However, 15 participants were excluded based on these criteria: five completed only one session (two of them due to measurement errors in their first session), five had excessive error rates, five were excluded due to incomplete oculomotor data.[3]

[3] A statistical comparison between included participants and participants that were excluded due to low performance or attendance of only one session is presented in Appendix D (Participant Characteristics). Participants differed significantly in age and trait self-control, but not in gender, boredom-proneness and

After exclusions, ninety-five participants without self-reported color blindness remained in the sample (*M* = 20.22 years, *SD* = 4.48, age range = 16 to 45 years; 75.79% women, 23.16% men, 1% other gender). All participants were recruited from the general population and from our institute.

Participants provided written informed consent and were free to end the experiment at any time. Parental consent was not required for participants under 18 years of age (i.e., 16 or 17 years old). They received 100 CHF as compensation. The local Ethics Committee approved the study, which was conducted according to the principles of the Declaration of Helsinki.

Materials and Design

In this within-subjects design, study participants took part in two sessions separated by at least one day, each consisting of a Stroop task and a subsequent flanker task. In one session, participants completed the Stroop task in its easy congruent-only version (easy Stroop), while in the other session, they completed the hard version with modified instructions (hard Stroop). Pupil size and thought probe data were recorded during the Stroop task only, while behavioral data (response time, accuracy) was recorded during both tasks.

Tasks

Stroop task. As primary task participants performed a color Stroop task (original Stroop task version first implemented by Stroop, (1935) comprising 34 practice trials and 360 experimental trials. Each trial consisted 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 was 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). Participants were instructed to press one of four keys on the

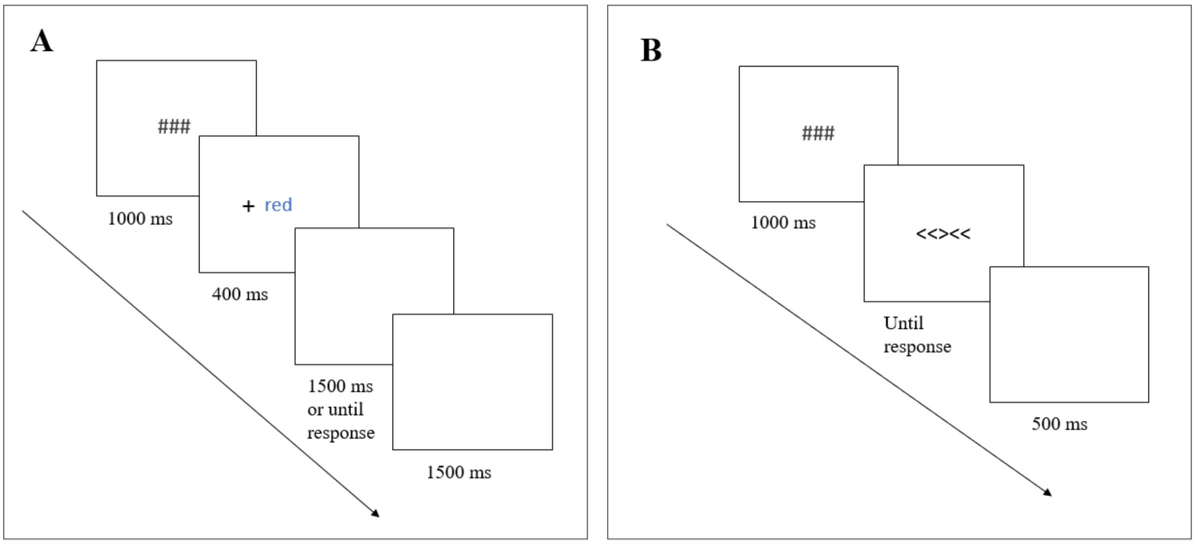
depression scores.

keyboard (“1” for green, “2” for blue, “9” for red, “0” for yellow) to indicate their response. The background of the screen was displayed in grey (RGB: 150,150,150,255). Self-control demands and the induction of boredom were manipulated using an easy version of the Stroop task and a hard version with modified instructions. While in the easy condition participants received the instruction to indicate the color of the presented color word, the task in the hard condition switched according to another stimulus presented on screen, which either demanded to indicate the color if a “+” was displayed (in 80 % of trials) or the word if a “x” was displayed (in 20% of trials). The instructions informed participants that in the easy condition all trials would be congruent (color and color word matching), whereas in the hard condition all trials would be incongruent (color and word not matching each other, see Figure 1 A). Response times and accuracy served as performance indicators.

Flanker task. As secondary task, participants performed a flanker task (original task version first implemented by Eriksen & Eriksen, 1974), comprising 24 practice trials and 140 experimental trials. Each trial consisted of fixation stimuli (###, displayed for 1000 ms on screen), five arrows presented on the middle of the screen, and an intertrial interval (displayed for 500 ms). The arrows remained on screen until participants pressed the keys “1” for right or “0” for left, indicating the direction of which the central arrow was pointing. The other four arrows were either congruent and pointed to the same direction as the central arrow (50 % of all trials) or incongruent and thus pointing to the contrary direction (see Figure 1 B). The task performance (accuracy, response times) was recorded. The flanker task was implemented to investigate potential effects of the primary tasks on secondary task performance.

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) was the task in 80 % of the trials of the hard Stroop task (A). In 20 % of the trials an "x" was presented and thus, the task was to indicate the word. All trials of the hard Stroop task were incongruent whereas all trials of the easy Stroop task were congruent. Flanker task trials (B) were either incongruent or congruent (50 % of all trials).

Thought Probes

During the Stroop task participants were prompted eleven times to report how bored they were ("How much boredom do you feel?"), as how difficult they were experiencing the task ("How difficult is the task?"), how much effort they were 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 were experiencing ("How much fatigue do you feel?"). They indicated their answer by pressing a key between 1 (not at all) and 9 (very much) on their keyboard. The probes were displayed randomly after blocks of 30 to 35 trials. The order of the probes was randomized for each presentation.

Pupil Size

Tonic (baseline) pupil size and phasic (stimulus-evoked) pupil size during the tasks were recorded at a sampling rate of 1000 Hz, a spatial resolution of 0.01°, and a gaze position accuracy of 0.05° using an EyeLink 1000 Plus eye-tracker. A chinrest was placed at a distance of 45 cm to the monitor (1920 x 1080 pixels) in order to prevent head movements. At the start of the experiment the eye-tracker was calibrated and validated using a 9-point grid. Participants, whose average horizontal and vertical error exceeded an angular accuracy of 0.8° were excluded. The testing room was cut off from external light sources, and it was not illuminated throughout all testing sessions.

The data was extracted using EyeLink DataViewer version 3.4.1. Missing data points due to off-screen fixations or eye blinks were removed before analyzing pupil diameter. Pupil size is reported for the right eye and reported in pixels.

Tonic (baseline) pupil size was averaged separately for each last 500 ms of the fixation stimuli (###) presentation. Phasic (stimulus-evoked) peak pupil dilation for each trial was 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 were averaged in blocks with each block consisting of the last five trials before each probe. A larger phasic (stimulus-evoked) peak pupil dilation served as indicator for higher stimulus-induced cognitive effort. Conversely, a lager tonic (baseline) pupil size served 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 the endurance of boredom on cognitive effort and could lead to an overlook of this effect.

Supplementary Measures

In addition to the measures needed to investigate the core research question of this paper we implemented 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 completed three questionnaires before taking 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 also indicated whether they had been diagnosed with ADHD. They completed the questionnaire online via a link provided by the experimenter, which directed them to a LimeSurvey form[4]. On average, 29.33 days (*SD* = 20.27) elapsed between completing the questionnaire and attending the first experimental session. Immediately before beginning the session, participants indicated if they had consumed any stimulants on the day of testing (caffeine, nicotine, amphetamine, others), and whether they felt sleep deprived ("Do you feel sleep deprived*?"* Answered with "yes" or "no"). If participants indicated that they had taken stimulants before the first session, they were asked to do the same before the second session[4]. Responses were included in the sample description. After completing the experiment, participants were offered to take with them some sweets. The amount chosen by each participant was measured. The corresponding analyses and characteristics of participants are provided in Appendix D.

Procedure

Before starting the first experimental session in the laboratory, participants answered the Short Boredom Proneness Scale (Struk et al., 2017), the Brief Self-Control Questionnaire (Tangney et al., 2018) and the BDI-II (Beck et al., 1996) as well as the other supplementary

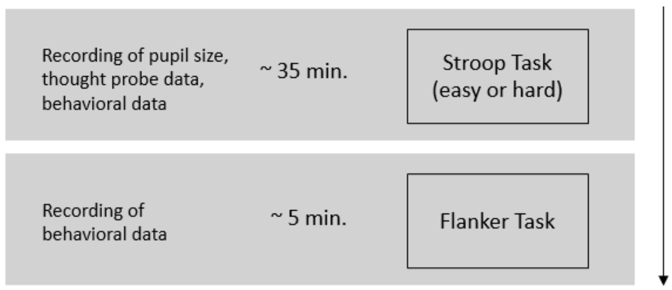
[4] Added as a clarification at Stage 2. This procedural detail does not alter any preregistered aspect of data collection, hypotheses, or analyses.

measures online. Afterwards they received an invitation to the laboratory together with the request not to drink any coffee at least for two hours before the start of the experiment. In each study part, online and offline, participants were informed about the experimental procedure and asked to consent into the study conditions. The experimental part of the study consisted of two sessions. Every participant took part in both sessions (within-subjects design). Before each session participants reported if they had consumed any stimulants on the day of testing. The procedure of both experimental sessions (see Figure 2) was similar with the difference that the first sequential paradigm task (Stroop task) was an easy version in one session and a hard in the other one. Both sessions involved the flanker task as secondary task. The order of the sessions was counterbalanced. Each session took approximately 40 to 50 minutes. During the task, participants answered to probes indicating their perception of boredom, task difficulty, effort they had to invest due to boredom and due to task difficulty, overall effort and fatigue. There was a break of at least one day in between the sessions (M = 11.88, SD = 11.32, range from 1 to 52).

In the experimental sessions participants placed their head on a chinrest and looked at the screen (at 45 cm distance to head) where the tasks were presented while oculomotor data (pupil size) and behavioral data (task performance, probe answers) was recorded. Participants were instructed to stay as still as possible, to direct their gaze to the center of the screen and to answer the questions (probes) presented on screen by pressing the corresponding number on the keyboard without removing their head from the chinrest.

Figure 2

Procedure for Each Session



Note. Behavioral data (reaction time, accuracy) was recorded during the Stroop task and during the Flanker task. Pupil and thought probe data was recorded during the Stroop task only.

Pre-Test

To test the feasibility of our study design we tested two participants and recorded pupil size (tonic and phasic pupil size) and the answers to the thought probes before starting the main study. See Appendix B for the corresponding descriptive data.

Transparency and Openness

We share the dataset of our online study as well as the R code to reproduce our analyses on Open Science Framework (OSF) available at https://osf.io/jn54q/. We further incorporated the research materials, the full non-aggregated dataset as well as the analysis scripts for the present study. The data was analyzed using R, version 4.3.1 (R Core Team, 2023).

Results

We first report the manipulation check, followed by temporal analyses of the thought probes, analyses of the relationship between effort and pupil size, and finally, the impact of completing the first task (Stroop) on performance in the second task (Flanker). Statistical significance for all analyses was evaluated at $\alpha = .05$ (two sided)[4]. For further details on preregistered analyses, see Appendix C (Study-Design Template).

Manipulation Check

We confirmed the effectiveness of our experimental manipulations by assessing their impact on participants' subjective perception and performance. This manipulation check allowed us to ensure that the intended differences between the conditions concerning the manipulated difficulty level were indeed present.

Task Difficulty

We hypothesized that the hard condition would exhibit significantly greater overall perceived task difficulty compared to the easy condition. To test this hypothesis, we performed a paired t-test comparing the overall task difficulty (dependent variable) in the easy and hard Stroop (difficulty level as independent variable). The overall perceived task difficulty was calculated as the mean of the perceived difficulty ratings for the answered probes. As expected, results revealed a significant lower perceived task difficulty in the easy (M = 3.02, SD = 1.77) compared to the hard (M = 6.66, SD = 1.73) task, with a large effect size, $t(94) = -18.98$, $p < .001$, $d = -1.95$.

Boredom

In the context of boredom levels, we expected the easy condition to induce significantly greater overall boredom compared to the hard condition. To evaluate this hypothesis, we conducted a paired t-test comparing overall boredom (dependent variable) in the easy and hard condition (difficulty level as independent variable). Overall boredom was obtained by calculating the mean of the boredom ratings for the answered probes. As expected, the analysis confirmed that perceived boredom was significantly greater in the easy (M = 5.47, SD = 1.75) compared to the hard (M = 3.99, SD = 1.94) task, with a moderate effect size, $t(94) = 6.81$, $p < .001$, $d = 0.70$.

Performance

Performance differences between the two conditions were assessed by examining error rates and reaction times. We expected participants to demonstrate a greater error rate and longer reaction time in the hard condition compared to the easy condition. We conducted two paired t-tests, one comparing the error rate and one comparing the reaction time between the easy and hard condition, implementing the error rate and the reaction time as dependent variables each in one analysis and the difficulty level (easy, hard) as independent variable. Error rates and reaction times were calculated as the mean values across all completed trials, including correct and incorrect responses[4]. Consistent with our hypothesis, results indicated significant lower error rates in the easy (M = 0.04, SD = 0.04) compared to the hard (M = 0.19, SD = 0.13) Stroop, $t(94) = -11.73$, $p < .001$, $d = -3.16$. Similarly, the paired t-test for reaction time showed that participants were significantly faster in the easy (M = 707.45 ms, SD = 108.57 ms) compared to the hard (M = 1138.90 ms, SD = 162.69 ms) Stroop, $t(94) = -30.79$, $p < .001$, $d = -3.16$.

Thought Probes

We expected changes over time in participants' self-reports for both the easy and the hard variant of the task. Changes over time concerning those variables were analyzed conducting six linear mixed models including the perception of boredom, boredom related effort, task difficulty and task difficulty related effort, overall effort and fatigue as outcome variables, task (easy, hard) 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. We expected a main effect of time. More precisely, we expected an increase in boredom, boredom related effort, fatigue, and overall effort and a decrease of task difficulty and task difficulty related effort over time for both Stroop versions (easy and hard). Further, we predicted a main effect of difficulty level (easy, hard). We anticipated greater levels of boredom and greater boredom related effort, but less task difficulty and task difficulty related effort for the easy Stroop version. We expected an interaction for perceived task difficulty and task difficulty related effort with a greater decrease over time in the hard Stroop. We further expected an interaction for perceived overall effort with a greater increase in overall effort in the easy Stroop version. Six plots (see Figure 3) showing the time (probe one to eleven) on the x-axis, the scores on the y-axis, and two different lines for the two tasks, illustrate how the perception of boredom, boredom related effort, task difficulty related effort, task difficulty, fatigue and effort change over time.

Results of our linear mixed models supported most hypotheses. As expected, the analysis for task difficulty revealed a significant interaction between task and time (γ_{03} = -0.15, t = -12.04, p < 0.001, 95% CI [-0.17, -0.12]), indicating a greater decrease in difficulty over time in the hard Stroop. However, in contrast to our expectations, difficulty did not decrease in the easy Stroop. In addition, main effects of time (γ_{01} = 0.04, t = 2.71, p = 0.008, 95% CI [0.01, 0.08]) and task (γ_{02} = 4.52, t = 54.35, p < 0.001, 95% CI [4.36, 4.69]) were observed, showing a general increase in difficulty over time and overall higher difficulty ratings in the hard Stroop.

As hypothesized, the model predicting effort due to task difficulty showed a significant interaction of time and task (γ_{03} = -0.12, t = -9.25, p < 0.001, 95% CI [-0.14, -0.09]), with a greater decrease in task difficulty related effort over time in the hard Stroop. However, in contrast to our predictions, and similar to the trajectory of task difficulty, difficulty related effort did not decrease in the easy Stroop. Additionally, main effects of time (γ_{03} = -0.12, t = -9.25, p < 0.001, 95% CI [-0.14, -0.09]) and task (γ_{02} = 4.16, t = 47.90, p < 0.001, 95% CI [3.99, 4.34) were significant, indicating a general increase in difficulty related effort over time as well as higher reported difficulty related effort in the hard Stroop task.

Regarding boredom, the analysis revealed a significant interaction between task and time (γ_{03} = -0.04, t = -3.00, p = 0.003, 95% CI [-0.07, -0.01]), revealing a greater increase in boredom over time in the easy Stroop. As expected, we observed significant main effects of time (γ_{01} = 0.03, t = 15.54, p < 0.001, 95% CI [0.26, 0.34]) and task (γ_{02} = -1.23, t = -12.80, p < 0.001, 95% CI [-1.41, -1.04]), with increases in boredom in both tasks and significantly lower boredom in the hard Stroop task.

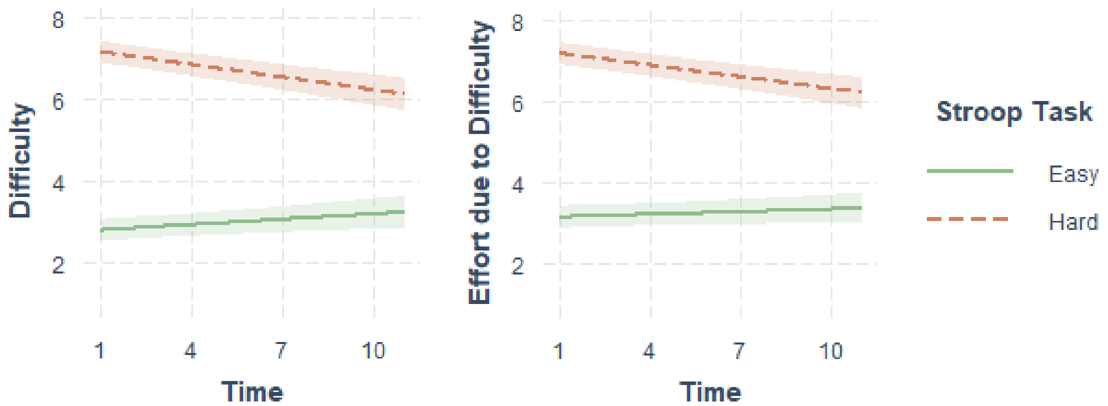
The model for boredom related effort showed a significant interaction between task and time (γ_{03} = -0.07, t = -4.36, p < 0.001, 95% CI [-0.10, -0.04]), indicating a greater increase in boredom related effort over time in the easy Stroop task. In line with our hypothesis, the main effect of time (γ_{01} = 0.22, t = 10.45, p < 0.001, 95% CI [0.18, 0.27]) was significant, indicating overall increases in boredom related effort over time. The main effect of task was not significant (γ_{02} = 0.19, t = 1.82, p = 0.069, 95% CI [-0.54, 0.39]).

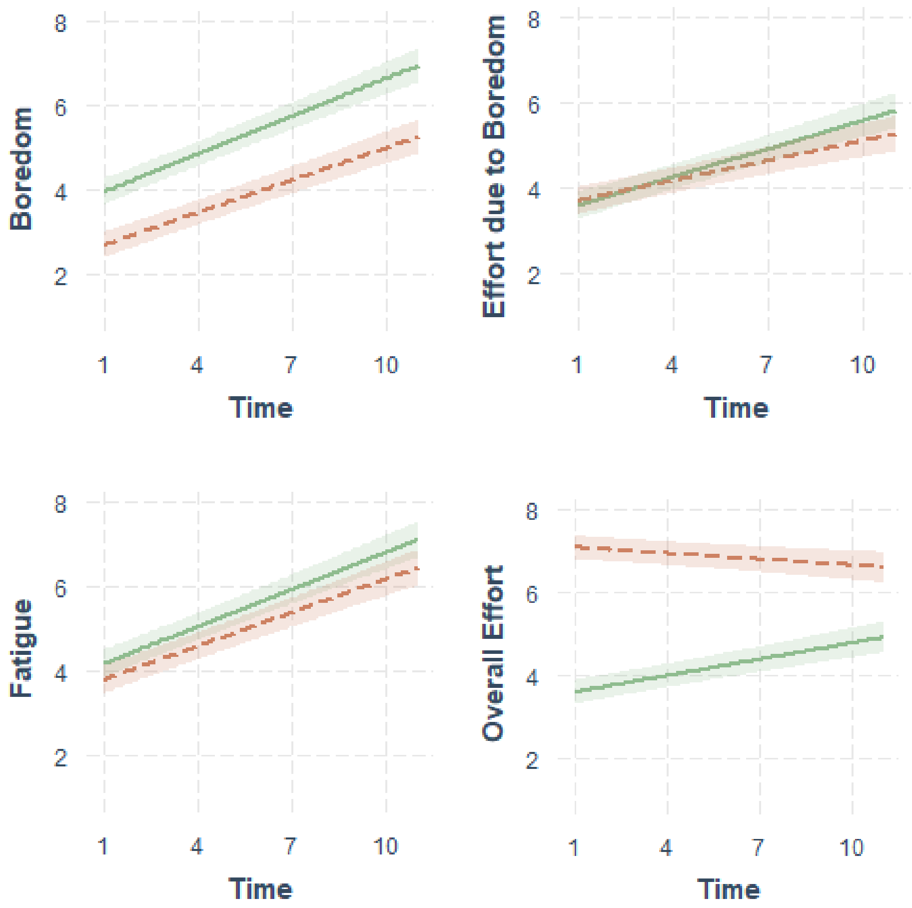
As hypothesized, the analysis for overall effort revealed a significant interaction between task and time (γ_{03} = -0.18, t = -14.40, p < 0.001, 95% CI [-0.20, -0.16]), indicating a greater increase in overall effort in the easy Stroop task over time while overall effort in the hard Stroop task even tend to decrease. Additionally, as predicted, a significant main effect of time (γ_{01} = 0.13, t = 8.81, p < 0.001, 95% CI [0.10, 0.16]) was observed, indicating increases in overall effort over time. Results further revealed a main effect of task (γ_{02} = 3.68, t = 43.36, p < 0.001, 95% CI [3.51, 3.84]), indicating greater overall effort reported in the hard Stroop task.

Finally, consistent with our expectations, a significant main effect of time (γ_{01} = 0.29, t = 15.91, p < 0.001, 95% CI [0.26, 0.33]) was observed for fatigue, with increasing levels of fatigue over time. Results further revealed a significant main effect of task (γ_{02} = -0.35, t = -3.51, p < 0.001, 95% CI [-0.54, -0.15]), with lower fatigue levels in the hard Stroop task. Our analysis indicated that the interaction between task and time was not significant (γ_{03} = -0.03, t = -1.95, p = 0.051, 95% CI [-0.06, 0.00]).

Figure 3

Effects of Task (Easy, Hard Stroop) and Time (Block One to Eleven) on Perceived Task Difficulty, Effort Related to Task Difficulty, Boredom, Effort Related to Boredom, Overall Effort, and Fatigue





Note. Linear mixed models revealed a significant main effect of task for all variables except for effort due to boredom (p = .07). They further demonstrated a significant main effect of time across all variables except for effort due to task difficulty (p = .19). Additionally, a significant interaction between time and task was observed for difficulty, effort related to difficulty, boredom, effort related to boredom, and overall effort (all $ps < 0.003$), but not for fatigue (p = .051). The confidence band indicates the 95% confidence interval.

Effort and Pupil Size

To test our prediction that pupil size shows a stronger association with perceived boredom related effort and task difficulty related effort together than with task difficulty related effort alone, we analyzed tonic (baseline) pupil size and phasic (stimulus-evoked) pupil size separately. We hypothesized that tonic (baseline) and phasic (stimulus-evoked) pupil sizes would be more accurately predicted when considering both perceived effort due to boredom and effort due to task difficulty, as opposed to considering perceived effort due to task difficulty alone.

To investigate this hypothesis, we employed eight linear mixed models (LMM) to assess the influence of these factors on tonic pupil size and phasic pupil size separately. Detailed model outputs are provided in Appendix E. Tonic pupil size for each probe was determined by calculating the average tonic (baseline) pupil size (pupil size during the last 500 ms of the presentation of the fixation stimuli) across the last five trials preceding the probe presentation. Phasic pupil size for each probe was determined by calculating the peak stimulus-evoked pupil size. After baseline correcting phasic pupil size it was averaged across the last five trials preceding the probe presentation. We constructed four different models for each outcome variable (tonic pupil size, phasic pupil size). All models were calculated using the lmer function from the *lme4* package in R version 4.3.1. (R Core Team, 2023) with the standard optimizer "nloptwrap"[4]. However, model three predicting tonic pupil size, as well as models one, three, and four predicting phasic pupil size, were estimated using the optimizer "bobyqa" due to convergence issues with the standard optimizer. The presented confidence intervals correspond to the non-standardized estimates.

Our first model predicted pupil size with the task's difficulty level (easy, hard), time (thought probe/block one to eleven), and perceived 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. Results of model one indicated a significant main effect of the task's difficulty level (easy, hard) (γ_{01} = 42.02, β = 0.14, t = 4.48, p < 0.001, 95% CI [3.22, 11.68]), time (γ_{02} = -6.25, β = -0.07, t = -2.96, p = 0.003, 95% CI [3.22, 11.68]), and effort related to task difficulty (γ_{03} = = 7.45, β = 0.06, t = 3.45, p < 0.001, 95% CI [3.22, 11.68]) on tonic pupil size. While the task's difficulty level (γ_{01} = 44.78, β = 0.75, t = 20.13, p < 0.001, 95% CI [40.42, 49.14]) and time (γ_{02} = -1.31, β = -0.07, t = -4.24, p < 0.001, 95% CI [-1.92, -0.71]) were also significantly related to phasic pupil size, contrary to our expectations, the main effect of effort related to task difficulty was not significant (γ_{03} = 0.71, β = 0.03, *t* = -0.27, *p* = 0.158, 95% CI [-0.27, 1.70]). Thus, tonic and phasic pupil size were larger during the hard Stroop version and decreased over time. The effects of the task's difficulty level and time remained similar in all models (see Appendix E for full estimates). Furthermore, tonic pupil size but not phasic pupil size was larger for participants reporting greater effort related to task difficulty.

Our second model differed from model one by implementing effort due to boredom as fixed predictor instead of effort due to task difficulty. The main effect of effort related to boredom was not significant for either tonic (γ_{03} = 3.25, β = 0.02, *t* = 1.79, *p* = 0.074, 95% CI [-0.32, 6.82]) or phasic pupil size (γ_{03} = -0.78, β = -0.03, *t* = -1.83, *p* = 0.067, 95% CI [-1.61, 0.05]).

Our third model integrated effort due to task difficulty and effort due to boredom as fixed predictors, while remaining with the other predictors. Similar to model one, larger tonic pupil size was significantly related to increased effort related to task difficulty (γ_{03} = 7.10, β = 0.06, *t* = 3.27, *p* = 0.001, 95% CI [-2.85, 11.36]), while phasic pupil size was not significantly related to effort related to task difficulty (γ_{03} = 0.85, β = 0.04, *t* = 1.66, *p* = 0.096, 95% CI [-0.15, 1.84]). In contrast to our expectations, boredom related effort was not significantly related to tonic pupil size (γ_{04} = 2.58, β = 0.02, *t* = 1.41, *p* = 0.158, 95% CI [-1.01, 6.17]), but significantly related to phasic pupil size (γ_{04} = -0.87, β = -0.03, *t* = -2.04, *p* = 0.042, 95% CI [-1.71, -0.03]), with smaller phasic pupil sizes when boredom related effort increased.

Our fourth model expanded upon model three by including additional covariates, namely participants' score of the Short Boredom Proneness Scale (SBPS), the Beck Depression Inventory II (BDI-II), and the Brief Self-Control Scale (BSCS; descriptive statistics and internal consistencies for these scales are reported in Appendix E, Table 6). This model continued to show similar results to model three. Neither the SBPS, nor the BSCS or BDI-II scores were significantly related to pupil sizes (see Appendix E for the detailed model output). Difficulty related effort was significantly related to larger tonic pupil sizes (γ_{03} =

7.17, $\beta = 0.06$, $t = 3.30$, $p = 0.001$, 95% CI [2.92, 11.43]) but not significantly related to phasic pupil size ($\gamma_{03} = 0.88$, $\beta = 0.04$, $t = 1.72$, $p = 0.085$, 95% CI [-0.12, 1.87]). As in model three, contrary to our expectations, no significant main effect of boredom related effort was observed for tonic pupil size ($\gamma_{04} = 2.57$, $\beta = 0.02$, $t = 1.40$, $p = 0.161$, 95% CI [-1.02, 6.16]), but phasic pupil sizes were significantly smaller with increased boredom related effort ($\gamma_{04} = -0.86$, $\beta = -0.03$, $t = -2.01$, $p = 0.044$, 95% CI [-1.70, -0.02]).

As not preregistered exploratory analyses, we included the interaction effects of task (easy, hard) and the effort variables (boredom related and difficulty related) in our third linear mixed models (included as model 3ex in Table 1). Results indicated significant interaction effects between task and difficulty related effort for tonic ($\gamma_{05} = -15.31$, $\beta = -0.13$, $t = -4.34$, $p < 0.001$, 95% CI [-22.22, -8.40]) and phasic pupil size ($\gamma_{05} = -3.76$, $\beta = -0.17$, $t = -4.44$, $p < 0.001$, 95% CI [-5.43, -2.10]). In the easy task, pupil sizes increased significantly with increasing difficulty related effort (tonic, $\gamma = 13.47$, SE = 2.81, 95% CI [7.92, 18.97], $z = 4.80$, $p < .001$; phasic, $\gamma = 2.63$, SE = 0.67, 95% CI [1.32, 3.94], $z = 3.94$, $p < .001$). In contrast, in the hard task, phasic pupil size was not significantly associated with difficulty related effort ($\gamma = -1.14$, SE = 0.66, 95% CI [-2.43, 0.16], $z = -1.72$, $p = .086$). Similarly, tonic pupil size was not significantly associated with difficulty related effort ($\gamma = -1.84$, SE = 2.80, 95% CI [-7.33, 3.65], $z = -0.66$, $p = .511$).

Our exploratory linear mixed models further revealed a significant interaction effect between task and boredom related effort for tonic pupil size ($\gamma_{06} = -8.58$, $\beta = -0.07$, $t = -2.80$, $p = 0.005$, 95% CI [-14.58, -2.58]), but not for phasic pupil size $\gamma_{06} = -0.20$, $\beta = -0.01$, $t = -0.27$, $p = 0.785$, $p = 0.785$, 95% CI [-1.64, 1.24]). While tonic pupil size significantly increased with greater boredom related effort in the easy task ($\gamma = 6.85$, SE = 2.52, 95% CI [1.92, 11.80], $z = 2.72$, $p = .007$), decreases in the hard task were not significant ($\gamma = -1.72$, SE = 2.26, 95% CI [-6.15, 2.70], $z = -0.76$, $p = .445$). Figure 4 illustrates these relationships.

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 are reported (see Table 1). When comparing models 1 and 3 as preregistered, model 1 was favored for tonic pupil size according to AIC and BIC, whereas WAIC indicated a preference for model 3. For phasic pupil size, AIC favored model 3, while BIC and WAIC favored model 1. When comparing all models, the models predicting tonic and phasic pupil size with the optimal balance between model fit and complexity according to the AIC, BIC, and WAIC were models 3ex, which included task, time, boredom related effort, difficulty related effort, as well as the interactions of the two effort variables with task, which was included for exploratory purposes.

Table 1

AIC, BIC, and WAIC Values of Linear Mixed Models Explaining Tonic Pupil Size (Left) and Phasic Pupil Size (Right)

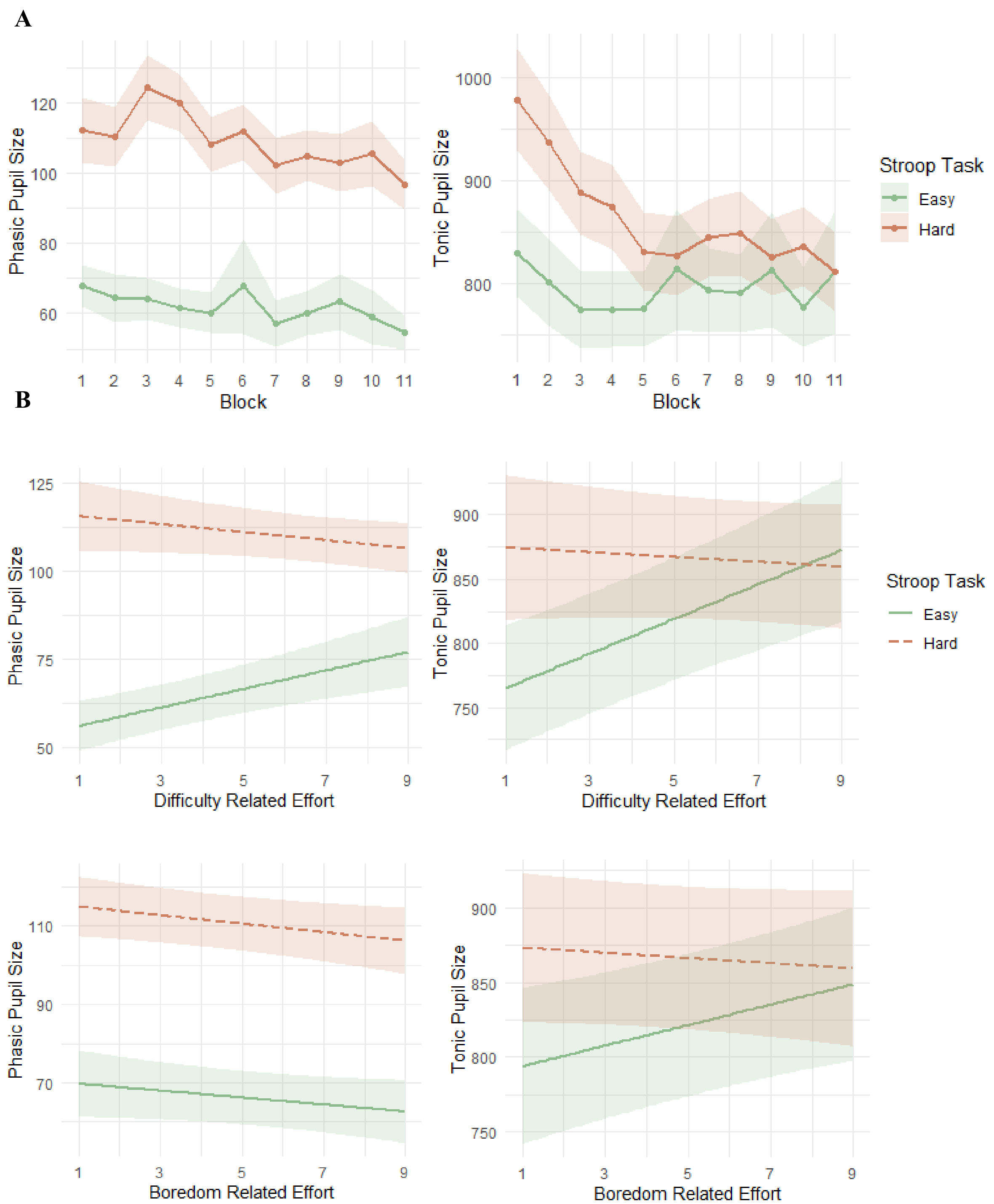
	Tonic Pupil Size			Phasic Pupil Size		
Model	AIC	BIC	WAIC	AIC	BIC	WAIC
Model 1	56024.8	56075.5	55701.2	44036.0	44086.7	43858.9
Model 2	56033.9	56084.6	55708.5	44035.0	44085.7	43851.9
Model 3	56078.8	56078.8	55700.5	44033.8	44090.8	43859.1
Model 3ex	**55979.2**	**56048.9**	**55653.6**	**44012.4**	**44082.1**	**43835.0**
Model 4	55995.6	56078.9	55684.9	44019.9	44095.9	43862.4

Note. Model 3ex, which includes task, time, boredom related effort, difficulty related effort and their interactions with task, exhibits the lowest AIC, BIC, and WAIC values for both tonic and phasic pupil size. Note, however, that 0.9% of the p_{WAIC} values exceeded 0.4, suggesting that the WAIC approximation may be less reliable for a small portion of the data. All models include task and time. Model 1 only adds difficulty related effort, model 2 only adds boredom related effort, model three includes both effort variables, model 4 extends model 3 by adding the questionnaire scores (Short Boredom Proneness Scale, Brief Self-Control Questionnaire, Becks Depression Inventory-II). The comparison between models 1 and 3 was preregistered. Model 3ex is an exploratory extension of model 3 that includes the interactions between the effort variables and task.

As an illustration three plots (see Figure 4) were implemented for each outcome variable (tonic pupil size, phasic pupil size) showing time and the scores of perceived boredom related effort and task difficulty related effort on the x-axis, the outcome variable on the y-axis and two different lines indicating the manipulated difficulty level (easy, hard).

Figure 4

Means of Phasic and Tonic Pupil Size for Block 1 to 11 (A), and Effects of Difficulty Related Effort and Boredom Related Effort on Tonic and Phasic Pupil Size (B).



Note. Exploratory analyses of linear mixed models revealed a significant interaction between task and difficulty related effort for phasic pupil size and tonic pupil size as well a significant interaction between task and boredom related effort for tonic pupil size. The confidence band indicates the 95% confidence interval.

Task Performance in Secondary Task

For assessing task performance in the secondary task, we focused on the error rate and reaction time separately. We aimed to analyze if participants' performance in the secondary task (flanker) is influenced by the difficulty level of the preceding Stroop type (easy Stroop, hard Stroop). We further intended to test whether perceived task difficulty related effort exerted in the Stroop task predicted task performance in the flanker task, and whether boredom and task difficulty related effort together predicted task performance more accurately.

Difficulty Manipulation's Influence on Performance

To assess whether the level of difficulty in the first task (Stroop task; easy, hard) influenced participants' performance in the secondary task (flanker), we conducted two 2 (preceding Stroop type: easy, hard) x 2 (flanker type: congruent, incongruent) repeated measures ANOVAs[5]. One ANOVA analyzed participants' error rate as dependent variable, while the other focused on the reaction time.

The ANOVA on error rates revealed significant effect of flanker congruency, $F(1, 94) = 31.85$, $p < .001$, $\eta^2_p = .25$, indicating that participants made more errors in incongruent ($M = 0.07$, $SD = 0.15$) compared to congruent trials ($M = 0.01$, $SD = 0.02$; flanker effect). The main effect of preceding Stroop type (easy Stroop, hard Stroop), $F(1, 94) = 1.29$, $p = .259$, $\eta^2_p = .01$, and the interaction between preceding Stroop type and flanker congruency was not significant, $F(1, 94) = 0.68$, $p = .412$, $\eta^2_p = .007$, suggesting that the difficulty level of the Stroop did not impact error rates in the flanker task.

Similar, the ANOVA on reaction times revealed a significant main effect of flanker

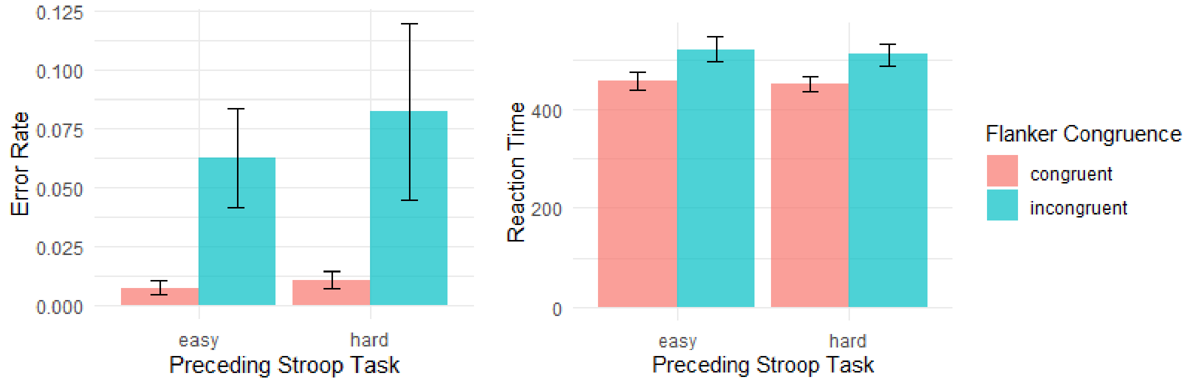
[5] At Stage 2 of the manuscript, we added "repeated measures" to clearly indicate the type of ANOVA used.

congruence, $F(1, 94) = 107.61$, $p < .001$, $\eta^2_p = .53$, indicating that participants required less time to respond in incongruent ($M = 515.80$, $SD = 118.66$) compared to congruent trials ($M = 454.13$, $SD = 78.01$). The main effect of preceding Stroop type (easy Stroop, hard Stroop), $F(1, 94) = 1.21$, $p = .274$, $\eta^2_p = .01$, and the interaction between previous Stroop type and flanker congruence were not significant, $F(1, 94) = 0.74$, $p = .393$, $\eta^2_p = .008$, suggesting that the difficulty of the previous Stroop task did not impact reaction times in the flanker task.

The results are 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).

Figure 5

Average Error Rate and Reaction Time in Congruent and Incongruent Flanker Trials after Completing the Easy and Hard Stroop



Note. Effects of the preceding Stroop type (easy Stroop, hard Stroop) on flanker task reaction time and error rate were not significant. The error bars represent the 95% confidence intervals.

Perceived Effort's Influence on Performance

We explored 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 investigated whether perceived effort due to task difficulty in the Stroop task predicted 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, provided 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 employed several linear mixed models (LMMs), using either the error rate or the reaction time in the flanker task as outcome variable. Perceived effort due to boredom and effort due to task difficulty were obtained by calculating the mean of these variables across the probes. Three distinct models were tested for each performance variable (error rate, reaction time). Detailed model outputs are provided in Appendix E. The first model assessed the effects of the preceding Stroop task type (easy Stroop, hard Stroop; fixed effect) and perceived effort due to task difficulty during the Stroop task (fixed effect), as well as the type of flanker trial (congruent, incongruent; fixed effect) on the reaction time in the flanker task, while including random intercepts for participants[6]. The second model differed from the first by incorporating boredom related effort as a fixed effect instead of task difficulty related effort. The third model included both perceived task difficulty and boredom related effort as fixed effects. AIC, BIC, and WAIC are reported. The model with the lowest BIC was considered the most optimal one. Results of linear mixed models as well as the AIC, BIC, and WAIC are represented in Table 2. The confidence intervals presented correspond to the non-standardized estimates.

All linear mixed models predicting error rate and reaction time revealed a significant main effect of congruence, with lower error rates in congruent trials compared to incongruent trials. Neither the preceding Stroop type nor the difficulty related effort experienced during

[6] Note that, although our preregistered protocol indicated that random slopes would be included, we did not include them in the analyses. Given the structure of the data, including them would not have been appropriate: the Stroop task and its associated effort ratings varied between flanker sessions but remained constant within sessions. Thus, there was no within-subject variation in these predictors that would permit the estimation of random slopes.

the Stroop task had a significant main effect in any of the three models predicting error rate. The main effect of boredom related effort, which was included in model two and three, reached significance in model two (γ_{03} = 0.01, β = 0.11, *t* = 2.12, *p* = 0.034, 95% CI [0.00, 0.01]), indicating a higher error rate in the flanker task with greater boredom related effort in the preceding Stroop. Although showing a similar trend, the main effect of boredom related effort did not reach significance in the third model (γ_{04} = 0.01, β = 0.10, *t* = 1.75, *p* = 0.081, 95% CI [-0.00, 0.01]).

Results of linear mixed models predicting reaction time revealed a significant main effect of the preceding Stroop type in model one (γ_{01} = -25.36, β = -0.24, *t* = -2.28, *p* = 0.023, 95% CI [-47.21, -3.51]), indicating faster reaction times in the flanker task after completing the hard Stroop task, but not in model two and three. The main effect of boredom, which was included in models two (γ_{03} = 9.73, β = 0.17, *t* = 4.14, $p < 0.001$, 95% CI [5.11, 14.34]) and three (γ_{04} = 9.31, *t* = 3.93, $p < 0.001$, 95% CI [4.65, 13.97]), was significant in both models, showing slower reaction times in the flanker task with greater boredom related effort in the Stroop.

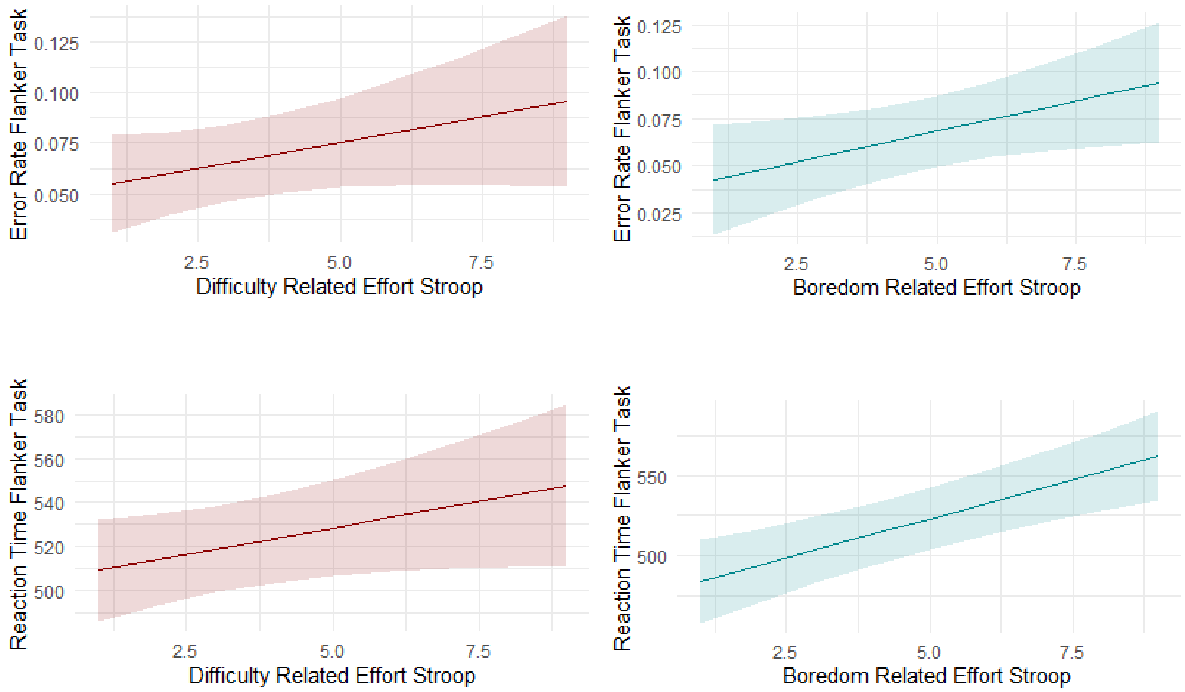
Model two, which included boredom related effort, but not difficulty related effort, emerged as the model with the lowest BIC for predicting both flanker task error rate and reaction time. This suggests that boredom related effort in the Stroop task plays a crucial role in subsequent flanker task performance, while difficulty related effort appears to be less relevant.

Two plots for each performance variable (error rate, reaction time) 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 represent the preceding Stroop type (easy Stroop, hard Stroop).

Figure 6

Effects of Difficulty Related Effort and Boredom Related Effort in the Stroop Task on Error Rate and Reaction

Time in the Flanker Task



Note. Analyses of linear mixed models revealed a significant main effect boredom related effort in the Stroop task on error rate and reaction time in the flanker task. The main effect of difficulty related effort was not significant when predicting error rate, neither in model one (p = .13, not including boredom related effort) nor in model three (p = .35, including boredom related effort). The confidence band indicates the 95% confidence interval.

Data Summary and Explorative Analysis

As preregistered, we calculated descriptive statistics, correlation analyses, and an explorative analysis concerning the number of sweets participants took with them after the experiment. Results are reported in Appendix F.

Discussion

This study examined the temporal dynamics of perceived effort (overall, boredom related, difficulty related), boredom, task difficulty, and fatigue during the performance of an easy congruent and a hard incongruent Stroop task with a task switching component (indicating the color vs. indicating the word). Participants' pupil sizes were measured during the task and served as an indicator for psychophysiological effort. It was tested if boredom and difficulty related effort predict tonic (baseline) and phasic (stimulus-evoked) pupil size during Stroop performance. Furthermore, this study investigated whether the Stroop task difficulty level as well as the perception of boredom and difficulty related effort in the Stroop task impacted performance in a secondary flanker task.

Thought Probes: Temporal Dynamics Differ by Task Type

The manipulation checks revealed that participants' error rates were lower, and reaction times were faster in the easy Stroop task. They reported greater overall boredom and perceived the easy Stroop as less difficult compared to the hard Stroop version. These findings confirm that our easy congruent Stroop version was indeed easier compared to our hard version.

Consistent with our predictions, ratings of perceived boredom, boredom related effort, difficulty, difficulty related effort, overall effort, and fatigue changed over time. Specifically, boredom, boredom related effort, and overall effort increased more strongly over time in the easy Stroop task. For perceived difficulty and difficulty related effort, ratings decreased more strongly over time in the hard Stroop task, whereas they did not decrease in the easy task. Given that boredom was particularly high in the easy Stroop, it is possible that the increased sensation of boredom contributed to a perception that, over time, the task was becoming more difficult or at least not easier. In contrast, no significant interaction between task and time emerged for fatigue. Instead, main effects indicated that participants reported greater fatigue over time in both tasks, and higher overall fatigue levels in the easy Stroop. Similarly, participants reported more boredom in the easy Stroop, and greater difficulty, difficulty related effort, and overall boredom in the hard Stroop.

These temporal dynamics align with results of previous research on fatigue and effort related to boredom and task difficulty (Radtke et al., 2025). The results suggest that regardless of whether a task is easy or difficult, it becomes more boring over time, likely increasing boredom related effort. This supports the idea that boredom prompts the exploration of alternative behaviors (Agrawal et al., 2022; Bench & Lench, 2013; Danckert, 2019; Geana et al., 2016), increasing the opportunity costs of continuing the current task (Bieleke & Wolff, 2021a; Bieleke et al., 2023). As a result, sustaining engagement in the boring task is presumably perceived as more effortful (Wolff & Martarelli, 2020). Meanwhile, even though task difficulty and difficulty related effort decrease over time in hard tasks, fatigue increases in both task conditions. Combined with the finding that overall fatigue is higher in the easy compared to the hard Stroop, this suggests that boredom and its associated effort may be even more fatiguing than task difficulty and difficulty related effort. This aligns with previous research reporting greater fatigue during and greater increases of fatigue following boring tasks (Mangin et al., 2021; Milyavskaya et al., 2019; Radtke et al., 2025), but contrasts with research suggesting greater fatigue after more cognitive demanding tasks (Kurzban et al., 2013; Marcora et al., 2009; O'Keeffe et al., 2019). While the temporal dynamics of fatigue mirror those of boredom related effort rather than those of difficulty related effort, participants' perception of the overall effort required follows the pattern of difficulty related effort in the hard task and boredom related effort in the easy task. Thus, whether a task is overall easy or hard seems to influence whether overall effort while performing a task increases or decreases over time. Another possible explanation is that participants in the hard condition were more engaged, as indicated by the pattern of overall effort, and may therefore not have perceived their exerted effort as fatiguing.

Overall, these results suggest that, over time, tasks, regardless of their difficulty level, become more boring and more fatiguing, while also the effort that has to be invested due to boredom increases. In contrast, the perceived difficulty, the effort that has to be invested due to the difficulty, and overall effort increase or decrease over time depending on the objective difficulty level of a task, highlighting the relevance of task characteristics in predicting how these variables change over time.

Effort and Pupil Size: Boredom and Difficulty Related Effort Reflect in Pupil Size

Preregistered comparisons between model 1, which predicted pupil size with task, time, and difficulty related effort, and model 3, which additionally included boredom related effort, did not provide consistent support for our hypothesis, as AIC, BIC, and WAIC partly favored model 1 and partly model 3. When extending the model comparisons to all models, including the exploratory model 3ex, which added interactions between task and the effort variables, model 3ex was preferred over all other models. This provided post-hoc support that tonic and phasic pupil size were predicted more accurately by task difficulty related effort and boredom related effort together than by task difficulty related effort alone. As pupil size serves as objective measure and physiological correlate of effort (van der Wel & van Steenbergen, 2018), this result emphasizes the importance of recognizing boredom as a potential confounding variable in research investigating effort and self-control. However, results from our linear mixed models revealed that the relationship between the level of task difficulty (easy, hard), effort related to boredom, effort related to difficulty, and pupil size is complex. Consistent with previous research, pupil size was larger in the hard Stroop and decreased over time in both tasks. This likely reflects the greater cognitive load and overall effort required in the hard task (van der Wel & van Steenbergen, 2018) as well as an adaptation and a reduction in task demands over time (Hopstaken et al., 2015; Timme et al., 2022).

Regarding the effect of boredom related effort, our findings highlight distinct patterns in tonic (baseline) and phasic (stimulus-evoked) pupillary responses. Tonic pupil size, reflecting the general state of an individual (Cohen Hoffing et al., 2020), increased with greater boredom related effort in the easy task. This suggests that boredom related effort during easy tasks is not only perceived as effortful but also has physiological correlates. Similar, difficulty related effort in the easy Stroop was related to larger tonic pupil sizes. This indicates that both boredom-related and difficulty-related effort contribute to an elevated baseline cognitive load, even in the absence of external stimuli.

However, these effects were not observed in the hard Stroop task. One possible explanation is that the high cognitive demands of the task, such as remembering four response keys (one for each color) and two different symbols (+ and x) indicating the task for the trial (color naming vs. word reading), resulted in overwhelm and consequently in reduced attention. This could have diminished the effect of boredom and difficulty related effort on tonic pupil size. Supporting this idea, smaller tonic pupil sizes have been associated with lapses in attention (Kristjansson et al., 2009; van den Brink et al., 2016). Another explanation is that the task may have been too difficult, exceeding the optimal challenge level for inducing flow. Supporting this idea, recent research has shown that flow during a cognitive task follows an inverted U-shaped relationship with pupil dilation (Lu et al., 2023).

Phasic pupil size, which reflects the reactions to the stimulus (Cohen Hoffing et al., 2020), showed a different pattern. Contrary to our expectations, phasic pupil size decreased with greater boredom related effort in both the easy and hard Stroop. This suggests that as boredom related effort increased, participants' stimulus-driven attention and engagement declined. Prior studies have reported similar associations between boredom and reduced on-task attention (Bench & Lench, 2013; Eastwood et al., 2012; Westgate & Wilson, 2018) as well as between attention and phasic pupil size (Unsworth & Miller, 2021; Unsworth & Robison, 2017).

In contrast to boredom related effort, increases in difficulty related effort were associated with larger phasic pupil sizes in the easy Stroop task, indicating that perceived effort related to difficulty was physiologically reflected in stronger pupillary responses to the stimulus. However, this effect was not observed in the hard Stroop task. One possible explanation is that the cognitive load in the hard Stroop was already so high that any additional effort (due to boredom or task difficulty), did not lead to increases in stimulus-driven responses, but rather to task disengagement and reduced attention, a state that has previously been linked to decreases in phasic pupil size (Unsworth & Miller, 2021; Unsworth & Robison, 2017).

These findings contribute to an emerging body of literature suggesting that enduring boredom adds additional requirements of self-control and may confound related research outcomes (Francis et al., 2018; Job et al., 2010; Mangin & Pageaux, 2024; Milyavskaya et al., 2019; Wolff & Martarelli, 2020). Our study extends this literature by providing preliminary evidence, that boredom related effort has a physiological impact on tonic and phasic pupil responses. This result complements recent findings linking boredom related effort to electrodermal activity (Radtke et al., 2025). Taken together, our results suggest that enduring boredom in cognitive tasks may not only lead to the perception of effort related to boredom but also be reflected in psychophysiological changes.

Task Performance in Secondary Task: No Ego Depletion Effect

As effects of investing effort and self-control in one task on performance in a secondary task are widely discussed (Carter & McCullough, 2014; Dang, 2018; Giboin & Wolff, 2019; Hagger et al., 2016; Hagger et al., 2010; Holgado & Mesquida, 2023; Wolff et al., 2018), we included the flanker task as secondary task in the present study and measured performance during this task. Repeated measures ANOVAs revealed no significant differences in flanker task performance (error rate, reaction time) after the easy and hard Stroop tasks. This finding aligns with prior research failing to support the ego depletion effect or mental fatigue (Holgado & Mesquida, 2023; Wolff et al., 2018) and indicate that differences in the difficulty of a preceding cognitive task do not necessarily affect performance on a subsequent cognitive task. According to ego depletion and mental fatigue theories, engaging in a cognitively demanding task should impair subsequent performance in another task requiring self-control or executive function (Baumeister et al., 1998; Kurzban et al., 2013). However, it has been suggested that these effects frequently remain undetected because both tasks require self-control and effort, just for different reasons (Wolff & Martarelli, 2020). The easy Stroop task may require more self-control due to boredom related effort, whereas the hard Stroop task requires more self-control due to task difficulty related effort.

To examine these effects more directly, we conducted linear mixed models to test whether effort related to task difficulty and effort related to boredom in the primary Stroop task predicted performance in the secondary flanker task. Our results indicate that effort related to boredom predicted higher error rates and longer reaction times in the flanker task, whereas the main effect of effort related to task difficulty was not significant. Model comparisons using BIC indicated that the best-fitting model included boredom related effort but not task difficulty related effort (in addition to Stroop task difficulty and flanker congruence). Thus, boredom related effort appears to have the stronger impact on performance.

As the increase in boredom related effort was greater in the easy Stroop compared to the hard Stroop, and effort due to difficulty was substantially higher in the hard Stroop (M = 6.66, compared to M = 3.02 in the easy Stroop), the weaker role of difficulty related effort in predicting secondary task performance may explain why no significant effect of Stroop task difficulty was observed.

These results support the notion that, although differences in secondary task performance after completing easy versus difficult tasks are often not observable, this does not imply that the allocation of effort in the secondary task is not affected by the first task (Wolff & Martarelli, 2020). However, the most relevant factor influencing secondary task performance seems not necessarily to be the difficulty level of a task, but rather other variables, such as boredom, that further increase the effort an individual has to invest.

Besides boredom related effort influencing mental fatigue or the depletion of self-control by being effortful (Baumeister, 2018; Kurzban et al., 2013; Marcora et al., 2009), other mechanisms are possibly contributing to the negative impact of boredom related effort on performance. It is possible that boredom related effort in the Stroop task reduces motivation, thereby increasing the perceived amount of effort that has to be invested in the flanker task. Consequently, the reward value of the flanker task might not justify the effort that has to be invested to perform well (inspired by the motivational intensity theory; Brehm & Self, 1989). Moreover, from the perspective of the opportunity cost model of effort (Kurzban et al., 2013), boredom may signal that more rewarding alternatives are available, leading to a rise in perceived opportunity costs. This subjective cost may render continued cognitive engagement increasingly aversive, possibly reflected in the increase of boredom related effort. The fact that boredom related effort in the Stroop task negatively predicted flanker task performance, while objective task difficulty, subjective difficulty, and difficulty related effort did not, suggests that subjective opportunity costs, such as those triggered by boredom, might be more relevant than objective task demands in driving performance decrements across sequential tasks. While the opportunity cost model appears to offer the most straightforward explanation for our findings, other theories may also contribute to an explanation for our results, especially when extended to incorporate the role of boredom related effort, as discussed above.

Additionally, the prolonged endurance of boredom (and the resulting effort invested), may increase stress levels (Chao et al., 2020; Parasuraman & Purohit, 2000) and negative emotions such as frustration (Wolff et al., 2024). These emotional responses, in turn, can reduce attentional control and increase cognitive load, as off-task thoughts related to negative emotions begin to compete for working memory resources (Plass & Kalyuga, 2019). Moreover, as boredom is often accompanied by mind-wandering (Zanesco et al., 2024), likely as a coping mechanism, this shift in attentional focus could further impair task performance (Randall et al., 2014).

Overall, the findings of the present study highlight the importance of considering boredom related effort in self-control and cognitive performance research and, more general, within the broader scope of study designs. However, future research is necessary to improve the understanding of the underlying mechanisms and address several limitations of the present study.

Limitations and Future Research

It is important to acknowledge limitations regarding the measurement of certain constructs in this study. Specifically, despite its successful implementation (i.e., participants did not raise questions about the items and ratings changed over time following different trajectories) in the present and in a related study (Radtke et al., 2025), the validity of items used to assess perceptions of task difficulty, boredom and related effort need further validation. The self-report items used to assess these variables consisted of single-item measures and have not undergone formal psychometric validation. Therefore, the construct validity of these measures cannot be fully established based on the present data. Especially since the temporal patterns did not fully evolve as predicted (i.e., no decrease over time in perceived difficulty and difficulty related effort in the easy Stroop) and, contrary to our hypotheses, greater difficulty related effort was associated with neither larger phasic nor tonic pupil sizes. Nevertheless, in the present data and in Radtke et al. (2025), these measures demonstrated sensitivity to within-task changes. Additionally, although the variance inflation factors (VIFs) in the linear mixed models remained below 5 for these variables, their mean scores (averaged across all blocks) were correlated. Correlational evidence (Appendix F, Table 5) provides some support for the distinction between difficulty-related effort and boredom-related effort. Their association was modest in the easy Stroop ($r = .26$, $p < .05$) but stronger in the hard Stroop ($r = .69$, $p < .01$), suggesting that the overlap between the two constructs increases under higher task demands. Moreover, difficulty-related effort was more strongly related to task difficulty and perceived effort, whereas boredom-related effort showed stronger associations with boredom and fatigue. Together, these patterns indicate that the two constructs share variance but also reflect separate processes. Future research is necessary to examine whether difficulty and boredom related effort are accurately captured by our variables and whether they demonstrate adequate construct validity. This could be addressed, for example, by evaluating these ratings and their trajectories in relation to other variables, examining them across different tasks, and incorporating additional physiological or behavioral measures (e.g., task persistence).

Different variants of the Stroop task exist, each with distinct neural and behavioral impacts (Müller et al., 2024). The results reported here may be specific to the methodological choices we made. In particular, a constraint lies in the task switching component of the hard Stroop version, which combined multiple cognitive demands, namely, Stroop interference and task-switching between color-naming and word-naming. On the one hand, this may result in varying levels of required cognitive control throughout the task. On the other hand, task difficulty and difficulty related effort are impacted by a composite of cognitive processes rather than purely by Stroop interference. 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. This variability in cognitive control could have contributed to the non-significant relationship between difficulty related effort and pupil size in the present study. 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. Moreover, isolating distinct cognitive demands in different tasks would clarify the distinct contributions of each cognitive process to effort related to task difficulty and physiological responses, improving the interpretability of findings.

Furthermore, although our results suggest carryover effects between primary and secondary tasks induced by boredom related effort, the exact mechanisms by which boredom related effort influences subsequent performance remain unclear. Future research should investigate underlying mechanisms that lead to an increase in boredom related effort, and mechanisms by which boredom related effort ultimately influences task performance. In summary, the present study demonstrates that while task difficulty is an important factor in determining cognitive effort, boredom can not only trigger additional cognitive effort but even affect performance in subsequent tasks. Increases in boredom related effort are not only subjectively perceived in easy and hard tasks but are also reflected in pupil size as a psychophysiological marker of effort during easy tasks. This highlights the importance of considering boredom as a dimension of effort in research on self-control and cognitive performance.

Constraints on Generality

The present findings should be interpreted with several constraints on generality (Simons et al., 2017). Our sample was WEIRD (Western, Educated, Industrialized, Rich and Democratic; Henrich et al., 2010), relatively young (M = 20.22 years, SD = 4.48, range = 16 - 45), predominantly female (75.79%), and consisted of Swiss and German citizens recruited mainly from the general population, including school students, university students, working adults, and unemployed individuals. Although this provides a somewhat broader range than typical university-only samples, caution is warranted when generalizing to older adults, more balanced gender distributions, or other cultural contexts. Especially older adults might experience effort and boredom in cognitive tasks differently due to age-related changes in cognitive control and motivational processes (Yee et al., 2019). Furthermore, the cognitive tasks employed were an easy (congruent) and hard (incongruent, with task-switching component) Stroop variant that consisted of 360 trials each. Therefore, the observed relationships may not generalize to other cognitive control tasks, Stroop versions with different congruency structures, or longer tasks. Additionally, the experiment was completed in the laboratory. Real-world environments may involve more complex dynamics, especially regarding boredom related effort, as the real world offers greater opportunities for distraction than the laboratory. These constraints should be kept in mind when interpreting the findings.

Beyond the factors described above, we have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

Declarations

CRediT authorship contribution statement

Vanessa C. Radtke: Writing – original draft, Methodology, Investigation, Data curation, Visualization, Formal analysis, Project administration, Conceptualization. **Wanja Wolff:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization, Funding acquisition. **Corinna S. Martarelli:** Writing – review & editing, Supervision, Methodology, Project administration, Conceptualization, Funding acquisition.

Acknowledgments

We thank Ilaria Raich for her support during data collection.

Ethics Approval

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of UniDistance Suisse (Ethics approval number: 2022-10-00005).

Informed Consent

Informed consent was obtained from all individual participants included in the study. Parental consent was not required for participants under 18 years of age (i.e., 16 or 17 years old).

Conflict of Interest

The authors of this article declare that they have no financial conflict of interest with the content of this article.

Funding

This work was supported by the Swiss National Science Foundation [grant number 10001CL_200468] and the German Research Foundation [grant number 460704829].

Artificial Intelligence Statement

The authors used ChatGPT to refine the wording in selected sections of the manuscript.

Data Availability

Data and scripts are shared on Open Science Framework (OSF) at https://osf.io/jn54q/.

References

Agrawal, M., Mattar, M. G., Cohen, J. D., & Daw, N. D. (2022). The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. *Psychological Review*, *129*(3), 564–585. https://doi.org/10.1037/rev0000309

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. (Ed.). (2018). *World library of psychologists*. *Self-regulation and self-control: Selected works of Roy Baumeister*. Routledge. https://doi.org/10.4324/9781315175775

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

Bench, S. W., & Lench, H. C. (2013). On the function of boredom. *Behavioral Sciences*, *3*(3), 459–472. https://doi.org/10.3390/bs3030459

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. (2021a). It's not a bug, it's boredom: Effortful willpower balances exploitation and exploration. *The Behavioral and Brain Sciences*, *44*, e33. https://doi.org/10.1017/S0140525X20001053

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

Bieleke, M., Wolff, W., & Bertrams, A. (2023). *On the virtues of fragile self-control: Boredom as a catalyst for adaptive behavior regulation*. PsyArXiv. https://doi.org/10.31234/osf.io/tgq95

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

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*, 823. https://doi.org/10.3389/fpsyg.2014.00823

Chao, M., Chen, X., Liu, T., Yang, H., & Hall, B. J. (2020). Psychological distress and state boredom during the COVID-19 outbreak in China: The role of meaning in life and media use. *European Journal of Psychotraumatology*, *11*(1), 1769379. https://doi.org/10.1080/20008198.2020.1769379

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. (2019). Boredom: Managing the delicate balance between exploration and exploitation. In J. Ros Velasco (Ed.), *Springer eBook Collection. Boredom Is in Your Mind: A Shared Psychological-Philosophical Approach* (1st ed. 2019, pp. 37–53). Springer. https://doi.org/10.1007/978-3-030-26395-9_3

Danckert, J., & Eastwood, J. D. (2020). *Out of my skull: The psychology of boredom*. Harvard University Press.

Dang, J. (2018). An updated meta-analysis of the ego depletion effect. *Psychological Research*, *82*(4), 645–651. https://doi.org/10.1007/s00426-017-0862-x

Dang, J., Barker, P., Baumert, A., Bentvelzen, M., Berkman, E., Buchholz, N., Buczny, J., Chen, Z., Cristofaro, V. de, Vries, L. de, Dewitte, S., Giacomantonio, M., Gong, R., Homan, M., Imhoff, R., Ismail, I., Jia, L., Kubiak, T., Lange, F., . . . Zinkernagel, A. (2021). A multilab replication of the ego depletion effect. *Social Psychological and Personality Science*, *12*(1), 14–24. https://doi.org/10.1177/1948550619887702

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

Eastwood, J. D., Frischen, A., Fenske, M. J., & Smilek, D. (2012). The unengaged mind: Defining boredom in terms of attention. *Perspectives on Psychological Science*, *7*(5), 482–495. https://doi.org/10.1177/1745691612456044

Elpidorou, A. (2014). The bright side of boredom. *Frontiers in Psychology*, *5*, 1245. https://doi.org/10.3389/fpsyg.2014.01245

Englert, C., & Bertrams, A. (2021). Again, No evidence for or against the existence of ego depletion: Opinion on "A multi-site preregistered paradigmatic test of the ego depletion effect". *Frontiers in Human Neuroscience*, *15*, 658890. https://doi.org/10.3389/fnhum.2021.658890

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

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

Francis, Z., Milyavskaya, M., Lin, H., & Inzlicht, M. (2018). Development of a within-subject, repeated-measures ego-depletion paradigm. *Social Psychology*, *49*(5), 271–286. https://doi.org/10.1027/1864-9335/a000348

Geana, A., Wilson, R. C., Daw, N., & Cohen, J. D. (2016). *Boredom, information-seeking and exploration*. https://www.princeton.edu/~ndaw/gwdc16.pdf

Giboin, L.-S., & Wolff, W. (2019). The effect of ego depletion or mental fatigue on subsequent physical endurance performance: A meta-analysis. *Performance Enhancement & Health*, *7*(1-2), 100150. https://doi.org/10.1016/j.peh.2019.100150

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., Ridder, D. T. D. de, Dewitte, S., . . . Zwienenberg, M. (2016). A Multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, *136*(4), 495–525. https://doi.org/10.1037/a0019486

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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). *The weirdest people in the world?* Behavioral and Brain Sciences 33(2-3), 61–83. https://10.1017/s0140525x0999152x

Hofmann, W., Baumeister, R. F., Förster, G., & Vohs, K. D. (2012). Everyday temptations: An experience sampling study of desire, conflict, and self-control. *Journal of Personality and Social Psychology*, *102*(6), 1318–1335. https://doi.org/10.1037/a0026545

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

Holgado, D., & Mesquida, C. (2023). *Assessing the evidential value of mental fatigue and exercise research.* https://doi.org/10.51224/SRXIV.245

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

Hunte, R., Cooper, S. B., Taylor, I. M., Nevill, M. E., & Boat, R. (2022). Boredom, motivation, and perceptions of pain: Mechanisms to explain the effects of self-control exertion on subsequent physical performance. *Psychology of Sport and Exercise*, *63*, 102265. https://doi.org/10.1016/j.psychsport.2022.102265

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

Job, V., Dweck, C. S., & Walton, G. M. (2010). Ego depletion—Is it all in your head? *Psychological Science*, *21*(11), 1686–1693. https://doi.org/10.1177/0956797610384745

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

Kool, W., McGuire, J. T., Wang, G. J., & Botvinick, M. M. (2013). Neural and behavioral evidence for an intrinsic cost of self-control. *PLoS ONE*, *8*(8), e72626. https://doi.org/10.1371/journal.pone.0072626

Kristjansson, S. D., Stern, J. A., Brown, T. B., & Rohrbaugh, J. W. (2009). Detecting phasic lapses in alertness using pupillometric measures. *Applied Ergonomics*, *40*(6), 978–986. https://doi.org/10.1016/j.apergo.2009.04.007

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

Lu, H., van der Linden, D. & Bakker, A.B. (2023). Changes in pupil dilation and P300 amplitude indicate the possible involvement of the locus coeruleus-norepinephrine (LC-NE) system in psychological flow. *Scientific Reports*, *13*, 1908. https://doi.org/10.1038/s41598-023-28781-z

Mangin, T., André, N., Benraiss, A., Pageaux, B., & Audiffren, M. (2021). No ego-depletion effect without a good control task. *Psychology of Sport and Exercise*, *57*, 102033. https://doi.org/10.1016/j.psychsport.2021.102033

Mangin, T., & Pageaux, B. (2024). It is time to stop using the terminology “passive” fatigue. *Motivation Science.* Advance online publication. https://doi.org/10.1037/mot0000375

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

Meier, M., Martarelli, C. S., & Wolff, W. (2023). Bored participants, biased data? How boredom can influence behavioral science research and what we can do about it. *PsyArXiv.* Advance online publication. https://doi.org/10.31234/osf.io/hzfqr

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

Milyavskaya, M., Inzlicht, M., Johnson, T., & Larson, M. J. (2019). Reward sensitivity following boredom and cognitive effort: A high-powered neurophysiological investigation. *Neuropsychologia*, *123*, 159–168.

https://doi.org/10.1016/j.neuropsychologia.2018.03.033
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
O'Keeffe, K., Hodder, S., & Lloyd, A. (2019). A comparison of methods used for inducing mental fatigue in performance research: individualised, dual-task and short duration cognitive tests are most effective. *Ergonomics*, *63*(1), 1–12. https://doi.org/10.1080/00140139.2019.1687940
Parasuraman, S., & Purohit, Y. S. (2000). Distress and boredom among orchestra musicians: The two faces of stress. *Journal of Occupational Health Psychology*, *5*(1), 74–83. https://doi.org/10.1037//1076-8998.5.1.74
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
Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review*, *31*(2), 339–359. https://doi.org/10.1007/s10648-019-09473-5
R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/
Radtke, V., Martarelli, C., & Wolff, W. (2025). *Effort and its reasons - Electrodermal activity as indicator of boredom vs. difficulty related effort in cognitive-physical sequential tasks*. PsyArXiv. https://doi.org/10.31219/osf.io/2t6p8
Rajkowski, J., Lu, W., Zhu, Y., Cohen, J., & Aston-Jones, G. (2000). Prominent projections from the anterior cingulate cortex to the locus coeruleus in Rhesus monkey. *Society for Neuroscience Abstract*, *26*, 838–15.
Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin*, *140*(6), 1411–1431. https://doi.org/10.1037/a0037428
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
Schwarz, W., & Ischebeck, A. (2003). On the relative speed account of number-size interference in comparative judgments of numerals. *Journal of Experimental Psychology*, *29*(3), 507–522. https://doi.org/10.1037/0096-1523.29.3.507
Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123-1128. https://doi.org/10.1177/1745691617708630
Solomon, S., Holmes, D. S., & McCaul, K. D. (1980). Behavior control over aversive events: Does control that requires effort reduce anxiety and physiological arousal? *Journal of Personality and Social Psychology*, *39*(4), 729–736. https://doi.org/10.1037/0022-3514.39.4.729

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

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

Tzelgov, J., Meyer, J., & Henik, A. (1992). Automatic and intentional processing of numerical information. *Journal of Experimental Psychology*, *18*(1), 166–179. https://doi.org/10.1037/0278-7393.18.1.166

Unsworth, N., & Miller, A. L. (2021). Individual differences in the intensity and consistency of attention. *Current Directions in Psychological Science*, *30*(5), 391–400. https://doi.org/10.1177/09637214211030266

Unsworth, N., & Robison, M. K. (2017). The importance of arousal for variation in working memory capacity and attention control: A latent variable pupillometry study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1962–1987. https://doi.org/10.1037/xlm0000421

van den Brink, R. L., Murphy, P. R., & Nieuwenhuis, S. (2016). Pupil diameter tracks lapses of attention. *PLOS ONE*, *11*(10), e0165274. https://doi.org/10.1371/journal.pone.0165274

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

Vassena, E., Silvetti, M., Boehler, C. N., Achten, E., Fias, W., & Verguts, T. (2014). Overlapping neural systems represent cognitive effort and reward anticipation. *PLoS ONE*, *9*(3), e91008. https://doi.org/10.1371/journal.pone.0091008

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., Baumann, L., & Englert, C. (2018). Self-reports from behind the scenes: Questionable research practices and rates of replication in ego depletion research. *PLoS ONE*, *13*(6), e0199554. https://doi.org/10.1371/journal.pone.0199554

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

Yee, D.M., Adams, S., Beck, A., & Braver, T. S. (2019) Age-Related Differences in Motivational Integration and Cognitive Control. *Cognitive Affective Behavioral Neuroscience*, *19*, 692–714. https://doi.org/10.3758/s13415-019-00713-3

Zanesco, A. P., Denkova, E., & Jha, A. P. (2024). Mind-wandering increases in frequency over time during task performance: An individual-participant meta-analytic review. *Psychological Bulletin.* Advance online publication. https://doi.org/10.1037/bul0000424

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

Appendix A

Online Study

To ensure a significant divergence in the difficulty level and the potential for boredom induction by our tasks, we conducted an online study consisting of four Stroop task versions. These versions included both easy and hard versions of the color Stroop and the numerical Stroop. Our primary objective was to assess the participant's perception of task difficulty and boredom across the four tasks, thereby enabling us to select the most suitable Stroop task (color or numerical Stroop) for our present study and determining the appropriate sample size based on the effect size observed in our online study. The R code to reproduce analyses as well as the data set is available on OSF https://osf.io/w9phy/.

Methods

Data collection for all four sessions of our within-subjects design study was done online via CloudResearch's Connect platform. Participants were assigned to one out of two participation orders (order one: easy color, easy numerical, hard color, hard numerical; order two: hard color, hard numerical, easy color, easy numerical). The interval between sessions varied from one to seven days. After completing the tasks for ten minutes each, participants were asked to indicate how much they agree or disagree with statements about their experience during the task (boredom and perceived task difficulty).

Participants

Starting with an initial cohort of 328 participants in session one, the study involved 210 participants (42.38 % female, 54.28 % male, 3.33 % other genders or no answer) who completed all four sessions of the study seriously (z-transformed error rate in the tasks less than three standard deviations away from the mean of all participants, attention checks answered correctly) and answered the relevant questions. The mean age of participants was 32.20 (SD = 5.31). Respondents were compensated with a fixed amount of $3.50 for their participation in each session. They had the opportunity to earn an additional variable compensation of up to $1 in two of the sessions. All participants gave informed consent before starting the experiment. The study was approved by the local Ethics Committee and carried out in accordance with the Helsinki Declaration.

Tasks

Stroop tasks. The color Stroop versions closely resemble those suggested in the present study. However, there are some distinctions: in the online study, a task reminder remained on screen, and the task was indicated not only by the "+" and "x" on screen during stimulus presentation but also with a fixation cross before stimulus onset which was either a "+" or an "x". To maintain participant's gaze at the center of the screen, we omitted the task reminder for the present study. Further, we showed the same fixation stimuli (###) in both tasks to keep the physiological influence on baseline (tonic) pupil size consistent across both task versions. To improve the quality of the pupillary data, we presented the stimuli for a fixed duration of 400 ms, instead of presenting them until participants react as we did in the online study. Lastly, while the hard Stroop task in the online study consisted of 50% incongruent trials, the hard Stroop task in the present study consisted of 100% incongruent trials. This change is implemented with the aim to (1) increase the overall difficulty of the hard task and (2) enhance the clarity of our analysis, particularly regarding pupillary reactions, as pupil size is greater in incongruent than in congruent trials. Otherwise, by calculating the mean of the five trials preceding the probe, congruent and incongruent trials would get mixed up.

In the numerical Stroop version (Schwarz & Ischebeck, 2003; Tzelgov et al., 1992) each trial consisted of a fixation cross (displayed for 1000 on screen), two presented numbers (one with a larger numerical value and one in a larger font size), and an intertrial interval lasting 500 ms. The numbers remained on screen until participants pressed one of two keys on the keyboard ("1" for left, "0" for right) to indicate their response. In the easy congruent-only version one number was both larger in value and in font size. Participants were instructed to identify the number with the larger font size. In the hard version, the numbers were presented either congruent or incongruent (50 % of the trials). In incongruent trials one number was larger in font size while the other was larger in value. The instructions switched during the task. In 80 % of the trials, participants were instructed to indicate the number with the lager font size, and in 20 % of the trials, they were instructed to indicate the number with the larger numerical value. A task reminder was displayed on screen. The task was further indicated by an “x” (indicate which number has the larger numerical value) or an “+” (indicate which number has a larger font size).

Subjective experience. After completing each task participants were asked to indicate how much they agree or disagree with statements about their experience during the task on a 9-point Likert scale (*1* = strongly disagree, *9* = strongly agree). Among these statements were one on boredom (“I was bored while performing the task”) and one on task difficulty (“the task was difficult for me”).

Supplementary measures. In addition to the tasks and questions described, we collected additional variables that are not pertinent to the present study.

Results

Results of a two-way repeated measures ANOVA which was performed to analyze the effect of task variant (color Stroop, numerical Stroop) and difficulty level (easy, hard) on boredom revealed a main effect of task variant ($F(1, 204) = 4.90$, $p = .028$, partial $\eta^2 = 0.002$), indicating that participants perceived the numerical Stroop ($M = 5.67$, $SD = 2.78$) as more boring than the color Stroop ($M = 5.42$, $SD = 2.78$). Additionally, a main effect of difficulty level was observed ($F(1, 204) = 35.51$, $p < .001$, partial $\eta^2 = 0.017$), highlighting that the easy tasks ($M = 5.90$, $SD = 2.72$) were more boring than the difficult tasks ($M = 5.19$, $SD = 2.74$). There was no significant interaction between the effect of task variant and level of difficulty ($F(1, 204) = 2.98$, $p = .086$, partial $\eta^2 = 0.001$).

Another two-way repeated measures ANOVA analyzing the effect of task variant (color Stroop, numerical Stroop) and difficulty level (easy, hard) on perceived task difficulty indicated a main effect of task version ($F(1, 204) = 96.46$, $p < .001$, partial $\eta^2 = 0.060$), with a higher perceived task difficulty in the color Stroop ($M = 3.75$, $SD = 2.58$) than in the numerical Stroop ($M = 2.68$, $SD = 2.03$). The ANOVA further revealed a main effect of difficulty level ($F(1, 204) = 170.31$, $p < .001$, partial $\eta^2 = 0.116$), with a higher perceived task difficulty in the hard ($M = 3.99$, $SD = 2.52$) than in the easy ($M = 2.44$, $SD = 1.94$) tasks. A significant interaction effect was found between task variant and difficulty level ($F(1, 204) = 76.80$, $p < .001$, partial $\eta^2 = 0.043$). This interaction suggests that the impact of difficulty level (easy, hard) on perceived task difficulty depends on the task variant (color, numerical). Specifically, in the color Stroop task, participants perceived the task to be more difficult in the hard ($M = 4.98$, $SD = 2.55$) compared to the easy condition ($M = 2.53$, $SD = 1.95$). In contrast, in the numerical Stroop task, the difference in perceived difficulty between the hard ($M = 3.00$, $SD = 2.07$) and easy ($M = 2.36$, $SD = 1.93$) conditions was less pronounced.

As the focus of the present study lies in the difference between the easy and hard versions of the color Stroop (see *Conclusion* for the rationale), we further conducted two t-tests comparing the perceived task difficulty and boredom between these two tasks. Results show that perceived task difficulty is greater in the hard version ($M = 4.98$, $SD = 2.54$) compared to the easy version of color Stroop task ($M = 2.54$, $SD = 1.97$), $t(209) = 13.34$, $p < .01$, $d = 0.92$. Boredom on the other hand is less pronounced in the hard ($M = 5.01$, $SD = 2.79$) than in the easy color Stroop ($M = 5.90$, $SD = 2.67$), $t(209) = 4.91$, $p < .01$, $d = 0.34$.

Conclusion

In both versions of the Stroop task (color, numerical) participants found the hard versions to be more difficult and less boring than the easy counterparts. Overall, the color Stroop was perceived as more difficult and less boring than the numerical Stroop. The difference in perceived task difficulty between the two difficulty levels was more pronounced in the color than in the numerical Stroop variant. Consequently, the color Stroop emerges as the more suitable choice for the present study. Thus, we focused our power analysis on the difference in boredom between the hard and the easy color Stroop ($d = 0.34$) to adopt a conservative approach in calculating the required sample size.

Appendix B

Pre-test

Using the described experimental procedure, we tested two participants. Descriptive analyses revealed for the easy Stroop version a mean in boredom of 5.14 (SD = 1.04), in boredom related effort 5.59 (SD = 1.06), in task difficulty 2.77 (SD = 1.31), in task difficulty related effort 1.86 (SD = 0.37), in fatigue 6.41 (SD = 0.48), and in perceived effort 5.64 (SD = 0.60). In the hard Stroop version, we calculated a mean in boredom of 4.02 (SD = 0.59), in boredom related effort 3.68 (SD = 1.06), in task difficulty 4.76 (SD = 1.16), in task difficulty related effort 4.82 (SD = 1.20), in fatigue 5.27 (SD = 1.41), and in perceived effort 5.95 (SD = 0.57). Across all participants and probes, the mean baseline pupil size during the easy Stroop task was 918.38 (SD = 47.78), while the peak stimulus-evoked baseline corrected pupil size was 85.92 (SD = 12.30). In the case of the hard Stroop task, the average baseline pupil size was 788.07 (SD = 108.32), and the peak stimulus-evoked baseline corrected pupil size was 124.84 (SD = 23.96). The averages of thought probes are presented in table 1, the average tonic and phasic pupil size (averaged across the five trials preceding the thought probe) are shown in table 2.

Table 1

Means and Standard Deviations of Variables Corresponding to Each Thought Probe in the Easy und Hard Stroop Version

Variable	Difficulty Level		1	2	3	4	5	6	7	8	9	10	11
BrE	Easy	*M*	4.50	5.50	5.00	5.50	5.50	5.50	6.00	6.00	6.00	6.00	6.00
		SD	(0.71)	(0.71)	(1.41)	(0.71)	(0.71)	(0.71)	(0.00)	(1.41)	(1.41)	(1.41)	(1.41)
	Hard	*M*	2.00	2.50	2.50	2.50	3.50	4.00	4.00	5.00	5.00	4.50	5.00
		SD	(0.00)	(0.71)	(0.71)	(0.71)	(0.71)	(1.41)	(1.41)	(1.41)	(1.41)	(0.71)	(1.41)
DrE	Easy	*M*	2.00	2.00	2.00	2.00	2.00	2.00	2.00	1.50	2.00	1.50	1.50
		SD	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.71)	(0.00)	(0.71)	(0.71)
	Hard	*M*	5.00	5.00	5.00	5.00	4.50	4.00	6.00	5.00	4.50	4.50	4.50
		SD	(1.41)	(0.00)	(0.00)	(1.41)	(2.12)	(1.41)	(0.00)	(0.00)	(0.71)	(0.71)	(0.71)
Effort	Easy	*M*	5.00	5.50	5.00	5.50	5.50	5.50	6.00	6.00	6.00	6.00	6.00
		SD	(0.00)	(0.71)	(0.00)	(0.71)	(0.71)	(0.71)	(0.00)	(1.41)	(0.00)	(0.00)	(0.00)
	Hard	*M*	5.50	5.00	5.00	4.50	5.50	6.00	6.00	6.50	6.50	7.50	7.50
		SD	(0.71)	(0.00)	(0.00)	(0.71)	(0.71)	(0.00)	(0.00)	(0.71)	(0.71)	(0.71)	(0.71)
Fatigue	Easy	*M*	5.50	6.50	6.00	6.00	6.00	6.50	6.50	7.00	7.00	6.50	7.00
		SD	(0.71)	(0.71)	(0.00)	(0.00)	(0.00)	(0.71)	(0.71)	(0.00)	(0.00)	(0.71)	(0.00)
	Hard	*M*	4.0	4.00	4.00	4.00	2.50	5.00	5.50	6.50	7.50	7.50	7.50
		SD	(1.41)	(1.41)	(1.41)	(1.41)	(2.12)	(1.41)	(2.12)	(0.71)	(0.71)	(0.71)	(0.71)
Difficulty	Easy	*M*	3.50	2.5	2.00	2.50	3.00	2.50	3.50	3.00	2.50	3.00	2.50
		SD	(2.12)	(0.71)	(0.00)	(0.71)	(1.41)	(0.71)	(2.12)	(1.41)	(0.71)	(2.83)	(0.71)
	Hard	*M*	5.00	5.00	5.00	4.50	5.00	4.50	5.50	5.00	4.33	4.50	4.50
		SD	(1.41)	(0.00)	(0.00)	(0.71)	(0.00)	(0.71)	(0.71)	(0.00)	(3.21)	(0.71)	(0.71)
Boredom	Easy	*M*	4.00	5.00	5.50	5.50	6.00	6.00	6.00	6.50	6.50	6.00	6.00
		SD	(0.00)	(0.00)	(0.71)	(0.71)	(1.41)	(1.41)	(1.41)	(0.71)	(0.71)	(1.41)	(1.41)
	Hard	*M*	2.5	2.00	2.50	2.00	4.00	4.00	5.0	5.50	5.67	5.50	5.50
		SD	(0.71)	(0.00)	(0.71)	(0.00)	(0.00)	(0.00)	(0.00)	(0.71)	(1.15)	(0.71)	(0.71)

Note. BrE = Boredom related Effort, DrE = Difficulty related Effort.

Table 2

Means and Standard Deviations of Tonic and Phasic Pupil Size Corresponding to Each Thought Probe in the Easy und Hard Stroop Version

Variable	Difficulty Level		1	2	3	4	5	6	7	8	9	10	11
Tonic PS	Easy	*M*	976.4	162.2	972.6	958.9	918.5	921.5	884.3	923.9	833.5	891.7	858.7
		SD	(174.7)	(168.7)	(161.8)	(132.8)	(106.1)	(45.3)	(17.9)	(82.0)	(75.0)	(45.5)	(113.5)
	Hard	*M*	970.3	915.1	811.5	865.1	828.8	784.8	795.2	747.1	673.5	644.8	632.0
		SD	(738.7)	(666.1)	(565.7)	(596.7)	(567.6)	(489.9)	(453.2)	(378.0)	(265.2)	(211.8)	(218.3)
Phasic PS	Easy	*M*	93.9	91.2	85.5	87.3	97.4	54.4	83.0	84.2	101.7	86.3	80.2
		SD	(16.7)	(11.4)	(6.4)	(2.3)	(4.6)	(20.1)	(0.2)	(1.7)	(2.7)	(6.6)	(1.9)
	Hard	*M*	156.1	160.6	160.1	124.2	105.1	118.8	107.2	113.5	123.2	115.6	88.9
		SD	(97.1)	(135.4)	(106.2)	(80.8)	(59.8)	(52.1)	(45.5)	(90.0)	(77.3)	(53.7)	(37.5)

Note. Tonic PS = Tonic (Baseline) Pupil Size, Phasic PS = Phasic (Stimulus-evoked) Peak Pupil Size, baseline corrected. Tonic Pupil Size and Phasic Pupil Size are averaged across the five trials preceding the thought probe and reported in pixels.

Appendix C

Study-Design Template

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**	**Theory that could be shown wrong by the 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 good someone performs in the tasks?	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 will be recruited[7].	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 hard Stroop version (as independent variable).		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 (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.	The premise that structurally more demanding tasks are also perceived as more difficult and yield lower performance scores. In addition: that self-reported boredom levels are higher in structurally easy tasks.

[7] 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.

					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.	
Main Research Question 1: Do boredom, effort demands due to boredom, perceived task-difficulty, effort demands due to perceived task difficulty, overall perceived effort and fatigue change over time during both tasks (easy and hard Stroop)?	We expect an increase in boredom, effort due to boredom, fatigue, and overall effort and a decrease of task difficulty and effort due to task difficulty over time for both Stroop versions (easy and hard). We expect a greater decrease of task difficulty and task-difficulty-related effort over time in the hard than in the easy version. We further expect an interaction for overall effort with a greater increase in overall effort in the easy Stroop version.		Six Linear Mixed Models including boredom, boredom related effort, task difficulty and 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.		If hypothesis is supported: Over time tasks become more boring which increases effort demands to continue with a boring task. When tasks become easier over time, this reduces effort demands due to task difficulty. At the same time overall effort and fatigue increase. This holds true for tasks with different difficulty levels. Nevertheless, task difficulty and the related effort decrease stronger in hard tasks (due to the initially higher task difficulty). Further, as task difficulty decreases to a greater extent in hard tasks than in easy tasks, and boredom increases in both tasks, overall effort (which is expected to include boredom and difficulty related boredom) increases less in harder tasks than in easy tasks.	N/A

					If our hypothesis is not supported: When results show that it depends on the tasks' difficulty level (easy, hard) if boredom, perceived task difficulty, the related effort, perceived overall effort, and fatigue decrease or increase over time, we would conclude that task characteristics are relevant for predicting how these variables change over time.	
Main Research Question 2: Are physiological measures of effort solely influenced by task-difficulty-related effort or are they also affected by boredom-related effort?	Pupil size (tonic and phasic), as physiological indicator of effort, is predicted more accurately by task-difficulty-related effort and boredom-related effort than by task-difficulty related effort alone.		We will construct four Linear Mixed Models (LMM) for each outcome variable (tonic pupil size, phasic pupil size). Difficulty level of the task (easy, hard), time (thought probe one to eleven), effort due to task difficulty and effort due to boredom will serve as fixed predictors and participants' score of the Short Boredom Proneness Scale (SBPS), the Beck Depression Inventory II (BDI-II), and the Brief Self-Control Scale (BSCS) as covariates in different models. We will account for general differences between participants in tonic and phasic pupil		Should our results, as hypothesized, show that pupil size as physiological indicator of effort is predicted more accurately when effort due to boredom is considered as predictor, we would characterize boredom as effortful. This would emphasize the importance of recognizing boredom as a confounding variable within self-control research and, more general, within the broader scope of study designs. Conversely, if pupil size is not predicted by cognitive effort due to boredom, it could be argued that being bored during task performance might not	Performing easy and hard tasks is not effortful on a physiological level. Enduring boredom is not effortful on a physiological level.

			size by including random intercepts and different effects of time on pupil size among participants by including random slopes for the variable time. To evaluate the relative performance of these models in predicting pupil size, the AIC, BIC, and WAIC will be reported. We will base our decision on the most optimal model regarding model fit and complexity on the BIC.		trigger changes in this physiological correlate of effort. However, this outcome could even indicate a connection between boredom and low arousal levels, a correlation that remains ambiguous. In this case, the perceived effort elicited by boredom might be less influential on the pupillary activity than the possible decrease in arousal.	
Research Question 3A: Is task performance in a secondary task influenced by the level of difficulty (easy, hard) in the first task?	Given the inconsistent results in ego depletion research (Giboin & Wolff, 2019; Hagger et al., 2016; Hagger et al., 2010), this research question will be tested exploratively.		We will conduct two paired t-tests. One t-test will analyze participants' error rate as dependent variable, while the other will focus on the reaction time. The difficulty level of the first task (easy, hard) will serve as the independent variable in both tests.		If secondary task performance is worse after performing the hard task: Working on a structurally more demanding task leads to reduced performance on a subsequent task. If secondary task performance is worse after performing the easy task: Working on a structurally less demanding task leads to reduced performance on a subsequent task. If secondary task performance does not differ: Structural demand	Ego Depletion: Self-Control needs in a preceding task leads to less remaining self-control for a secondary task which manifests in a worsened task performance. (Please note that even if a reduction in performance is observed, this would not constitute evidence for a depletion of a specific resource, as we do not measure such a resource in this research.) Mental Fatigue Theories: Mental fatigue occurs after a period of prolonged

					differences in a prior mental task do not differentially carry over to performance on a secondary mental task.	mental activity which leads to decrements in task performance.
Research Question 3B: Can task performance in a secondary task be explained by the perceived amount of effort due to task difficulty invested in the first task? And does considering perceived boredom related effort in the first task provide a more accurate prediction of performance in the second task than considering task difficulty related effort alone?	This research question will be tested exploratively.		Three distinct LMMs will be tested for each performance variable in the flanker task (error rate, reaction time). The first model will assess the impact of the difficulty level (easy, hard; fixed effect) and the perception of 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 perceived boredom related effort as a fixed effect instead of perceived task difficulty related effort. The third model will include both perceived task difficulty and boredom related effort as fixed effects.		If secondary task performance can be explained by task difficulty related effort in the preceding task: Subjectively investing effort due to the difficulty of a task reduces performance on a subsequent task. If secondary task performance can be explained more accurately by boredom related effort and task difficulty related effort together: Both the perception of a higher task difficulty related effort and boredom related effort in a first task, reduce performance in a subsequent task. Thus, performing boring as well as difficult tasks can lead to a worsened performance in a secondary task if effort (related to boredom or task difficulty) is perceived to be invested in the preceding task.	Ego Depletion: Self-Control needed in a preceding task leads to less remaining self-control for a secondary task which manifests in a worsened task performance. (Please note that even if a reduction in performance is observed, this would not constitute evidence for a depletion of a specific resource, as we do not measure such a resource in this research.) Mental Fatigue Theories: Mental fatigue occurs after a period of prolonged mental activity which leads to decrements in task performance.

			The model with the lowest BIC will be considered as the most optimal one. Note: While our preregistered protocol states that we would include random slopes, we did not include them, as these would not be appropriate given the structure of the data (no within-subject variation in the predictors).		If performance in the secondary task cannot be explained by task difficulty or boredom related effort in a first task: Performance in a secondary is not impacted by the perceived amount of effort (either task difficulty or boredom related) invested in a preceding task.	
Exploratory: Does boredom proneness affect the quantity of sweets participants take with them after completing the experiment?	1) Controlling for trait self-control, boredom proneness affects the amount of sweets taken. 2) No other hypothesis. Other analysis with key variables fully exploratory.		1) We will perform a multiple regression analysis with the amount of sweets taken as dependent variable, boredom proneness as predictor and trait self-control as control variable. 2) We will conduct correlation analyses between the amount of sweets taken and our key variables.		If boredom proneness affects the amount of sweets taken: Reward sensitivity after the performance of a prolonged task differs with higher levels of boredom proneness. General alterations in reward sensitivity are likely. This might impact behavior of individuals with higher boredom proneness in various ways. If boredom proneness does not affect the amount of sweets taken: Reward sensitivity after the performance of a prolonged task is not affected by boredom proneness. There are either	Individuals with higher boredom proneness levels show alterations in their reward sensitivity (which impacts need for rewards after experiencing boredom).

					no alterations in the reward system of individuals with higher boredom proneness or the alterations do not manifest in the need for sweets after performing prolonged tasks.	

Appendix D

Participant Characteristics

As preregistered, we assessed self-reported ADHD diagnosis, sleep deprivation, and intake of stimulants prior to the experimental sessions. None of the participants reported being diagnosed with ADHD. Nine participants reported feeling sleep deprived before both sessions, four reported sleep deprivation only before the easy Stroop session, and six only before the hard Stroop session. Although participants were advised not to consume any stimulants before the experimental sessions, twenty-six participants drank coffee, twelve smoked, and two consumed other stimulants before both sessions. Two participants drank coffee only before the easy session, and one participant drank coffee only before the hard session. Three participants smoked only before the easy session.

Fifteen participants were excluded based on our preregistered criteria: five completed only one session (two of them due to measurement errors in their first session), five had excessive error rates, five were excluded due to incomplete oculomotor data. Statistical comparisons between included participants (95) and participants that were excluded due to low performance or only attending one session (8) were conducted, to compare relevant variables. Mann-Whitney U tests indicated that participants differed significantly in age and trait self-control (with excluded participants being older and reporting higher self-control, see Table 1), but not in boredom-proneness and depression scores. A Fisher’s exact test indicated that gender did not differ significantly.

Table 1

Comparison of Included and Excluded Participants (Means, Standard Deviations, and Test Statistics)

Variable	Included Participants (N = 95)	Excluded Participants (N = 8)	*p*	W
Age	20.22 (4.48)	24.20 (8.61)	.04	211.5
Gender	75.79% female	62.50% female	.45	
Beck Depression Inventory Score (BDI-II; Beck et al., 1996)	10.18 (7.77)	6.38 (5.48)	.06	530.5
Short Boredom Proneness Scale Score (SBPS; Struk et al., 2017; German: Martarelli et al., 2021)	2.75 (1.11)	2.56 (1.37)	.53	432
Brief Self-Control Scale Score (BSCS; Tangney et al., 2018; German: Bertrams & Dickhäuser, 2009)	3.16 (0.71)	3.82 (0.70)	.02	193

Note. The excluded participants that were included in this comparison had extreme error rates or completed only one session. Excluded participants who completed only one session due to measurement errors are not included in this comparison. Gender was compared with the Fisher's exact test. Age, BDI-II Score, SBPS-Score, and BSCS-Score were compared using the Mann-Whitney U Test. *W* = Mann-Whitney U test Statistic.

Appendix E

Linear Mixed Models Predicting Pupil Size

Table 2

Predictors, Estimates, Confidence Intervals, and P-Values for Linear Mixed Models 1 to 4 of Tonic (Upper) and Phasic (Lower) Pupil Size

Linear Mixed Models Predicting Tonic Pupil Size

	Model 1						**Model 2**					
Predictors	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*t*	*p*	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*t*	*p*
(Intercept)	810.67	-0.07	753.10 – 868.23	-0.23 – 0.09	27.61	**<0.001**	825.08	-0.11	767.77 – 882.40	-0.27 – 0.04	28.22	**<0.001**
task [hard]	42.02	0.14	23.63 – 60.40	0.08 – 0.20	4.48	**<0.001**	68.43	0.23	57.23 – 79.63	0.19 – 0.26	11.98	**<0.001**
block	-6.45	-0.07	-10.73 – -2.18	-0.11 – -0.02	-2.96	**0.003**	-7.36	-0.08	-11.67 – -3.04	-0.12 – -0.03	-3.34	**0.001**
effort difficulty	7.45	0.06	3.22 – 11.68	0.03 – 0.10	3.45	**0.001**						
effort boredom							3.25	0.02	-0.32 – 6.82	-0.00 – 0.05	1.79	0.074
Random Effects												
σ^2	33887.66						33954.98					
τ_{00}	72203.33 participant						72876.56 participant					
τ_{11}	374.83 participant.block						371.86 participant.block					
ρ_{01}	-0.53 participant						-0.52 participant					
Marginal R^2 / Conditional R^2	0.020 / 0.632						0.018 / 0.634					

	Model 3						**Model 3ex (including interaction)**						
Predictors	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*t*	*p*	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*t*	*std. Statistic*	*p*
(Intercept)	802.66	-0.07	744.03 – 861.30	-0.23 – 0.08	26.84	**<0.001**	765.66	-0.04	705.64 – 825.69	-0.19 – 0.12	25.01	-0.47	**<0.001**
task [hard]	43.75	0.14	25.21 – 62.29	0.08 – 0.20	4.63	**<0.001**	164.01	0.16	121.73 – 206.29	0.10 – 0.22	7.61	5.09	**<0.001**
block	-6.96	-0.07	-11.28 – -2.64	-0.12 – -0.03	-3.16	**0.002**	-7.60	-0.08	-11.92 – -3.29	-0.12 – -0.03	-3.45	-3.45	**0.001**
effort difficulty	7.10	0.06	2.85 – 11.36	0.02 – 0.10	3.27	**0.001**	13.47	0.12	7.96 – 18.97	0.07 – 0.16	4.80	4.80	**<0.001**
effort boredom	2.58	0.02	-1.01 – 6.17	-0.01 – 0.05	1.41	0.158	6.85	0.05	1.92 – 11.79	0.01 – 0.09	2.72	2.72	**0.007**
task [hard] × effort difficulty							-15.31	-0.13	-22.22 – -8.40	-0.19 – -0.07	-4.34	-4.34	**<0.001**
task [hard] × effort boredom							-8.58	-0.07	-14.58 – -2.58	-0.11 – -0.02	-2.80	-2.80	**0.005**
Random Effects													
σ^2	33885.22						33586.48						
τ_{00}	72217.33 participant						72405.30 participant						
τ_{11}	371.85 participant.block						370.24 participant.block						
ρ_{01}	-0.53 participant						-0.53 participant						
Marginal R^2 / Conditional R^2	0.020 / 0.632						0.024 / 0.635						

	Model 4							
Predictors	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*t*	*std. Statistic*	*p*	*std. p*
(Intercept)	1074.67	-0.07	677.09 – 1472.24	-0.23 – 0.08	5.30	-0.90	**<0.001**	0.368

task [hard]	43.50	0.14	24.96 – 62.05	0.08 – 0.20	4.60	4.60	**<0.001**	**<0.001**
block	-6.95	-0.07	-11.27 – -2.63	-0.12 – -0.03	-3.15	-3.15	**0.002**	**0.002**
effort difficulty	7.17	0.06	2.92 – 11.43	0.03 – 0.10	3.30	3.30	**0.001**	**0.001**
effort boredom	2.57	0.02	-1.02 – 6.16	-0.01 – 0.05	1.40	1.40	0.161	0.161
sbps mean	-23.74	-0.09	-86.96 – 39.48	-0.32 – 0.14	-0.74	-0.74	0.462	0.462
bdi sum	-3.96	-0.10	-11.66 – 3.74	-0.30 – 0.10	-1.01	-1.01	0.314	0.314
bsc mean	-52.20	-0.12	-139.75 – 35.35	-0.32 – 0.08	-1.17	-1.17	0.242	0.242
Random Effects								
σ^2	33885.16							
τ_{00}	71080.20 $_{\text{participant}}$							
τ_{11}	371.87 $_{\text{participant.block}}$							
ρ_{01}	-0.52 $_{\text{participant}}$							
Marginal R^2 / Conditional R^2	0.037 / 0.638							

Linear Mixed Models Predicting Phasic Pupil Size

	Model 1						Model 2					
Predictors	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*t*	*p*	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*t*	*p*
(Intercept)	67.45	-0.37	58.56 – 76.34	-0.49 – -0.26	14.87	**<0.001**	72.70	-0.39	63.93 – 81.48	-0.50 – -0.28	16.25	**<0.001**
task [hard]	44.78	0.75	40.42 – 49.14	0.68 – 0.82	20.13	**<0.001**	47.07	0.79	44.34 – 49.80	0.74 – 0.83	33.78	**<0.001**
block	-1.31	-0.07	-1.92 – -0.71	-0.10 – -0.04	-4.24	**<0.001**	-1.19	-0.06	-1.82 – -0.56	-0.10 – -0.03	-3.70	**<0.001**

effort difficulty	0.71	0.03	-0.27 – 1.70	-0.01 – 0.07	1.41	0.158						
effort boredom							-0.78	-0.03	-1.61 – 0.05	-0.06 – 0.00	-1.83	0.067
Random Effects												
σ^2	2023.44						2020.53					
τ_{00}	1403.27 participant						1427.39 participant					
τ_{11}	4.42 participant.block						4.63 participant.block					
ρ_{01}	-0.66 participant						-0.65 participant					
Marginal R^2 / Conditional R^2	0.161 / 0.435						0.161 / 0.439					

	Model 3						**Model 3ex (including interaction)**						
Predictors	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*t*	*p*	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*t*	*std. Statistic*	*p*
(Intercept)	70.09	-0.37	60.81 – 79.36	-0.48 – -0.26	14.82	**<0.001**	64.93	-0.32	55.18 – 74.67	-0.43 – -0.20	13.06	-5.42	**<0.001**
task [hard]	44.13	0.74	39.72 – 48.54	0.66 – 0.81	19.63	**<0.001**	64.15	0.74	53.99 – 74.30	0.67 – 0.82	12.38	19.69	**<0.001**
block	-1.14	-0.06	-1.77 – -0.51	-0.09 – -0.03	-3.53	**<0.001**	-1.24	-0.07	-1.88 – -0.60	-0.10 – -0.03	-3.79	-3.79	**<0.001**
effort difficulty	0.85	0.04	-0.15 – 1.84	-0.01 – 0.08	1.66	0.096	2.63	0.12	1.32 – 3.94	0.06 – 0.17	3.94	3.94	**<0.001**
effort boredom	-0.87	-0.03	-1.71 – -0.03	-0.07 – -0.00	-2.04	**0.042**	-0.89	-0.03	-2.06 – 0.28	-0.08 – 0.01	-1.49	-1.49	0.137
task [hard] × effort difficulty							-3.76	-0.17	-5.43 – -2.10	-0.24 – -0.09	-4.44	-4.44	**<0.001**
task [hard] × effort boredom							-0.20	-0.01	-1.64 – 1.24	-0.06 – 0.05	-0.27	-0.27	0.785
Random Effects													
σ^2	2019.99						2008.91						
τ_{00}	1415.84 participant						1403.55 participant						
τ_{11}	4.61 participant.block						4.79 participant.block						
ρ_{01}	-0.65 participant						-0.64 participant						

Marginal R^2 / Conditional R^2	0.161 / 0.439	0.166 / 0.441

	Model 4					
Predictors	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*t*	*p*
(Intercept)	88.68	-0.37	35.66 – 141.70	-0.48 – -0.26	3.28	**0.001**
task [hard]	44.03	0.74	39.61 – 48.44	0.66 – 0.81	19.55	**<0.001**
block	-1.14	-0.06	-1.77 – -0.51	-0.09 – -0.03	-3.53	**<0.001**
effort difficulty	0.88	0.04	-0.12 – 1.87	-0.01 – 0.08	1.72	0.085
effort boredom	-0.86	-0.03	-1.70 – -0.02	-0.07 – -0.00	-2.01	**0.044**
sbps mean	-1.33	-0.02	-9.71 – 7.06	-0.18 – 0.13	-0.31	0.756
bdi sum	-0.66	-0.08	-1.68 – 0.37	-0.22 – 0.05	-1.26	0.209
bsc mean	-2.57	-0.03	-14.19 – 9.05	-0.17 – 0.11	-0.43	0.665
Random Effects						
σ^2	2020.03					
τ_{00} participant	1371.74					
τ_{11} participant.block	4.61					
ρ_{01} participant	-0.63					
Marginal R^2 / Conditional R^2	0.168 / 0.441					

Note. Effort Difficulty = Difficulty Related Effort, Effort Boredom = Boredom Related Effort, BP = Short Boredom Proneness Scale Score , BDI-II = Beck Depression Inventory-II Score, BSC = Brief Self-Control Scale Score. These tables were generated using the tab_model function of the R package *sjPlot*.

Linear Mixed Models Predicting Flanker Task Performance

Table 3

Results of Linear Mixed Models predicting Error Rate and Reaction Time in the Flanker Task

	Model 1 Error Rate						**Model 2 Error Rate**					
Predictors	*Estimates*	*Std. Beta*	*CI*	*Std. CI*	*t*	*p*	*Estimates*	*Std. Beta*	*CI*	*Std. CI*	*t*	*p*
Intercept	0.05	0.31	0.02 – 0.08	0.12 – 0.51	3.46	**0.001**	0.04	0.23	0.00 – 0.07	0.06 – 0.40	2.11	**0.036**
task [hard]	-0.01	-0.05	-0.04 – 0.02	-0.33 – 0.22	-0.39	0.698	0.01	0.12	-0.01 – 0.03	-0.07 – 0.30	1.24	0.215
congruence flanker [yes]	-0.06	-0.57	-0.08 – -0.04	-0.76 – -0.39	-6.09	**<0.001**	-0.06	-0.57	-0.08 – -0.04	-0.76 – -0.39	-6.12	**<0.001**
effort difficulty	0.01	0.11	-0.00 – 0.01	-0.03 – 0.26	1.52	0.128						
effort boredom							0.01	0.11	0.00 – 0.01	0.01 – 0.21	2.12	**0.034**
Marginal R^2 / Conditional R^2	0.090 / 0.168						0.096 / 0.177					
AIC	-591.340						-593.303					
BIC	-567.699						-569.662					
WAIC	-595.3						-601.8					

	Model 3 Error Rate						**Model 3 Reaction Time**					
Predictors	*Estimates*	*Std. Beta*	*CI*	*Std. CI*	*t*	*p*	*Estimates*	*Std. Beta*	*CI*	*Std. CI*	*t*	*p*
Intercept	0.03	0.28	-0.01 – 0.07	0.08 – 0.48	1.59	0.112	465.85	0.38	433.59 – 498.11	0.18 – 0.58	28.39	**<0.001**
task [hard]	0.00	0.01	-0.03 – 0.03	-0.27 – 0.30	0.08	0.935	-18.05	-0.17	-39.78 – 3.68	-0.38 – 0.04	-1.63	0.103
congruence flanker [yes]	-0.06	-0.57	-0.08 – -0.04	-0.76 – -0.39	-6.12	**<0.001**	-61.67	-0.59	-72.96 – -50.38	-0.69 – -0.48	-10.74	**<0.001**
effort difficulty	0.00	0.07	-0.00 – 0.01	-0.08 – 0.23	0.94	0.347	3.24	0.08	-2.09 – 8.57	-0.05 – 0.20	1.20	0.232
effort boredom	0.01	0.10	-0.00 – 0.01	-0.01 – 0.20	1.75	0.081	9.31	0.17	4.65 – 13.97	0.08 – 0.25	3.93	**<0.001**
Marginal R^2 / Conditional R^2	0.098 / 0.183						0.123 / 0.721					

AIC	-582.710	4336.505
BIC	-555.128	4364.086
WAIC	-599.3	4242.3

	Model 1 Reaction Time						Model 2 Reaction Time					
Predictors	*Estimates*	*Std. Beta*	*CI*	*Std. CI*	*t*	*p*	*Estimates*	*Std. Beta*	*CI*	*Std. CI*	*t*	*p*
(Intercept)	504.59	0.41	478.39 – 530.80	0.21 – 0.62	37.86	**<0.001**	474.47	0.33	445.51 – 503.43	0.14 – 0.51	32.21	**<0.001**
task [hard]	-25.36	-0.24	-47.21 – -3.51	-0.45 – -0.03	-2.28	**0.023**	-6.77	-0.06	-18.12 – 4.58	-0.17 – 0.04	-1.17	0.242
congruence flanker [yes]	-61.67	-0.59	-73.20 – -50.14	-0.70 – -0.48	-10.52	**<0.001**	-61.67	-0.59	-72.98 – -50.36	-0.69 – -0.48	-10.72	**<0.001**
effort difficulty	4.79	0.11	-0.59 – 10.16	-0.01 – 0.24	1.75	0.081						
effort boredom							9.73	0.17	5.11 – 14.34	0.09 – 0.26	4.14	**<0.001**
Marginal R^2 / Conditional R^2	0.093 / 0.708						0.118 / 0.717					
AIC	4353.205						4339.763					
BIC	4376.846						4363.404					
WAIC	4257.8						4244.9					

Appendix F

Data Summary

Descriptive Statistics

To comprehensively understand our variables, we conducted an analysis of descriptive statistics and presented them in a table (see Table 3). We calculated the means and standard deviations for various measures across eleven different time points. These statistics were 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, are reported in eleven blocks corresponding to the thought probes. Reaction times and error rates in the flanker task are reported separately for congruent and incongruent trials, as well as combined. Phasic peak pupil dilation during the Stroop task is reported as an average across all trials for both tasks. Additionally, it is 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.

Table 4

Means and Standard Deviations of Variables Corresponding to Each Thought Probe in the Easy und Hard Stroop Version

Variable	Task		1	2	3	4	5	6	7	8	9	10	11
Difficulty	Easy	*M*	2.87	2.88	2.86	2.88	2.91	2.97	3.11	3.04	3.19	3.17	3.32
		SD	(1.88)	(1.89)	(1.86)	(1.84)	(1.89)	(1.82)	(2.1)	(2.05)	(2.19)	(2.13)	(2.19)
	Hard	*M*	7.41	7.11	7.01	6.68	6.61	6.53	6.42	6.46	6.43	6.36	6.2
		SD	(1.47)	(1.59)	(1.7)	(1.91)	(1.96)	(1.97)	(2.03)	(2.15)	(2.15)	(2.15)	(2.3)
Difficulty Related Effort	Easy	*M*	3.15	3.32	3.2	3.19	3.12	3.13	3.27	3.36	3.48	3.4	3.27
		SD	(1.89)	(1.94)	(1.95)	(1.94)	(1.85)	(1.86)	(2.08)	(2.2)	(2.29)	(2.16)	(2.08)
	Hard	*M*	7.38	7.17	7.01	6.86	6.65	6.59	6.56	6.42	6.46	6.38	6.39
		SD	(1.55)	(1.62)	(1.73)	(1.77)	(1.96)	(1.92)	(1.98)	(2.13)	(2.15)	(2.21)	(2.13)
Boredom	Easy	*M*	3.69	3.95	4.51	5.07	5.59	5.75	5.94	6.16	6.32	6.51	6.65
		SD	(2.1)	(2.12)	(2.13)	(2.09)	(2.1)	(2.08)	(2.04)	(1.99)	(2.02)	(2.02)	(2.11)

Variable	Task		1	2	3	4	5	6	7	8	9	10	11
Boredom Related Effort	Hard	*M*	2.59	2.91	3.23	3.43	3.74	4.2	4.36	4.56	4.8	4.86	5.17
		SD	(1.63)	(1.75)	(1.88)	(2.02)	(2.07)	(2.34)	(2.27)	(2.45)	(2.46)	(2.54)	(2.47)
	Easy	*M*	3.33	3.65	4.11	4.47	4.59	4.75	5.11	5.21	5.49	5.51	5.53
		SD	(1.83)	(1.94)	(1.94)	(2.06)	(2.11)	(2)	(2.31)	(2.18)	(2.22)	(2.22)	(2.35)
	Hard	*M*	3.74	3.92	3.91	4.24	4.34	4.4	4.69	4.78	5.03	5.07	5.29
		SD	(2.27)	(2.22)	(2.23)	(2.35)	(2.24)	(2.27)	(2.4)	(2.48)	(2.45)	(2.52)	(2.52)
Effort	Easy	*M*	3.69	3.78	3.85	3.87	4.03	4.17	4.36	4.59	4.86	4.99	4.67
		SD	(2.04)	(1.97)	(2.11)	(1.99)	(2.03)	(1.96)	(2.14)	(2.19)	(2.1)	(2.24)	(2.34)
	Hard	*M*	7.33	7.03	6.98	6.99	6.69	6.73	6.68	6.83	6.68	6.78	6.72
		SD	(1.49)	(1.67)	(1.72)	(1.75)	(1.88)	(1.86)	(1.92)	(1.85)	(1.96)	(1.9)	(2.01)
Fatigue	Easy	*M*	3.78	4.34	4.83	5.21	5.68	5.82	6.19	6.28	6.44	6.8	6.79
		SD	(2.13)	(2.06)	(2.18)	(2.22)	(2.18)	(2.06)	(2.09)	(2.12)	(2.12)	(2.08)	(2.09)
	Hard	*M*	3.69	3.95	4.4	4.66	4.94	5.26	5.46	5.61	5.93	6.18	6.36
		SD	(2.08)	(2.12)	(2.27)	(2.29)	(2.4)	(2.41)	(2.42)	(2.5)	(2.56)	(2.55)	(2.64)
Tonic Pupil	Easy	*M*	829.71	801.31	775.04	774.99	775.56	814.31	794.39	791.26	813.19	777.42	811.37
		SD	(300.61)	(294)	(259.02)	(254.91)	(258.63)	(408.62)	(286.27)	(265.57)	(390.56)	(265.95)	(417.14)
	Hard	*M*	978.24	936.93	888.26	874.76	831.27	827.31	844.5	848.77	825.39	836.22	812.04
		SD	(346.07)	(320.65)	(283.82)	(287.9)	(265.62)	(268.99)	(264.1)	(288.63)	(261.27)	(266.74)	(267.76)
Phasic Pupil	Easy	*M*	67.95	64.55	64.21	61.68	60.31	67.77	57.23	60.26	63.34	59.06	54.62

Variable	Task		1	2	3	4	5	6	7	8	9	10	11
		SD	(41.04)	(48.15)	(40.67)	(38.51)	(40.61)	(94.05)	(47.58)	(43.65)	(56.28)	(55.44)	(32.81)
	Hard	*M*	112.32	110.48	124.46	120.13	108.33	111.91	102.22	105.13	103.09	105.67	96.8
		SD	(64.07)	(58.89)	(65.48)	(56.07)	(54.27)	(56.38)	(56.48)	(51.09)	(56.83)	(64.41)	(50)
Phasic Pupil Meaning Trials Only	Hard	*M*	140.81	131.53	160.35	147.53	153.32	141.89	135.05	136.66	125.63	130.44	129.75
		SD	(121.69)	(101.51)	(112.25)	(94.91)	(101.5)	(95.25)	(104.82)	(84.3)	(87.83)	(85.62)	(86.2)
Phasic Pupil Color Trials Only	Hard	*M*	107.06	105.78	115.8	112.73	98.69	104.71	95.29	97.53	97.54	94.18	88.71
		SD	(93.81)	(90.06)	(90.99)	(92.67)	(88.46)	(85.22)	(90.03)	(82.54)	(81.97)	(77.97)	(79.09)
Stroop Reaction Time	Easy	*M*	722.59	733.55	711.14	707.39	705.07	697.89	697.4	703.65	696.03	709.29	697.92
		SD	(114.45)	(113.63)	(111.99)	(116.29)	(111.84)	(118.61)	(125.07)	(137.08)	(120.17)	(130.55)	(126.48)
	Hard	*M*	1281.76	1226.84	1199.1	1163.98	1132.22	1119.99	1109.99	1088.01	1075.21	1069.84	1060.93
		SD	(181.07)	(210.93)	(198.3)	(193.79)	(189.83)	(189.52)	(189.07)	(171.85)	(169.85)	(164.19)	(173.5)
Stroop Error Rate	Easy	*M*	0.04	0.03	0.04	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05
		SD	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.06)	(0.07)	(0.07)	(0.08)	(0.07)
	Hard	*M*	0.33	0.27	0.22	0.18	0.16	0.15	0.15	0.16	0.16	0.15	0.16
		SD	(0.23)	(0.2)	(0.17)	(0.15)	(0.16)	(0.16)	(0.17)	(0.16)	(0.17)	(0.15)	(0.16)
Flanker Reaction Time	Easy	*M*	489.38	489.38	489.38	489.38	489.38	489.38	489.38	489.38	489.38	489.38	489.38
		SD	(114.53)	(114.53)	(114.53)	(114.53)	(114.53)	(114.53)	(114.53)	(114.53)	(114.53)	(114.53)	(114.53)
	Hard	*M*	480.55	480.55	480.55	480.55	480.55	480.55	480.55	480.55	480.55	480.55	480.55
		SD	(94.96)	(94.96)	(94.96)	(94.96)	(94.96)	(94.96)	(94.96)	(94.96)	(94.96)	(94.96)	(94.96)

Variable	Task		1	2	3	4	5	6	7	8	9	10	11
Flanker Reaction Time Congruent Trials Only	Easy	*M*	457.64	457.64	457.64	457.64	457.64	457.64	457.64	457.64	457.64	457.64	457.64
		SD	(87.08)	(87.08)	(87.08)	(87.08)	(87.08)	(87.08)	(87.08)	(87.08)	(87.08)	(87.08)	(87.08)
	Hard	*M*	450.62	450.62	450.62	450.62	450.62	450.62	450.62	450.62	450.62	450.62	450.62
		SD	(68.46)	(68.46)	(68.46)	(68.46)	(68.46)	(68.46)	(68.46)	(68.46)	(68.46)	(68.46)	(68.46)
Flanker Reaction Time Incongruent Trials Only	Easy	*M*	521.11										
		SD	(129.44)										
	Hard	*M*	510.48										
		SD	(107.85)										
Flanker Error Rate	Easy	*M*	0.03										
		SD	(0.08)										
	Hard	*M*	0.05										
		SD	(0.14)										
Flanker Error Rate Congruent Trials Only	Easy	*M*	0.01										
		SD	(0.01)										
	Hard	*M*	0.01										
		SD	(0.02)										
Flanker Error Rate Incongruent Trials Only	Easy	*M*	0.06										
		SD	(0.1)										
	Hard	*M*	0.08										

Variable	Task		1	2	3	4	5	6	7	8	9	10	11
		SD	(0.18)										

Note. The easy Stroop task consisted only of color-naming trials. Therefore, phasic pupil responses were calculated separately for color and meaning trials only in the hard Stroop task. The flanker task was not divided into blocks, so only a single average value was calculated.

Correlation Analysis

To elucidate the relationships between the key variables in our study, we conducted correlation analyses and presented them in a correlation table (see Table 5). Boredom, perceived task difficulty, effort due to boredom, effort due to difficulty, overall effort, and fatigue which were assessed with the thought probes were included in the analysis. Concerning the tasks, the error rate and the reaction time in the easy version and in the hard version of the Stroop task as well as the error rate and reaction time in the flanker task were integrated. Regarding the pupil size, the average tonic pupil size during the easy and during the hard version of the Stroop as well as the average phasic pupil size during the easy and hard version of the Stroop were encompassed. It should be noted that, although effort due to boredom and effort due to task difficulty were correlated when averaged across the entire Stroop task, variance inflation factors (VIF) indicated that multicollinearity was not problematic in our models predicting pupil size that included both variables. In model 3, which included the two effort variables, the VIFS for both variables were below 3 (tonic pupil size: VIF for effort due to difficulty = 2.75, VIF for effort due to boredom = 1.04; phasic pupil size: VIF for effort due to difficulty = 2.61, VIF for effort due to boredom = 1.09). In the exploratory version of model 3, which additionally included the interactions with task, the VIFS for both variables remained below 5 (tonic pupil size: VIF for effort due to difficulty = 4.64, VIF for effort due to boredom = 1.99; phasic pupil size: VIF for effort due to difficulty = 4.30, VIF for effort due to boredom = 2.12).

Table 5

Means, standard deviations, and correlations for the hard (upper triangle) and easy Stroop (lower triangle)

Variable	M_{ES}	M_{HS}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Diff.	3.02 (1.77)	6.66 (1.73)	**1**	.94**	.09	.21*	.88**	.22*	-.06	-.12	-.07	.40**	.05	-.01	.16	.05	.05	-.02
2. DrE	3.26 (1.78)	6.72 (1.70)	.92**	**1**	.12	.26*	.95**	.29**	.03	-.06	.00	.31**	.05	.01	.15	.02	.03	-.08
3. Bored	5.47 (1.75)	3.99 (1.94)	.10	.10	**1**	.84**	.19	.58**	.02	.07	.17	.20*	.25*	.09	.02	.07	-.27**	-.11
4. BrE	4.70 (1.75)	4.49 (2.03)	.39**	.41**	.69**	**1**	.31**	.54**	-.00	.03	.24*	.22*	.29*	.17	.04	.08	-.20*	-.08
5. Effort	4.26 (1.68)	6.86 (1.61)	.84**	.81**	.26*	.60**	**1**	.38**	.04	-.06	.03	.29**	.09	.01	.16	.05	.01	-.10
6. Fatigue	5.65 (1.89)	5.31 (2.14)	.33**	.32**	.68**	.62**	.51**	**1**	.06	.03	.31**	.25*	.19	.00	.26**	.12	-.26*	.12
7. Tonic	796.2 (250.8)	864.0 (261.1)	.32**	.30**	-.01	.14	.33**	.14	**1**	.56**	-.04	-.26*	-.15	-.04	-.18	-.08	-.08	-.18
8. Phasic	61.91 (30.68)	109.1 (41.22)	.27**	.29**	.03	.16	.28**	.08	.67**	**1**	.09	-.29**	-.17	-.13	-.22*	-.17	.08	-.09
9. RT S	707.5 (108.6)	1138.9 (162.7)	.05	.09	.16	.11	.11	.19	.05	.06	**1**	.03	.38**	-.05	.02	-.07	-.04	-.00
10. Error S	0.04 (0.04)	0.19 (0.13)	.35**	.32**	.21*	.20	.34**	.30**	.21*	.11	-.06	**1**	.10	.05	.05	.02	.02	.06
11. RT F	489.4 (105.9)	480.6 (85.0)	.06	.09	.22*	.13	.20	.34**	.06	.13	.47**	.05	**1**	.39**	-.05	-.00	-.10	.05
12. Error F	0.03 (0.05)	0.05 (0.09)	.24*	.28**	.08	.09	.23*	.06	.00	.12	-.05	.27**	.36**	**1**	.12	.04	-.11	-.00

13. BDI	10.61 (7.77)	10.61 (7.77)	.10	.06	.11	.15	.13	.19	-.07	-.13	.03	.00	.09	.10	**1**	.63**	-.46**	.18
14. SBPS	2.75 (1.11)	2.75 (1.11)	-.08	-.05	.19	.17	.03	.20	-.09	-.12	-.02	.07	.01	-.01	.63**	**1**	-.66**	.06
15. BSCS	3.16 (0.71)	3.16 (0.71)	.05	.07	-.09	-.13	-.06	-.24*	.05	.12	.02	.03	-.01	-.01	-.46**	-.66**	**1**	-.13
16. Sweets	1.20 (0.85)	1.16 (0.93)	.22*	.20	.07	.17	.14	.08	.11	-.06	-.06	.03	-.05	.04	.30**	.19	-.16	**1**

Note. Diff. = Difficulty, DrE = Difficulty Related Effort, BrE = Boredom Related Effort, RT S = Reaction Time Stroop, Error S = Error Rate Stroop, RT F = Reaction Time Flanker Task, Error F = Error Rate Flanker Task, BDI = Beck Depression Inventory-II Score, SBPS = Short Boredom Proneness Scale Score, BSCS = Brief Self-Control Scale Score, Sweets = Amount of Sweets Taken. *MES* and *MHS* represent the means for the easy and hard Stroop, respectively. Standard deviations are presented in brackets. * indicates $p < .05$. ** indicates $p < .01$.

Descriptive Information and Internal Consistencies of Questionnaires

The table below presents the number of items, response range, mean scores with standard deviations, and internal consistencies (Cronbach's α) for all questionnaires used as covariates in the linear mixed models.

Table 6

Descriptive Information and Internal Consistencies of Questionnaires

Questionnaire	Number of Items	Scale	Score (SD)	Cronbach's α
Beck Depression Inventory II (BDI-II; Beck et al., 1996)	21	4 statements, best-fitting selected (coded 0 – 3)	10.61 (7.77)	0.89

Short Boredom Proneness Scale (SBPS; Struk et al., 2017; German: Martarelli et al., 2021)	8	1 (strongly disagree) to 7 (strongly agree)	2.75 (1.11)	0.86
Brief Self-Control Scale (BSCS; Tangney et al., 2018; German: Bertrams & Dickhäuser, 2009)	13	1 (not at all like me) to 5 (very much like me)	3.16 (0.71)	0.85

Note. Scores represent the mean across participants: SBPS and BSCS are reported as mean item scores, BDI-II as the total sum across items.

Explorative Analysis: Sweets

To assess whether boredom proneness affected the quantity of sweets participants took with them after completing the experiment, a multiple linear regression was conducted to predict the quantity of sweets taken predicted by boredom proneness while controlling for self-control and the Stroop tasks' difficulty level. The overall model was not significant, $F(3,186) = 1.41$, $p = .242$, and explained only a small proportion of variance in sweets quantity ($R^2 = .022$, adjusted $R^2 = .006$). None of the predictors were significant, neither the Stroop tasks' difficulty level ($B = -0.04$, $\beta = -0.05$, $t = -0.33$, $p = 0.742$), nor self-control ($B = -0.13$, $\beta = -0.10$, $t = -1.09$, $p = 0.278$), or boredom proneness ($B = 0.04$, $\beta = 0.06$, $t = 0.58$, $p = 0.563$). As the model assumptions of normality of the residuals and independence of errors were not met, a robust regression was conducted, leading to similar results regarding significance.

We further explored associations between our key variables and the quantity of sweets participants took with them by including this variable in our correlation analysis. The results indicated that participants' depression scores ($r(93) = .22$, $p < .05$) and perceived task difficulty ($r(93) = .30$,

$p < .01$) were positively correlated with the number of sweets taken after the easy Stroop task. No significant correlations between the number of sweets and other variables were observed in the hard Stroop task.

Results suggest that there were either no alterations in the reward system of individuals with higher boredom proneness or the alterations did not manifest in the need for sweets after performing boring and effortful tasks. Instead, a significant relationship between BDI-scores and the amount of sweets taken after the easy Stroop was observed. This suggests that, for individuals with more symptoms of depression, boring activities might trigger reward-seeking behavior. However, various explanations are plausible, and further research is necessary to confirm this interpretation.