## Test Boredom While Working on Difficult versus Easy Tasks: The Same Emotion but Different Effects on Performance

# Test Boredom While Working on Difficult versus Easy Tasks: The Same Emotion but Different Effects on Performance

Thomas Goetz[1], Sarah Stoll[1], Caroline A. Adam[1], Maik Bieleke[2], Anne C. Frenzel[3], Jonathan Fries[1], Lukas Kraiger[1], Lisa Stempfer[1], Takuya Yanagida[1], and Reinhard Pekrun[4, 5, 3]

[1] Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna

[2] Department of Sport Science, University of Konstanz

[3] Department of Psychology, Ludwig-Maximilians-Universität München

[4] Department of Psychology, University of Essex

[5] Institute for Positive Psychology and Education, Australian Catholic University

Citation for this article:

Thomas Goetz, T., Stoll, S., Adam, C. A., Bieleke, M., Frenzel, A. C., Fries, J., Kraiger, L., Stempfer, L., Yanagida, T., & Pekrun, R. (2026, in press). Test boredom while working on difficult versus easy tasks: The same emotion but different effects on performance. *Journal of Educational Psychology.*

## Author Note

Thomas Goetz: https://orcid.org/0000-0002-8908-2166

Sarah Stoll: https://orcid.org/0000-0003-4930-0277

Caroline A. Adam: https://orcid.org/0009-0005-8903-6513

Maik Bieleke: https://orcid.org/0000-0003-2586-1416

Anne C. Frenzel: https://orcid.org/0000-0002-9068-9926

Jonathan Fries: https://orcid.org/0000-0003-3253-5109

Lukas Kraiger: https://orcid.org/0009-0006-6792-0444

Lisa Stempfer: https://orcid.org/0000-0003-2748-3738

Takuya Yanagida: https://orcid.org/0000-0001-9052-4841

Reinhard Pekrun: https://orcid.org/0000-0003-4489-3827

We have no conflicts of interest to disclose. The data used for this research and the findings have not previously been disseminated. The data, research materials, and analysis code are available at https://doi.org/10.17605/OSF.IO/WS23C.

Correspondence concerning this article should be addressed to Thomas Goetz, Department of Developmental and Educational Psychology, University of Vienna, Universitaetsstrasse 7 (NIG), 1010 Vienna, Austria. E-Mail: thomas.goetz@univie.ac.at.

**Abstract**

The focus of this work was on test boredom experienced while working on difficult versus easy tasks. In Study 1 ($N$ = 208 8th graders), we investigated the structural validity of test boredom while working on difficult tasks (i.e., overchallenge boredom) versus easy tasks (i.e., underchallenge boredom). Supporting our hypotheses, the Test Boredom Scale-State (TBS-S) demonstrated scalar invariance across both situations, with weak-to-moderate correlations of boredom between the two situations. Also consistent with our hypotheses, a prototypical single boredom item from the TBS-S showed measurement invariance but weak correlations across situations. These findings suggest that test boredom represents the same emotion in both situations and can be validly assessed with a single item. Based on these findings, in Study 2 ($N$ = 132 university students), we used this item to test the hypothesis that test boredom reduces cognitive resources and reduces performance when working on difficult tasks but not when working on easy tasks (i.e., the abundance hypothesis). We used an experimental within-person design in which participants worked on both difficult and easy digit span memory tasks. As expected, the results indicated that boredom experienced during difficult tasks negatively affected performance by reducing cognitive resources (measured via reaction time), a consequence of processing boredom. In contrast, boredom experienced during easy tasks had no negative impact on performance, likely because cognitive resources remained abundant even when processing boredom. Implications for future research and practice are discussed.

*Keywords:* abundance hypothesis, boredom, cognitive resources, performance, test, control-value theory

## Educational Impact and Implications Statement

Our research indicates that boredom that arises from working on difficult tasks versus easy tasks represents the same emotion. However, distinguishing between the experiences of boredom in these differing situations is crucial for understanding their effects. Unlike boredom experienced while working on easy tasks (i.e., underchallenge boredom), boredom when working on difficult tasks (i.e., overchallenge boredom) impairs performance by depleting cognitive resources. Thus, while boredom as an unpleasant emotional state is generally undesirable, educators should recognize that boredom caused by working on difficult tasks is particularly detrimental.

# Test Boredom While Working on Difficult versus Easy Tasks: The Same Emotion but Different Effects on Performance

During the last 15 years, there has been a significant increase in studies on boredom (Bieleke et al., 2024). A key reason for this growing interest is the accumulating empirical evidence on boredom's negative effects on numerous important outcomes, including physical and psychological health, eating habits, drug use, motivation, and performance (Stempfer et al., 2025). The present research contributes to the field by examining potentially different effects of boredom based on its specific antecedents.

A recent study by Goetz et al. (2023) has shown that boredom arises both when working on difficult tasks and when working on easy tasks. However, it is quite unclear whether boredom represents the same emotional experience across both types of situations. In Study 1, we addressed this issue and hypothesized that boredom experienced when working on difficult versus easy tasks is the same emotion. We also hypothesized that, due to their different situational antecedents, correlations between the two experiences of boredom would be moderate in size. Since both multi-item scales and single items are commonly used in boredom research, we tested both hypotheses using a multi-item boredom scale and a prototypical single boredom item that is part of this scale.

Even if boredom experienced when working on difficult versus easy tasks is the same emotion, its effects on performance might nevertheless differ. This assumption was recently formulated in the "abundance hypothesis" (Goetz et al., 2023). In their work, Goetz et al. (2023) found, in line with this hypothesis, that boredom when working on difficult tasks has negative effects on performance outcomes, whereas boredom when working on easy tasks does not. However, Goetz et al. did not investigate the reasons for these differing effects, even though they

were outlined in the theoretical part of their work. As such, we focused on this aspect in our Study 2. When working on difficult tasks, all cognitive resources are needed to successfully complete the task. As such, a reduction of resources due to processing test boredom should lead to lower performance. In contrast, when working on easy tasks, cognitive resources might be abundant. Thus, resources used for processing test boredom should have no effect on performance outcomes.

In summary, Study 1 examines whether test boredom experienced when working on difficult versus easy tasks is the same emotion. Study 2 aims to enhance our understanding of why boredom has different effects on performance outcomes depending on whether one is working on difficult or easy tasks.

## Theoretical Background

### Definition, Occurrence, and Importance of Test Boredom

Goetz et al. (2023) defined test boredom as boredom experienced in situations labeled and/or perceived as tests. To conceptualize boredom in an operationalizable manner, we draw on the component process model of emotions (Scherer, 2000; Scherer & Moors, 2019). This model posits that emotions are best understood through their underlying processes. Based on this framework, boredom is defined as an emotional process comprising four component processes that together create a unique, boredom-specific profile. These comprise affective (an unpleasant, aversive feeling), cognitive (an altered perception of time, mind wandering), motivational (a desire to withdraw from the current situation), and physiological/expressive components (low arousal, yawning, looking tired; Goetz et al., 2019, 2024; Pekrun et al., 2010, 2014). Test boredom can be conceptualized either as a trait or a state. This distinction is consistent with research on test anxiety, which has traditionally differentiated between trait and state test anxiety

(Zeidner, 1998). Trait test boredom is defined as habitual boredom in situations labeled and/or perceived as tests - that is, boredom that recurs across such situations and over time. State test boredom refers to the current experience of boredom in situations labeled and/or perceived as tests.

Concerning the occurrence of test boredom, the existing scattered empirical evidence refers to low-stakes mathematics tests in schools, assessed via 5-point response scales (i.e., 1-5). Raccanello et al. (2019) reported a mean level of $M = 1.96$ for trait test boredom in elementary students. In a sample of 6$^{th}$ graders, Goetz et al. (2007) found mean levels of state test boredom to be $M = 1.98$ and $M = 2.11$ for two assessments within a single test. Goetz et al. (2023) reported mean levels of $M = 1.91$ for state test boredom and $M = 1.48$ for trait test boredom in 8$^{th}$ graders (Study 1). For 5$^{th}$ to 10$^{th}$ graders (Study 2), a mean level of $M = 1.84$ for state test boredom was reported in this study. The authors noted that, despite these relatively low means, the scores were distributed across a wide range, with some students reaching the highest possible score. Asseburg and Frey (2013) reported a mean level of state test boredom of $M = 2.36$ for 9$^{th}$ graders as experienced during the PISA 2006 mathematics test (Prenzel & Blum, 2007).

In sum, in students from elementary school to 10$^{th}$ grade, state levels of test boredom appear to be around a value of 2 on a 5-point response scale, while trait levels are reported to be somewhat lower. This difference is likely due to the influence of subjective beliefs on emotions (e.g., “A test cannot be boring”), which tend to have a stronger influence on trait assessments than state assessments (e.g., Goetz et al., 2013; Robinson & Clore, 2002). Compared to other negative emotions, the levels of state test boredom are relatively high. For example, in the study by Goetz et al. (2007), the values for anger and anxiety were $M = 1.44$/ 1.57 and 1.32/1.31, respectively (based on two assessments within a single test using a 5-point response scale; 1-5).

Roos et al. (2021) reported levels of anxiety during a low-stakes math test ranging from $M = 1.24$ to $M = 1.60$ (assessed on a 6-point response scale [0-5] across six assessments during the test).

Concerning the importance of test boredom, only one study addressed effects of this emotion. Goetz et al. (2023) demonstrated significant negative effects of test boredom on performance - however, only when working on difficult tasks. This finding aligns with a large body of research showing that boredom, beyond testing situations, relates negatively to achievement outcomes (e.g., Camacho-Morles et al., 2021; Daniels et al., 2009; Goetz et al., 2010; Pekrun et al., 2010, 2011, 2014). Given that boredom has further negative effects beyond performance (e.g., on health, eating habits, drug use, and motivation; Stempfer et al., 2025), and considering the very high number of testing situations people encounter throughout their lives (especially during the school years), test boredom can be regarded as an important emotion worthy of further investigation.

**Boredom while Working on Difficult versus Easy Tasks: The Same Emotion?**

Boredom can be expected to arise when working on both difficult and easy tasks (Goetz et al., 2023). This assumption is based on core theories regarding the antecedents of boredom, which state that boredom arises in situations with levels of control that are inappropriate for the individual. Specifically, boredom occurs in situations of low control (i.e., working on difficult tasks - overchallenge) as well as in situations of high control (i.e., working on easy tasks - underchallenge; Pekrun, 2006, 2018, 2024; Westgate & Wilson, 2018). Thus, boredom can be assumed to arise in both situations. Accordingly, when task difficulty and individual ability are well matched, low levels of boredom are expected (cf. research on “flow”; Csikszentmihalyi, 1975/2000).

Related to the two situational antecedents of boredom, a key question is whether test

boredom arising from working on difficult versus easy tasks represents the same emotion. If they are different, then comparing their effects on performance, for example, would be akin to comparing apples to oranges. From a measurement perspective: Do measures of boredom show measurement invariance across both situations? Invariance would indicate structural validity of boredom measures across these situations, suggesting that boredom is the same construct in both situations. In contrast, a lack of invariance would suggest the existence of different types of boredom specific to each situation. Knowledge of the structural validity of boredom measures is crucial in determining whether boredom has differential effects on outcomes depending on the situational context, or whether it is different types of boredom that generate these differences. To our knowledge, no study has investigated the structural validity of boredom in relation to situations of working on difficult versus easy tasks. We assume that the same components constitute boredom with equal importance, regardless of whether it is caused by working on difficult or easy tasks. As such, we hypothesized that the two types of boredom represent the same emotion.

An important related question is whether it is possible to draw conclusions about an individual's disposition to experience boredom when working on difficult tasks based on their disposition to experience boredom when working on easy tasks, and vice versa. If the correlation between these dispositions is high, it would be sufficient to investigate boredom in just one situation (e.g., related to difficult tasks), as it would be possible to draw conclusions about the levels of boredom in the other situation. To our knowledge, this question has also not yet been answered by empirical research. However, test takers may vary in the extent to which they experience boredom when working on difficult versus easy tasks.

**Boredom During Difficult versus Easy Tasks: Different Effects on Performance?**

Until recently, theories of boredom did not differentiate between the effects of boredom caused by working on difficult tasks (i.e., overchallenge boredom) and those caused by working on easy tasks (i.e., underchallenge boredom). Core theories of boredom generally assumed, regardless of the antecedents, that boredom has negative effects on performance outcomes. For instance, the cognitive-motivational model of emotion effects within the control-value theory (CVT; Pekrun, 2006, 2018, 2024) suggests that boredom negatively impacts performance outcomes through its detrimental effects on cognition (e.g., mind-wandering), motivation, and behavior (e.g., disengagement and reliance on superficial learning strategies). Numerous studies (e.g., Daniels et al., 2009; Goetz et al., 2010; Pekrun et al., 2010, 2011, 2014) and several meta-analyses (e.g., Camacho-Morles et al., 2021; Li et al., 2025; Stempfer et al., 2025; Tze et al., 2016) support the proposition that boredom has negative effects on performance outcomes.

However, Goetz et al. (2023) proposed the abundance hypothesis, which states that boredom negatively affects performance outcomes when experienced while working on difficult tasks (i.e., overchallenge boredom) but has no detrimental effects on test performance when working on easy tasks (i.e., underchallenge boredom). The abundance hypothesis considers the mediating role of cognition in the effects of boredom on performance (Goetz et al., 2023). A core mediator, and perhaps the most important one for cognitive performance, likely is the reduction in cognitive resources caused by processing boredom (Eastwood et al., 2012; Stempfer et al., 2025). The processing of boredom is expected to require cognitive resources, regardless of whether it arises from working on difficult or easy tasks.

The strain on cognitive resources caused by boredom can result in slower cognitive processing of task information, with slower reaction times often serving as a key indicator of reduced speed (Wechsler, 2006; see also Claros-Salinas et al., 2013; Neumann et al., 2014). The

abundance hypothesis implies that the boredom-generated reduction of available resources will have different consequences for test performance, depending on whether test boredom arises when working on difficult or easy tasks. If boredom is caused by working on easy tasks (i.e., being underchallenged), negative effects are likely to be small or nonexistent. Even if the available cognitive resources are significantly reduced due to processing boredom, they may still be sufficient to successfully complete easy tasks. In other words, resources may still be abundantly available, making it possible to simultaneously process boredom and successfully complete the task. In contrast, if boredom arises when working on difficult tasks (i.e., when overchallenged), the reduction in available cognitive resources caused by boredom is expected to have a significant negative impact on performance. For difficult tasks, the resources used for processing boredom would have been critical for successfully solving the tasks. Based on these considerations, the abundance hypothesis states that boredom is more detrimental to performance when students work on difficult tasks than when they work on easy tasks.

There are clear connections between the abundance hypothesis and cognitive load theory (CLT; Sweller, 2023; Sweller et al., 2014). CLT is built on the assumption that the processing of information in working memory is limited in terms of both capacity and duration. If cognitive load exceeds the capacity of working memory, performance is impaired. We argue that the abundance hypothesis is not a separate theory but rather a specific application of CLT: when tasks are easy, intrinsic load is low and working memory resources remain abundant, allowing the extraneous load added by boredom to be accommodated without harming performance. However, when tasks are difficult, elevated intrinsic load means boredom-induced processing more readily tips working memory into overload and harms performance. This reasoning is in line with recent work that has argued to integrate emotion into CLT (Plass & Kalyuga, 2019).

A study with high school students (5th to 10th graders; Goetz et al., 2023) provided initial support for the abundance hypothesis: While boredom when working on easy tasks had no impact on math performance, boredom when working on difficult tasks showed a significant negative effect on performance outcomes. However, that study did not examine the mechanisms generating these differences. As such, empirical evidence is lacking to explain why boredom when working on easy versus difficult tasks exerts differential effects on performance. In the present research, we tested the role of cognitive resources available for task performance. These resources are expected to be insufficient, leading to reduced performance, when boredom is experienced while working on difficult tasks. In contrast, when boredom is experienced while working on easy tasks, sufficient cognitive resources should be available (i.e. abundant) so that there is no negative impact on performance.

## Aims and Hypotheses of the Present Research

### Study 1: Structural Validity and Intercorrelations of Test Boredom During Difficult versus Easy Tasks

We investigated the structural validity (i.e., construct validity) of test boredom when working on difficult versus easy tasks. Since both multi-item scales and single items are frequently used in boredom research, we used both the Test Boredom Scale – State (TBS-S; a 12-item measure; Goetz et al., 2023) and a prototypical single boredom item from the TBS-S (i.e., “I was bored”). We hypothesized that both the TBS-S and the single item would demonstrate measurement invariance across the two situations (*H1*). Measurement invariance would confirm that the same construct is assessed in both contexts. The analyses of the structural validity of the single item are especially important for Study 2, in which a single test-boredom item was used to assess experiences while working on difficult versus easy tasks.

Furthermore, we tested hypothesis *H2*, which posits that the correlation between boredom caused by difficult tasks and boredom caused by easy tasks is moderate, using both the TBS-S and the prototypical single boredom item.

In sum, we tested the following two hypotheses:

*H1_Total_Scale*: The TBS-S demonstrates measurement invariance across situations of working on difficult and easy tasks.

*H1_Single_Item*: The single boredom item "I was bored" demonstrates measurement invariance across situations of working on difficult and easy tasks.

*H2_Total_Scale*: The correlation of test boredom as assessed via the TBS-S across situations of working on difficult and easy tasks is moderate.

*H2_Single_Item*: The correlation of test boredom as assessed via the single boredom item "I was bored" across situations of working on difficult and easy tasks is moderate.

## Study 2: Differential Effects of Boredom During Difficult versus Easy Tasks: The Role of Cognitive Resources

Based on the abundance hypothesis, which assumes differential effects of test boredom on performance when working on difficult versus easy tasks, Study 2 examined the role of cognitive resource availability during task performance in this context. Cognitive resources were expected to be insufficient for task completion when boredom was generated by working on difficult tasks (i.e., overchallenge boredom), leading to reduced performance. In contrast, cognitive resources were expected to be abundant when test boredom was due to working on easy tasks (i.e., underchallenge boredom), resulting in no negative effects on performance. Reaction times were used as an indicator of the availability of cognitive resources. We tested the following hypothesis.

*H3:* Test boredom negatively impacts cognitive resources and performance during difficult tasks but has no effect on cognitive resources and performance during easy tasks.

## Transparency and Openness

In line with standards of openness and transparency (Nosek et al., 2015), we describe the sample and procedure and report any data exclusions in detail. The data from Study 1 were analyzed and graphically illustrated using Mplus 8.10 (Muthén & Muthén, 2017) and R 4.5.1 (R Core Team, 2025) using the dmacs package v 0.1.0.9002 (Dueber, 2025) and qgraph package v1.9.8 (Epskamp et al., 2012). The data from Study 2 were analyzed using the statistical programming environment R 4.5.1 (R Core Team, 2025). We used the package *lme4* to compute linear mixed-effect models (Bates et al., 2015). The package *ggplot2* was used for data visualizations (Wickham, 2016). For both studies, all data, measures, and analysis codes are available at the online repository associated with this article (Fries & Yanagida, 2025).

Both studies were conducted in accordance with the ethical standards described in the World Medical Association (WMA) Declaration of Helsinki. Study 1 is a re-analysis of data published in Goetz et al. (2023). There is no overlap of the current study with the analysis reported by Goetz et al. Study 1 was not pre-registered. Study 2 was pre-registered on OSF (Adam, 2024). Both studies were approved by the Institutional Ethics Review Board of the first author's institution, and all study procedures were deemed appropriate. Study 2 was part of a larger research project. No findings from this project have been published as yet.

## Study 1: Structural Validity and Intercorrelations of Test Boredom During Difficult versus Easy Tasks

The key goal of this study was to investigate whether boredom represents the same emotion in situations involving easy versus difficult tasks. Specifically, we examined the

structural validity of both the TBS-S and a prototypical single boredom item from this scale (i.e., "I was bored"), as well as the correlations of TBS-S scores across the two situations and the correlations of single-item scores across the two situations. This analysis was based on data from Goetz et al.'s (2023; Study 1) project.

## Method

### *Participants*

The sample comprised 208 students (54% female; mean age = 13.73 years, *SD* = 0.44, range: 12.65–15.55) from nine 8$^{th}$-grade math classes. These classes were drawn from four different schools within the high-achieving track of Germany's three-track secondary school system (i.e., Gymnasium), which enrolls approximately 40% of the total student cohort (Federal Statistical Office [Statistisches Bundesamt], 2020). 78% of the students were German, with 17% of these German students having a second nationality. The remaining 22% reported a total of 15 different nationalities.

### *Procedure*

Participants completed a low-stakes mathematics achievement test (paper-and-pencil format; 45 minutes) during their regular math classes. The test consisted of two parts: one with several very difficult tasks and another with several very easy tasks. This design aimed to elicit boredom from working on difficult tasks (i.e., overchallenge boredom) and easy tasks (i.e., underchallenge boredom). The order in which students completed the difficult and easy parts was fully counterbalanced. As reported in Goetz et al. (2023), from which the data for the current analysis were drawn, students reported significantly higher levels of overchallenge when working on difficult tasks compared to easy tasks. Boredom was assessed immediately after each part using the Test Boredom Scale-State (TBS-S), with students providing retrospective reports

of the boredom experienced while working on the math tasks. For more information about the procedure of this study, see Goetz et al. (2023; Study 1).

### *Measure of Boredom*

The TBS-S was developed by Goetz et al. (2023), based on the Achievement Emotions Questionnaire (AEQ; Pekrun et al., 2011, 2023). The scale consists of 12 items divided into four subscales, each representing a distinct component of boredom with three items per subscale. The subscales measure the affective (e.g., "*I was bored*"), cognitive (e.g., "*I was so bored that I found myself daydreaming*"), motivational (e.g., "*I would have preferred not to start at all with the math tasks because of boredom*"), and physiological/expressive (e.g., "*I was so bored that I was tired*") components of boredom. Responses are recorded on a 5-point scale ranging from 1 *(not at all true*), 2 (*slightly true*), 3 (*partly true*), 4 (*mostly true*), to 5 (*completely true*). Reliabilities (easy/difficult part) were α = .90/.91, .80/.87, .80/.83, and .82/.85 for the affective, cognitive, motivational, and physiological/expressive components, respectively. The reliabilities for the overall test boredom scores (i.e., comprising the mean scores of the four subscales [components]) were α = .91/.93 for the easy/difficult part. A complete list of items is available in Appendix A (in English and German [used in the present study]; for the TBS-S to be used in a concurrent assessment and for the Test Boredom Scale – Trait [TBS-T], see Goetz et al., 2023). We did not analyze the TBS-T data from Goetz et al. (2023) because the scale was not administered during difficult versus easy tasks. Consequently, these data do not permit testing the structural validity of the TBS-T across both situations.

### *Analytic Strategy*

**Structural validity – *H1***. To address hypothesis *H1_Total_Scale*, a series of confirmatory factor analyses (CFA) was conducted to test the TBS-S for measurement invariance

across situations involving easy versus difficult tasks. By using latent-variable estimates, we account for measurement error and reduce attenuation (Trafimow, 2016). Measurement invariance was tested for two measurement models: a first-order factor model including four factors representing the four subscales, and a second-order factor model including the same primary factors and a second-order boredom factor. First, configural invariance models were estimated, with freely estimated factor loadings and intercepts across the two situations, followed by metric invariance models with equal factor loadings, and finally scalar invariance models with both factor loadings and intercepts constrained to be equal across situations. Model fit was evaluated using CFI, TLI, RMSEA, and SRMR based on common cut-off criteria (Kline, 2023). Measurement invariance evaluation was based on model selection using BIC, which is recommended when the goal is to detect substantial non-invariance. The BIC shows a low rejection rate to small non-invariance (Liang & Luo, 2020). Note that a lower BIC value indicates a better trade-off between fit and complexity (Van de Schoot et al., 2012). All models were estimated using maximum likelihood estimation method with test statistics and standard errors robust to non-normality (MLR estimation) in Mplus 8.10 (Muthén & Muthén, 2017). The percentage of missing values across the 24 items ranged from 0.00% to 0.96%. Overall, 0.44% of data were missing, stemming from 17 participants with incomplete data. Full information maximum likelihood (FIML) estimation was used to deal with the missing data (see Enders, 2022).

Concerning hypothesis *H1_Single_Item*, no standard procedure exists to investigate measurement invariance for single items. Therefore, we drew on the work of Nye and Drasgow (2011) and Nye et al. (2019), which recommend effect-size indices for quantifying measurement invariance (i.e., the effect size $d_{\text{MACS}}$) that can be applied to single items. Low levels of $d_{\text{MACS}}$

indicate little measurement non-invariance. To calculate $d_{MACS}$, they recommend using referent items that demonstrate high construct validity across situations. Because our single item "I was bored" (item 1; see Appendix A) is part of the affective component of the TBS-S, which consists of three items, the two other affective items were appropriate referents (items 2 and 3; see Appendix A). The affective component subscale of the boredom scale was assumed to, and later shown to, demonstrate construct validity across the situations involving easy versus difficult tasks. We did not use items from other components, as their content is less closely related to the two affective items. Based on the two referent items, we computed two $d_{MACS}$ effect sizes, each quantifying the degree of measurement invariance of our single item across the situations

To calculate $d_{MACS}$ (Nye et al., 2019; Nye & Drasgow, 2011) relative to the reference items, we used the first-order factor model (see above, analyses on *H1_Total_Scale*). To define the metric of the latent variable for the affective component, the factor loading and intercept of the corresponding reference item were constrained to 1 and 0, respectively. We computed $d_{MACS}$ for item 1 by using both item 2 ("The math tasks seemed monotonous and dull to me from boredom") and item 3 ("The math tasks bored me to death"; see Appendix A) as referents. The effect size $d_{MACS}$ ranges between 0 and 1 and quantifies the degree of measurement non-invariance at the item level. It represents the standardized mean difference between the expected item scores for an individual in one situation (i.e., easy part - underchallenge) and the corresponding predicted item scores for the same individual in the other situation (i.e., difficult part – overchallenge). The effect size $d_{MACS}$ is interpreted according to benchmark values (Nyle et al., 2019) as follows: small ($d_{MACS}$ = 0.2), medium ($d_{MACS}$ = 0.4), and large ($d_{MACS}$ = 0.7). As such, a small effect size indicates a low degree of measurement non-invariance, which is the assumption in our study.

**Correlations – *H2***. As for hypothesis *H2_Total_Scale*, we investigated the relationship of the TBS-S between boredom caused by working on difficult versus easy tasks by estimating the correlation between the latent mean scores of boredom for these two types of tasks as assessed with the TBS-S. We calculated the correlations separately based on the first-order and second-order scalar invariance models. Additionally, we estimated the latent mean differences between boredom as experienced while working on easy versus difficult tasks.

For testing hypothesis *H2_Single_Item*, we computed the correlation of the single item between the two situations involving easy and difficult tasks. Additionally, we applied a dependent samples *t*-test to examine mean differences between the two types of tasks.

## Results

### *Structural validity – H1*

Regarding our analyses on the structural validity of the TBS-S (*H1_Total_Scale*), the CFAs revealed that scalar measurement invariance was supported for the TBS-S across situations involving difficult versus easy tasks, for both the first-order and second-order factor measurement models. The scalar invariance models showed the lowest BIC values (see Table 1). Furthermore, these models showed acceptable model fit, both for the first-order factor model, $\chi^2$ (223) = 351.19, $p < .001$, CFI = .940, TLI = 0.926, RMSEA = 0.053, SRMR = 0.054, and for the second-order factor model, $\chi^2$ (248) = 374.09, $p < .001$, CFI = .941, TLI = 0.935, RMSEA = 0.049 and SRMR = 0.060. Thus, our findings fully supported *H1_Total_Scale*.

Concerning our analyses on the prototypical single boredom item (*H1_Single_Item*), the effect size measure $d_{MACS}$ revealed that measurement invariance was supported (i.e., item 1: “I was bored”) across situations of working on easy versus difficult tasks. More specifically, results showed a negligible effect size of $d_{MACS} = 0.110$ relative to referent item 2 and $d_{MACS} = 0.086$

relative to referent item 3, according to benchmark values for $d_{MACS}$ (for more details, including correlations of the single items and all other items of the scale, see online supplemental material [S2]). This finding supports hypothesis *H1_Single_Item* and suggests that the impact of item non-invariance on subsequent analyses was negligible.

***Correlations – H2***

With respect to the correlations of the TBS-S across situations involving difficult versus easy tasks (*H2_Total_Scale*), latent correlations between the subscales were estimated based on the scalar invariance first-order factor model (see Table 2). The results showed high correlations between the different components of boredom, with $\hat{\rho}$ = .80 to .98 for boredom while working on easy tasks, and .87 to .95 for boredom while working on difficult tasks. However, the correlations across situations were much lower, ranging from $\hat{\rho}$ = .19 to .43 ($M$ = .30, $Md$ = .30). All correlations were statistically significant ($p < .05$). In Figure 1, the correlation matrix is visualized as a network. Based on the scalar invariance second-order factor model, the latent correlation between the second-order factors representing overall boredom while working on easy versus difficult tasks was $\hat{\rho}$ = .35 ($p < .05$). These results supported *H2_Total_Scale*.

Based on the scalar invariance first-order factor model, we estimated the latent mean differences between boredom as experienced while working on easy versus difficult tasks (Table 2). There were no statistically significant latent mean differences between the situations, with negligible effect sizes ranging from $d$ = 0.03 to 0.09. Likewise, there was no statistically significant latent mean difference in the second-order boredom factors between the situations; $M(SD)_{overchallenge}$ = 1.49(0.76); $M(SD)_{underchallenge}$ = 1.45(0.66); $d$ = 0.04 ($p$ = 0.49).

Supporting *H2_Single_Item*, the correlation between the prototypical single boredom item across situations involving difficult versus easy tasks was $\hat{\rho}$ = .27 ($p < .001$). The dependent

*t*-test did not indicate any mean difference across situations involving difficult ($M$ = 1.82, $SD$ = 1.16) versus easy tasks ($M$ = 1.75, $SD$ = 1.05), $t(207) = 0.73$, $p = .466$, $d = 0.06$.

**Discussion**

In support of *H1_Total_Scale* and *H1_Single_Item*, the results demonstrate invariance of both the TBS-S and the single item "I was bored," which is part of this scale, across situations involving difficult versus easy tasks. This finding suggests that boredom experienced during difficult and easy tasks represents the same emotion. As implied by scalar invariance, if someone experiences the same level of boredom during difficult and easy tasks, they would respond similarly to the TBS-S items and similar to the item "I was bored" as part of this scale. Our findings supported *H2_Total_Scale* and *H2_Single_Item* by revealing relatively weak correlations between reports of boredom across situations involving difficult versus easy tasks, both for the TBS-S and the single item as part of this scale. As such, the boredom experiences in the two situations can be clearly distinguished, from a correlational perspective. With respect to the TBS-S, we observed much stronger correlations among the components of boredom within situations (i.e., difficult or easy) than between these situations.

Furthermore, the findings indicate that the mean levels of both the boredom scale (i.e., components as well as the overall boredom scores) and the prototypical single boredom item did not differ between situations involving difficult and easy tasks. This finding indicates that the difficult tasks and the easy tasks led to the same level of boredom. However, changing the levels of difficulty in the two conditions might lead to different mean levels of boredom in each condition.

## Study 2: Differential Effects of Boredom During Difficult versus Easy Tasks: The Role of Cognitive Resources

Study 2 used a randomized counterbalanced within-subject experimental design (e.g., Shadish et al., 2002) to investigate the mechanisms underlying the abundance effect, focusing on the role of cognitive resources. We expected to find that test boredom negatively impacts cognitive resources (as measured by reaction time) and performance during difficult tasks, but not during easy tasks (H3).

Study 2 builds on the findings of Study 1, particularly the structural validity of measuring the prototypical single boredom item across situations involving difficult versus easy tasks. Furthermore, it expands the age range of research on test boredom by focusing on university students.

## Method

### *Participants*

The original sample consisted of 141 psychology students from all phases of the bachelor and master programs at the <first author's institution>. Nine participants were excluded due to invalid data, primarily resulting from technical issues during testing (i.e., internet failure). The final sample size was $N = 132$ (72% female, 26.5% male, 0.8% diverse, 0.8% not specified; mean age 22.08 years, $SD = 3.02$; range 18-38 years). In total, 109 participants (82.58%) reported their highest educational attainment as equivalent to the International Standard Classification of Education level 3–4 (ISCED-11; UNESCO Institute for Statistics, 2012). Among the remaining participants, 19 (14.39%) held a Bachelor's degree, three (2.27%) had completed only compulsory education, and one (0.76%) held a Master's degree. Participants generally reported strong German language proficiency: 123 (93.18%) were native speakers, five (3.79%) reported skills at the C1 level, three (2.27%) at the C2 level, and one (0.76%) at the A1 level.

Data were collected in February and March 2024 at the <institution of first author>. Participants were recruited through the official student recruiting system for experimental studies at the Faculty of Psychology, the Laboratory Administration for Behavioral Sciences (LABS) system. Participants received course credits for their participation. All participants provided informed consent prior to participation.

***Study Design and Experimental Manipulation***

We employed a fully counterbalanced two-group experimental design to manipulate participants' levels of boredom while completing a performance test. The manipulation was based on the theoretical proposition and prior empirical findings that boredom arises both during difficult tasks (i.e., overchallenge boredom) and easy tasks (i.e., underchallenge boredom). Participants completed both the easy and the difficult task blocks. The order of blocks was randomly assigned and counterbalanced across participants to control for sequence effects. All other procedures (instructions, timing, materials) were held constant across blocks. This experimental design is sometimes referred to as "crossover design" (e.g., Fleiss, 1986, p. 263).

With respect to manipulation tasks of varying difficulty, we used digit span memory tasks from the Wechsler Adult Intelligence Scale (WAIS-III, Wechsler, 1997; German version: WIE [Wechsler Intelligenztest für Erwachsene], Wechsler, 2006). For the difficult part of the test, the most difficult tasks of the WAIS-III were used - namely, digit span tasks consisting of 8-9 digits. For the easy part of the test, the easiest tasks of the WAIS-III were used - namely, digit span tasks consisting of 2–3 digits. This selection was based on empirical evidence indicating that working memory in young adults has a capacity limit of approximately three to five meaningful items (Cowan, 2010). Accordingly, number sequences containing eight or more digits can be classified as very difficult, whereas sequences containing only two or three digits can be

considered very easy. The manipulation trials, which were either very difficult or very easy - and thus assumed to be solved by nearly everyone or almost no one, respectively - were exclusively used to manipulate participants' boredom. It is important to note that these trials were not used to calculate performance scores (for more information on the underlying rationale, see online supplemental materials [S3]).

### *Measures*

**Test Boredom.** Using the same measure as Goetz et al. (2023; Study 2), we assessed boredom during the test using the single item "*I am bored*". We chose to use a single item to limit administration time and ensure the validity of our boredom measure (Gogol et al., 2014), given the number of boredom assessments during the test (i.e., four assessments in total, see Procedure). Based on our findings from Study 1, in which *H1_Single_Item* was supported, it can be strongly assumed that this single item measures the same construct of boredom when assessed during difficult and easy tasks, which was a crucial aspect of Study 2. Participants responded to the item on a 5-point rating scale (1 = *completely disagree* to 5 = *completely agree*).

**Perception of Over- and Underchallenge**. Following Goetz et al. (2023; Study 1), we assessed the extent of perceived over- and underchallenge during the difficult and easy tasks using the items "*I am feeling overchallenged*" and "*I am feeling underchallenged*". We chose to use single items for the same reasons as in our assessment of test boredom (i.e., again four assessments in total, see Procedure). Participants responded to the items on a 5-point rating scale ranging from 1 (*completely disagree*) to 5 (*completely agree*).

**Performance**. We assessed performance using the Digit Span subtest of the WAIS-III, similar to how boredom was manipulated (see above). The WAIS-III is one of the most widely used intelligence tests for adolescents and adults worldwide. The manual recommends

administering the test orally. In the present study, we used a computerized format. The test assesses the maximum length of sequences of digits that can be memorized.

The test comprises eight difficulty levels of digit sequences of increasing length, starting with two digits and increasing by one digit per difficulty level to a maximum of nine digits. The digit spans were presented for one second per digit and the participants had 30 seconds to enter their responses via a computer keyboard, either via the numeric keypad or the number row on the main keyboard. Each level consists of two consecutive sequences of equal length. As such, each test-taker has two attempts per level. If neither attempt at a level is completed or both are incorrect, the level is considered failed. It is considered correct if at least one of the two attempts is correct. The last correctly completed level is defined as the test score. For example, if at least one series is correctly completed at the six-digit difficulty level, but both series are incorrect at the seven-digit difficulty level, the test score would be six. In contrast to the original test version in the WAIS-III, the test was not aborted when participants failed at both attempts on a certain difficulty level. We did this to ensure that the duration and conditions were the same for all participants. Nevertheless, the test score was defined by the difficulty level prior to the difficulty level where participants failed both attempts for the first time, in line with the scoring procedures for the WAIS-III.

**Reaction Time**. We aimed to use a reaction time measure in which fast reactions clearly indicate high cognitive resources, while slow reactions indicate poor cognitive resources. To achieve this, we opted for a simple reaction task rather than a more complex one, avoiding the possibility that high cognitive resources might actually slow down reactions due to the initiation of elaboration processes or deeper thinking during task performance. Additionally, to ensure a clear distinction between reaction time and performance when testing our hypothesis (i.e., H3),

we did not include reaction time as part of the performance measure. Instead, reaction time was assessed using a completely separate task.

The reaction time (RT) assessment was implemented in a task in which participants looked at a white screen with a black fixation point in the centre. A black cross appeared irregularly in intervals between 1500ms and 5200ms at the location of the fixation point. As soon as the cross appeared, participants had to press the space bar on the computer keyboard as quickly as possible. The RT was recorded in milliseconds. For each participant, the arithmetic mean of the RTs across all RT tasks was calculated, representing their individual RT score.

***Procedure***

The experiment lasted about 90 minutes. On arrival at the laboratory, participants were informed that the study served to assess their emotions while they performed tasks on a computer. The experiment was conducted in groups, with an experimenter always present, on Windows 10 PCs using a custom website (Questionnaire Master, 2024) programmed specifically for this study using PHP (Hypertext Preprocessor), JavaScript, HTML, and CSS. Prior to the start of the experiment, participants were instructed to read the instructions for each task carefully, to answer any question spontaneously, and to aim for good scores on the performance tasks (all instructions are presented in the online supplemental material [S4]).

The experiment began with a five-minute warm-up session on the reaction time (RT) task to allow participants to become accustomed to the laboratory environment. Following the warm-up session, baseline RT was measured for five minutes. In the next step, participants were randomly assigned to one of the two experimental groups (i.e., Group A and Group B; see Figure 2, upper part).

*Group A* then underwent the manipulation designed to induce boredom by working on difficult manipulation trials (i.e., Block Difficult Trials; see Figure 2). The block lasted 20 minutes and included difficult manipulation trials, test trials, and a self-report assessment of overchallenge, underchallenge, and boredom. Participants were instructed that they would see digit spans of varying length that would disappear after a few seconds and that they would have to enter the digits using the computer keyboard (for the wording of the instructions, see online supplemental material [S4]).

The Block Difficult Trials began with a five-minute period of difficult manipulation trials. Then, there was a performance assessment, starting with test trials featuring digit spans of 2, 3, and 4. The test trials were presented sequentially in ascending order of difficulty (see the measures section). After the performance assessment, there was another continuous series of difficult manipulation trials. In addition to the difficult manipulation trials, test trials (i.e., performance assessments) with digit spans ranging from 5 to 9 digits were interspersed throughout this part of the block (for more details, see online supplemental material [S5]).

Within the Block Difficult Trials, perceived overchallenge, perceived underchallenge, and boredom were also assessed. Over- and underchallenge were assessed at 13 minutes, and boredom at 15 minutes after the block began (for more details, see online supplemental material [S6]). Both assessments took place while participants were working on the difficult manipulation trials.

After the Block Difficult Trials, perceived overchallenge, underchallenge, and boredom were assessed again. Subsequently, RT was again assessed for five minutes.

Next, participants began with the Block Easy Trials. The procedure and assessments were similar to the Block Difficult Trials, the only difference being that the easy manipulation trials

were used. The test trials in the Block Easy Trials were identical with those in the Block Difficult Trials, with an identical number of trials in both blocks.

However, there were more manipulation trials in the Block Easy Trials (i.e., Group A/B: $M_{trials}$ = 172.32/176.67, $SD_{trials}$ = 12.25/11.04), compared to the Block Difficult Trials (i.e., Group A/B: $M_{trials}$ = 71.97/76.33, $SD_{trials}$ = 5.59/6.22). The reason is that the overall time for both blocks was similar, but the digit trials were presented for a shorter duration in the easy trials.

After the Block Easy Trials, perceived overchallenge, underchallenge, and boredom were assessed again. RT was then assessed again for five minutes.

*Group B* started with the Block Easy Trails and then progressed to the Block Difficult Trials. This was the only difference between Groups A and B.

Finally, demographic data were assessed in both *Group A* and *Group B*. The experiment ended with a debriefing in which the variables analyzed were disclosed and the opportunity to contact the test administrator was offered.

***Analytic Strategy***

To test H3 which proposed that there would be negative effects of boredom on performance outcomes during difficult trials due to reduced cognitive resources (measured by reaction time), but no effects of boredom during easy trials, we examined the effects of boredom on task performance and deterioration of RT both for the difficult and easy parts of the test. RT deterioration was calculated as the difference between RT before and after the manipulation. For Group A, the difference for working on the difficult trials was calculated as RT after Block Difficult Trials minus Baseline RT, and the difference for working on the easy trials as RT after Block Easy Trials minus RT before Block Easy Trails (see Figure 2). For Group B, the difference for working on the easy trials was calculated as RT after Block Underchallenge minus

Baseline RT, and the difference for working on the difficult trials was calculated as RT after Block Difficult Trials minus RT before Difficult Trials.

We used linear mixed effects regression analysis to estimate the relations between boredom and performance as well as RT deterioration. We used the arithmetic mean of the within-block and post-block boredom scores as predictors in the regression models. The scores for boredom and the outcomes were standardized using *z*-transformations. To control for order effects, we included participants' group identifier (Group A or Group B) as a random effects term (i.e., as random intercepts). To quantify the variance in the outcome variable explained by the fixed-effect portion of our models, we calculated marginal R-squared following the procedure proposed by Nakagawa and Schielzeth (2013; for additional details on the statistical modeling, see online supplementary material [S8]).

**Results**

First, we analyzed whether our experimental conditions (i.e., working on difficult versus easy manipulation trials) produced the intended levels of over- and underchallenge. For the Block Easy Trials, the mean levels for perceived underchallenge, assessed during/after the block, were $M$ = 3.58/3.43 ($SD$ = 0.98/1.03), and the mean levels (during/after) for perceived overchallenge were $M$ = 2.00/2.27 ($SD$ = 0.96/1.08). For the Block Difficult Trials, mean levels during/after the block were $M$ = 1.89/2.24 (0.80 /0.85) for perceived underchallenge and $M$ = 3.28/3.33 ($SD$ = 1.04/1.03) for perceived overchallenge. In the Block Easy Trials, underchallenge scores (during/after) were significantly higher, $t$(131/131) = 16.88/13.09, $p$s < .001, Cohen's $d$ = 1.47/1.14, and overchallenge scores (during/after) were significantly lower, $t$(131/131) = -13.77/-12.59, $p$s < .001, Cohen's $d_z$ = -1.20/-1.10, compared with the Block Difficult Trials of the test.

Since the order of the Blocks Easy and Difficult Trials differed between the two experimental groups, we also examined whether the differences in levels of over- and underchallenge in the two blocks differed between the two groups. Cohen *d*s showed no significant differences (all *p*s > .15; range of z-scores: -0.57 to 1.04), indicating that the conditions had similar effects in both groups. In sum, our findings suggest that students in both groups actually experienced the Block Easy Trials as less challenging than the Block Difficult Trials. From these results, we conclude that our experimental manipulation successfully produced the intended over- and underchallenge boredom. In the online supplemental material (S7), we graphically illustrate the responses to the over- and underchallenge measures, each assessed during the test and separated by the two conditions (i.e., difficult and easy conditions; see Figure 2). Additionally, we report the percentage of participants who indicated low (i.e., values of 1 or 2 on the 5-point response scale) or high (i.e., values of 4 or 5) levels of overchallenge and underchallenge in the two conditions (see online supplemental material [S7]).

Table 3 shows the descriptive statistics for the study variables. The results show relatively high mean levels of boredom (i.e., $M = 3.56/4.02$ when working on the difficult/easy trials, respectively. The difference in means was significant, $p < .001$. Both when working on the difficult and easy manipulation trials, mean RT deteriorated ($M = 2.59/16.43$ for the difficult/easy manipulation trials, respectively), indicating on average an increase in RT from the beginning to the end of the experimental block. Mean levels of working memory performance were similar in both situations ($M = 6.55/6.95$ for the difficult/easy manipulation trials, respectively).

Table 4 presents the results for H3 (for additional analysis on the effects of group membership, see online supplemental material [S8]). Consistent with the abundance hypothesis,

we found a significant negative effect of boredom on working memory performance during the difficult part of the test (β = -0.18; *p* = .044). In contrast, the effect for the easy part of the test was not significant (β = -0.08; *p* = .353). Additionally, we observed a significant positive effect of boredom on RT deterioration (i.e., a decrease in RT) for the difficult part (β = 0.19; *p* = .023). In contrast, the effect for the easy part was not significant (β = 0.07; *p* = .407). This finding is consistent with our hypothesis that cognitive resources are still available (i.e., abundant) when working on easy tasks, even when some resources are devoted to processing boredom, and that sufficient resources are allocated to completing the tasks. Figure 3 provides a graphical illustration of these results, with both panels showing steeper regression slopes when working on difficult tasks. This visualizes that performance and RT were more strongly impacted by boredom during difficult tasks compared to easy tasks.

**Discussion**

Using our newly developed experimental manipulation, in which the manipulation trials used similar tasks as the performance trials (i.e., digit span memory tasks), we successfully achieved the targeted level differences in our boredom antecedent variables - namely, being overchallenged and underchallenged in the corresponding conditions. Our self-report assessments can be considered valid due to the use of single-item self-reports, which minimally disrupted the manipulation and were quick to complete. Regarding the assessment of our core variable, boredom, based on the results of Study 1, it can be inferred that our single-item measure of boredom captured the same construct across situations involving difficult versus easy tasks.

In line with *H3*, boredom experienced during difficult tasks (i.e., overchallenge boredom) led to a significant increase in reaction time (RT), whereas boredom during easy tasks (i.e.,

underchallenge boredom) did not. This finding suggests that boredom during difficult tasks negatively impacts performance by depleting cognitive resources that are essential for successfully completing the tasks. Although boredom during easy tasks also had negative effects on both outcomes (i.e., RT deterioration and performance), these effects were not statistically significant. As anticipated, our findings indicate that when working on easy tasks, cognitive resources remain sufficient to successfully complete the tasks while simultaneously processing boredom.

## General Discussion

In this research, we aimed to gain a deeper insight into the experience and effects of boredom when working on difficult as compared to easy tasks. In Study 1, we tested the hypothesis that test boredom demonstrates measurement invariance across situations involving working on difficult and easy tasks. We tested this hypothesis with respect to a boredom scale (i.e., the TBS-S) and a prototypical single-item measure of boredom included in this scale (i.e., "I was bored"). In addition, we tested the hypothesis that the relationship between boredom experienced during easy and difficult tasks is moderate, both for the total scale and the single item.

In Study 2, we tested the abundance hypothesis of boredom and investigated the role of cognitive resources available for task completion. These resources are expected to be insufficient, leading to reduced performance when boredom is experienced while working on difficult tasks (i.e., overchallenge boredom). However, they should be abundant and, therefore, have no negative effect on performance when boredom is experienced while working on easy tasks.

### Occurrence of Test Boredom

Study 2 contributes to the few existing studies on the occurrence of test boredom. Levels of test boredom are typically around a value of 2 on a 5-point response scale (1–5). In this study, we extended prior research on test boredom to university students and contexts beyond mathematics testing. The mean levels of test boredom while working on digit span memory tasks were above the midpoint of the 1-5 response scale we used (*M*s = 3.56 for difficult tasks, i.e., overchallenge boredom, and 4.02 for easy tasks, i.e., underchallenge boredom). To our knowledge, these are the highest test boredom scores ever reported. These values suggest that test boredom can, in fact, be an important emotion, due to its relatively high levels compared to other negative test emotions (e.g., state test anxiety, which typically ranges around *M* = 1.5; e.g., Roos et al., 2021) and its negative effects on performance, especially when working on difficult tasks.

**Boredom while Working on Difficult versus Easy Tasks: The Same Emotion?**

Based on our findings, the answer to this question is "yes." Latent measurement invariance tests revealed scalar invariance of the boredom scale (i.e., TBS-S) across situations of working on difficult versus easy tasks. Furthermore, the prototypical single boredom item "I was bored," which is part of the TBS-S, also demonstrated measurement invariance across the two situations . Thus, boredom experienced while working on difficult versus easy tasks reflects the same latent construct of boredom. The measurement invariance of the single item is an important finding for assessing state test boredom. Use of a single item is typically preferable because it minimizes the risk of disrupting the testing process, as the assessment of boredom itself might interfere with test performance.

Our findings suggest that, due to their similarity, it may not be possible to determine whether boredom results from being overchallenged or underchallenged based solely on its

measurement (i.e., inferring over- or underchallenge from the structure of item responses). Thus, to investigate the potentially different effects of boredom due to over- versus underchallenge on performance, over- and underchallenge need to be assessed alongside boredom. Considering indicators of being over- or underchallenged (e.g., as moderator variables) will also be important in meta-analyses on the relationship between boredom and performance.

In support of our hypotheses, we found weak-to-moderate correlations between components of the TBS-S across working on difficult versus easy tasks. In contrast, correlations between components of the TBS-S within these situations were strong. In line with these findings and with our hypotheses, the correlation of the prototypical single boredom item was also moderate across working on difficult versus easy tasks. Thus, from a correlational point of view, boredom assessments during work on difficult versus easy tasks can be clearly distinguished.

Confirming our expectations, our findings suggest that test takers differ in the extent to which they feel bored when working on difficult versus easy tasks (i.e., when being over- versus underchallenged) – some are primarily bored by difficult, others by easy tasks. Consequently, only limited inferences can be drawn from levels of boredom while working on difficult tasks to levels of boredom while working on easy tasks.

Confirming our expectations, our findings suggest that test takers differ in the extent to which they feel bored when working on difficult versus easy tasks (i.e., when being over- versus underchallenged) - some are primarily bored by difficult tasks, while others are primarily bored by easy tasks. Consequently, only limited inferences can be drawn from levels of boredom experienced during difficult tasks to levels of boredom experienced during easy tasks.

**Differential Effects of Boredom During Difficult versus Easy Tasks: The Role of Cognitive Resources**

Our results were fully consistent with our hypothesis, showing that test boredom while working on difficult tasks significantly reduced cognitive resources (i.e. measured by reaction time deterioration) as well as performance outcomes, whereas test boredom while working on easy tasks showed no significant effects on either construct. Our findings align closely with the assumptions outlined in the abundance hypothesis and with the previous findings of Goetz et al. (2023). However, our findings extend this research significantly.

First, unlike the study by Goetz et al. (2023), which focused on high school students, our study included university students. Thus, our findings attest to the generalizability of the abundance hypothesis across age (12 to 15 years [$M$ = 13.73; $SD$ = 0.44] in the study by Goetz et al., 2023; versus 18 to 38 years [$M$ = 13.73; $SD$ = 0.44] in the present study) and academic context (i.e., high school versus university).

Second, and most importantly, our study demonstrated a core mechanism posited in the abundance hypothesis, as suggested (but not tested) by Goetz et al. (2023). Specifically, the hypothesis posits that the negative effects of test boredom on cognitive performance are likely to be minimal or even non-existent when working on easy tasks (i.e., being underchallenged). Even though some resources are devoted to processing boredom, resources remain abundant for completing the task. In contrast, when working on difficult tasks (i.e., being overchallenged), processing boredom leads to a reduction in available cognitive resources, which has adverse effects on performance. In such situations, resources consumed by processing boredom are no longer available for solving the task, even though they would be needed for successful task completion. In our study, this lack of available cognitive resources in overchallenging situations

was reflected in longer RTs (i.e., slower reactions). Slow reactions can be assumed to reflect reduced cognitive performance.

Our understanding of the mechanisms underlying the abundance effect aligns with core propositions of cognitive load theory (CLT; Sweller, 2023). Although recent work has argued for fully integrating emotion into CLT (Plass & Kalyuga, 2019), few studies have directly addressed such an integration (e.g., Fraser et al., 2015; Knörzer et al., 2016). Our work contributes to closing this gap. In the present study, a larger proportion of working memory capacity is presumably occupied by working on the difficult tasks, as compared to working on the easy tasks. In both conditions, boredom causes additional, extraneous cognitive load (Sweller, 2023) due to mind-wandering, inattention, or conscious effort dedicated to maintain concentration (Eastwood et al., 2012; Goetz et al., 2024), which are typical of the experience of boredom. However, if the cognitive load imposed by the task (i.e., intrinsic load) is higher, then the additional cognitive load generated by boredom more likely results in working memory overload (Paas et al., 2003), because it further burdens already strained cognitive capacities. Therefore, it is plausible that the additional cognitive load resulting from boredom is more likely to impact performance when tasks are cognitively demanding as compared to easy tasks. Interestingly, CLT does not predict that emotion necessarily reduces cognitive resources. It has been argued that emotion can facilitate cognitive performance by improving the availability of working memory (Plass & Kalyuga, 2019) and enhancing motivation and effort (Heidig et al. 2015; Isen and Reeve 2005). According to our results, however, this does not appear to be the case for boredom.

However, CLT does not specifically address the different situational antecedents that might explain instances where additive cognitive load does not result in performance

decrements. As such, the abundance hypothesis complements CLT by proposing that performance remains unaffected as long as cognitive resources are abundant enough to accommodate both the task demands and the additional burden of boredom.

## Limitations

A number of limitations of the present study should be noted and can inform directions for future research. First, the sample size of Study 1 was relatively small for testing measurement invariance. Future studies could focus on larger samples. To test generalizability, samples in future studies could also be more diverse in terms of age and cultural background. Future studies might also examine additional demographic characteristics, including race/ethnicity and disability status.

Second, in terms of the mechanisms underlying the abundance hypothesis, we have focused on the role of cognitive resources. Although these resources can be expected to play a central role in the effects of boredom on cognitive performance, other potentially relevant variables could be investigated in future studies, such as reduced motivation and sub-optimal use of learning and working strategies due to the experience of boredom. Boredom might lead to the use of suboptimal strategies, which, in turn, could reduce performance when working on difficult tasks. However, when working on easy tasks, the use of suboptimal strategies may still be sufficient to complete the tasks successfully.

Third, our experiment did not include a condition with an optimal fit between task difficulty and individual ability. In such a condition, low levels of both over- and underchallenge - and consequently low levels of boredom - might be expected (Csikszentmihalyi, 1975/2000). Adding such a condition in future studies on the potentially differential effects of boredom during work on difficult versus easy tasks could help strengthen the argument that boredom

indeed arises from suboptimal levels of control (see Krannich et al., 2019, for initial work on this topic). Technically, such an “optimal challenge” condition could be implemented through computerized adaptive (tailored) testing (CAT; e.g., Asseburg & Frey, 2013).

Fourth, in Study 2, measuring reaction time after the achievement task did not fully align with our assumption that cognitive resources mediate boredom’s effect on achievement, which is a significant limitation. We avoided measuring reaction time during the task to prevent disrupting the induction of over-/underchallenge and boredom. Nevertheless, future work on the abundance hypothesis should attempt to align the timing of the assessment of working memory load with theory by administering it during task performance rather than afterwards.

Fifth, our studies assessed boredom during low-stakes testing. Future research should aim to replicate our findings in high-stakes contexts. Assessing boredom during high-stakes tests may pose challenges, as interruptions by administering questionnaires could affect test outcomes for some students. However, administering boredom assessments immediately after the test could mitigate this issue.

## Implications for Research

Our study has several implications for research on test boredom. First, since the experimental conditions in Study 1 (i.e., working on difficult versus easy tasks) produced similar levels of boredom (i.e., similar levels of over- and underchallenge boredom), future studies could adopt this design. For example, using this design, researchers could explore differential effects of test boredom on variables not assessed in our study, such as creativity and study behavior (see Krannich et al., 2025; Stempfer et al., 2025).

Second, our studies have focused on the structure and effects of test boredom during difficult versus easy tasks. It may be helpful to also assess other emotions in these situations,

such as anxiety, anger, hopelessness, enjoyment, hope, and pride (see Pekrun et al., 2023). Such studies would enable examination of the joint effects of boredom and other emotions on performance outcomes. For example, an important research question is whether specific combinations of emotional experiences (e.g., boredom and anxiety) are especially detrimental in overchallenging situations. Furthermore, this approach would allow analysis of differences in the component structures of test boredom and other test-related emotions, such as test anxiety (e.g., Lange & Zickfeld, 2021). There may be overlaps between specific components of different test emotions - for instance, the motivational components of both test boredom and test anxiety might include tendencies to escape from the situation (Goetz et al., 2024). Thus, researchers could investigate whether similar components across different emotions (e.g., the motivational component) explain the negative effects of different emotional experiences on achievement in overchallenging situations.

Third, regarding the structural validity of test boredom in situations involving difficult versus easy tasks, we have focused on state test boredom. Future studies could also examine the structure of trait test boredom in this context. Trait-like items could be used that assess boredom typically experienced when working on easy or difficult tasks (e.g., “How strongly do you typically experience boredom while working on very easy [very difficult] tasks?”).

Furthermore, the abundance hypothesis and the underlying mechanisms could be investigated beyond academic contexts, such as the workplace, sports, and the performing arts. For instance, it is plausible to assume that physical education activities can lead to perceptions of over- and underchallenge, similar to test situations in education. This could prompt emotional experiences similar to test boredom and affect performance outcomes (e.g., Velasco & Jorda, 2020). For example, during endurance sessions, runners could be verbally prompted via an app

on their headsets to rate their level of boredom and the extent to which they feel over- or underchallenged at that moment; responses could be given verbally via the headset microphone.

Finally, the differential effects of boredom on performance when working on difficult versus easy tasks should be considered in the context of large-scale assessments. For example, international student assessments such as PISA (Programme for International Student Assessment; OECD, 2019) may underestimate students' actual competencies due to the negative impact of boredom on test performance, especially when students work on difficult tasks. A study by Asseburg and Frey (2013) showed that average item difficulties exceeded students' individual abilities in about two-thirds of the sample in the PISA 2013 assessment, suggesting that overchallenge boredom may have biased test results. Accordingly, investigating experiences of boredom while working on the tasks - particularly on difficult tasks - could allow achievement levels to be adjusted.

## Implications for Practice

Our study also has several implications for practice. First, if teachers notice that a student is often bored while working on tests, it is crucial to identify the reasons for their boredom. A direct conclusion might be drawn from the student's test score: A low score could indicate that the tasks are too difficult for the student, while a high score could suggest that the tasks are too easy. However, in both cases, it is important to note that factors beyond ability, such as low interest in the topic, might also contribute to test boredom. While boredom during difficult tasks can be assumed to be particularly detrimental to performance outcomes, actions are appropriate in both scenarios. In the case of boredom arising from difficult tasks, tutoring might be helpful. In contrast, for boredom caused by easy tasks, additional challenges or enrichment opportunities might be beneficial (for strategies to regulate boredom, see also Goetz et al., 2024).

Second, parents should also be aware of their children's test boredom - and, of course, other emotions experienced during test-taking. Talking with children about overchallenge and underchallenge as potential causes of boredom might be helpful and could initiate appropriate actions (e.g., organizing tutoring or peer support when the difficulty exceeds the child's individual abilities).

Third, students themselves should also be aware of their possible test boredom. This can be encouraged if teachers and parents talk with them about their experiences of boredom - not only related to tests but also to other school-related situations (e.g., classroom activities, independent work at school, or homework). Discussing the potential causes of boredom can help students become more aware of it and, in turn, work to prevent this feeling (Stockinger et al., 2025). Boredom caused by excessive demands, in particular, seems especially harmful - as our study has shown that it is not only an unpleasant feeling but also a drain on the resources needed for successful performance.

Finally, boredom experienced during tests and other educational situations should be a topic in teacher education. In particular, strategies for preventing boredom should be taught, such as individualization to prevent students from being either overchallenged or underchallenged. It is important for educators to understand the negative effects of boredom, including its impact on achievement outcomes, particularly in cases where boredom arises from working on difficult tasks. This is especially important given the common but empirically unsupported claim that boredom in school has its benefits (see Vodanovich, 2003).

**Appendix A: Test Boredom Scale – State (TBS-State)**

| Nr. | English | German |
|---|---|---|
| 1 (a) | I was bored. | Ich war gelangweilt. |
| 2 (a) | The math tasks seemed monotonous and dull to me from boredom. | Vor Langeweile erschienen mir die Matheaufgaben eintönig und grau. |
| 3 (a) | The math tasks bored me to death. | Die Matheaufgaben haben mich zu Tode gelangweilt. |
| 4 (c) | I was so bored that I found myself daydreaming. | Ich habe mich so gelangweilt, dass ich mich beim Tagträumen ertappt habe. |
| 5 (c) | My mind was wandering. | Ich war mit den Gedanken woanders. |
| 6 (c) | I couldn't focus on the math tasks because I was so bored. | Ich konnte mich nicht auf die Matheaufgaben konzentrieren, weil ich so gelangweilt war. |
| 7 (m) | I would have preferred not to start at all with the math tasks because of boredom. | Vor lauter Langeweile hätte ich am liebsten gar nicht erst mit den Matheaufgaben angefangen. |
| 8 (m) | I constantly looked at my watch because time did not pass. | Ich habe ständig auf die Uhr geschaut, weil die Zeit nicht verging. |
| 9 (m) | I would have liked to leave the classroom out of boredom. | Aus Langeweile hätte ich die Klassenarbeit am liebsten verlassen. |
| 10 (p) | I was yawning because I was so bored. | Vor Langeweile musste ich gähnen. |
| 11 (p) | I was so bored that I was tired. | Ich langweilte mich so, dass ich ganz matt wurde. |
| 12 (p) | I could hardly keep awake because of boredom. | Vor Langeweile konnte ich mich kaum wach halten. |

*Note.* a = affective, c = cognitive, m = motivational, p = physiological component of boredom. The German version was used in the present study.

## References

Adam, C. A. (2024, January 23). *Unter- und Überforderungslangeweile, Erschöpfung und Leistung*. https://doi.org/10.17605/OSF.IO/5H34W

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*(1), 92–104.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

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

Bieleke, M., Wolff, W., Martarelli, C. S., Artak, C. E. T., Asani, N., Baillifard, A., Bertrams, A., Brielmann, A., Caldwell, L. L., Chan, C. S., Coppin, G., Danckert, J., Dang, V., Daniels, L., Dayan, P., Drody, A., Elpidorou, A., Erdemli, A., Fischer, U., Goetz, T., … Yakobi, O. (2024). Overview of current directions in boredom research. In M. Bieleke, W. Wolff, & C. S. Martarelli, *The Routledge International Handbook of Boredom* (pp. 382–391). Routledge.

Camacho-Morles, J., Slemp, G. R., Pekrun, R., Loderer, K., Hou, H., & Oades, L. G. (2021). Activity achievement emotions and academic performance: A meta-analysis. *Educational Psychology Review, 33*, 1051–1095. https://doi.org/10.1007/s10648-020-09585-3

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Csikszentmihalyi, M. (1975/2000). *Beyond boredom and anxiety: Experiencing flow in work and play* (2nd ed.). Jossey Bass.

Daniels, L. M., Stupnisky, R. H., Pekrun, R., Haynes, T. L., Perry, R. P., & Newall, N. E. (2009). A longitudinal analysis of achievement goals: From affective antecedents to emotional effects and achievement outcomes. *Journal of Educational Psychology, 101,* 948–963. https://doi.org/10.1037/ a0016096

Daschmann, E. C., Goetz, T., & Stupnisky, R. H. (2011). Testing the predictors of boredom at school. Development and validation of the Precursors to Boredom Scales. *British Journal of Educational Psychology*, *81*, 421–440. https://doi.org/10.1348/000709910X526038.

Claros-Salinas, D., Dittmer, N., Neumann, M., Sehle, A., Spiteri, S., Willmes, K., Schoenfeld, M. A., & Dettmers, C. (2013). Induction of cognitive fatigue in MS patients through cognitive and physical load. *Neuropsychological Rehabilitation, 23*(2), 182–201. https://doi.org/10.1080/09602011.2012.726925

Dueber, D. (2025). *dmacs: Measurement nonequivalence effect size calculator* (Version 0.1.0.9002) [Computer software]. https://github.com/ddueber/dmacs

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

Enders, C. K. (2022). *Applied missing data analysis* (2nd ed.). The Guilford Press.

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18. https://doi.org/10.18637/jss.v048.i04

Federal Statistical Office [Statistisches Bundesamt]. (2020). *Schnellmeldungsergebnisse zu Schülerinnen und Schülern der allgemeinbildenden und beruflichen Schulen – Schuljahr 2019/20* [Preliminary results of general and vocational school students: 2019–20 academic year] [Statistical report]. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Schulen/Publikationen/Downloads-Schulen/schnellmeldung-schueler-5211003208004.htm

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. Wiley.

Fraser, K. L., Ayres, P., & Sweller, J. (2015). Cognitive load theory for the design of medical simulations. *Simulation in Healthcare, 10*(5), 295–307. https://doi.org/10.1097/SIH.0000000000000097

Fries, J., & Yanagida, T. (2025, October 14). *Test boredom due to over- versus underchallenge: The same emotion but different effects on performance*. https://doi.org/10.17605/OSF.IO/WS23C

Goetz, T., Bieleke, M., Yanagida, T., Krannich, M., Roos, A.-L., Frenzel, A. C., Lipnevich, A. A., & Pekrun, R. (2023). Test boredom: Exploring a neglected emotion. *Journal of Educational Psychology, 115*(7), 911–931. https://doi.org/10.1037/edu0000807

Goetz, T., Cronjaeger, H., Frenzel, A. C., Lüdtke, O., & Hall, N. C. (2010). Academic self-concept and emotion relations: Domain specificity and age effects. *Contemporary Educational Psychology, 35*, 44–58. https://doi.org/10.1016/j.cedpsych.2009.10.001

Goetz, T., Hall, N. C., & Krannich, M. (2019). Boredom. In K. A. Renninger & S. E. Hidi (Eds.), *The Cambridge handbook on motivation and learning* (pp. 465–486). Cambridge University Press.

Goetz, T., Preckel, F., Pekrun, R., & Hall, N. C. (2007). Emotional experiences during test taking: Does cognitive ability make a difference? *Learning and Individual Differences, 17*(1), 3–16. https://doi.org/10 .1016/j.lindif.2006.12.002

Goetz, T., Stempfer, L., Pekrun, R., van Tilburg, W. A. P., & Lipnevich A. A. (2024). Academic boredom. In M. Bieleke, W. Wolff, & C. Martarelli, *The Routledge International Handbook of Boredom* (pp. 225–249). Routledge.

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Fischbach, A., Keller, U., & Preckel, F. (2014). 'My questionnaire is too long!' The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology, 39*, 188–205. https://doi.org/10.1016/j.cedpsych.2014.04.002

Heidig, S., Müller, J., & Reichelt, M. (2015). Emotional design in multimedia learning: Differentiation on relevant design features and their effects on emotions and learning. *Computers in Human Behavior, 44*, 81–95. https://doi.org/10.1016/j.chb.2014.11.009

Isen, A. M., & Reeve, J. (2005). The influence of positive affect on intrinsic and extrinsic motivation: Facilitating enjoyment of play, responsible work behavior, and self-control. *Motivation and Emotion, 29*(4), 295–323. https://doi.org/10.1007/s11031-006-9019-8

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Knörzer, L., Brünken, R., & Park, B. (2016). Facilitators or suppressors: Effects of experimentally induced emotions on multimedia learning. *Learning & Instruction, 44*, 97–107. https://doi.org/10.1016/j.learninstruc.2016.04.002

Krannich, M., Calik, B., Goetz, T., Ullrich, A.-L., & Lipnevich, A. A. (2025). Investigating the relationship between boredom and creativity: The role of academic challenge. *Education Sciences, 15*(3), 330. https://doi.org/10.3390/educsci15030330

Krannich, M., Goetz, T., Lipnevich, A. A., Bieg, M., Roos, A.-L., Becker, E. S., & Morger, V. (2019). Being over- or underchallenged in class: Effects on students' career aspirations via self-concept and boredom. *Learning and Individual Differences, 69*, 206–218. https://doi.org/10.1016/j.lindif.2018.10.004

Lange, J., & Zickfeld, J. H. (2021). Emotions as overlapping causal networks of emotion components: Implications and methodological approaches. *Emotion Review, 13*(2), 157–167. https://doi.org/10.1177/1754073920988787

Li, C., Feng, E., & Li, S. (2025). Boredom and achievement in L2 learning: A meta-analysis. *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2024-0266

Liang, X., & Luo, Y. (2020). A comprehensive comparison of model selection methods for testing factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(3), 380–395. https://doi.org/10.1080/10705511.2019.1649983

Little, T. D. (2013). *Longitudinal structural equation modeling.* Guilford Press.

Little, T. D., Siegers, D. W., & Card. N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13,* 59–72. https://doi.org/10.1207/s15328007sem1301_3

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Muthén, L. K., & Muthén, B. O. (1998–2018). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining $R^2$ from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nett, U., Goetz, T., & Daniels, L. (2010). What to do when feeling bored? Students‘ strategies for coping with boredom. *Learning and Individual Differences, 20*, 626–638. https://doi.org/10.1016/j.lindif.2010.09.004

Neumann, M., Sterr, A., Claros-Salinas, D., Gütler, R., Ulrich, R., & Dettmers, C. (2014). Modulation of alertness by sustained cognitive demand in MS as surrogate measure of fatigue and fatigability. *Journal of the Neurological Sciences, 340*, 178–82. https://doi.org/10.1016/j.jns.2014.03.024

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *The Journal of Applied Psychology*, *96*, 966–980. https://doi.org/10.1037/a0022955

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence.

*Organizational Research Methods*, *22*, 678–709. https://doi.org/10.1177/1094428118761122

Organization for Economic Cooperation and Development. (2019). PISA 2018 mathematics framework. In *PISA 2018 assessment and analytical framework*. https://doi.org/10.1787/13c8a22c-en

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–54). Cambridge University Press.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*, 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pekrun, R. (2018). Control-value theory: A social-cognitive approach to achievement emotions. In G. A. D. Liem & D. M. McInerney (Eds.), *Big theories revisited 2: A volume of research on sociocultural influences on motivation and learning* (pp. 162–190). Information Age Publishing.

Pekrun, R. (2021). Self-appraisals and emotions: A generalized control-value approach. In T. Dicke, F. Guay, H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds). *Self – a multidisciplinary concept* (pp. 1–30). Information Age Publishing.

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control-value antecedents and performance outcomes of

a neglected emotion. *Journal of Educational Psychology, 102*(3), 531–549. https://doi.org/10.1037/a0019243

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, *37*(2), 91–105. https://doi.org/10.1207/S15326985EP3702_4

Pekrun, R., Hall, N. C., Goetz, T., & Perry, R. P. (2014). Boredom and academic achievement: Testing a model of reciprocal causation. *Journal of Educational Psychology, 106*(3), 696–710. https://doi.org/10.1037/a0036006

Pekrun, R., Marsh, H. W., Elliot, A. J., Stockinger, K., Perry, R. P., Vogl, E., Goetz, T., van Tilburg, W. A. P., Lüdtke, O., & Vispoel, W. P. (2023). A three-dimensional taxonomy of achievement emotions. *Journal of Personality and Social Psychology, 124*(1), 145–178. https://doi.org/10.1037/pspp0000448

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

Prenzel, M., & Blum, W. (2007). *Entwicklung eines Testverfahrens zur Überprüfung der Bildungsstandards in Mathematik für den mittleren Schulabschluss*. [Test development for the examination of the Educational Standards for the Intermediate School Leaving Certificate in mathematics]. Kiel: IPN.

Questionnaire Master (2024, 29. April). https://intelligenz.bpnh.de/test/6584c847b5eb8 and https://intelligenz.bpnh.de/test/6595a726788f8

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Scherer, K. R. (2000). Emotions as episodes of subsystems synchronization driven by nonlinear appraisal processes. In M. D. Lewis & I. Granic (Eds.), *Emotion, development, and self-organization* (pp. 70–99). Cambridge University Press.

Scherer, K. R., & Moors, A., (2019). The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology, 70,* 719–745. https://doi.org/10.1146/annurev-psych-122216-011854

Shadish, W., Cook, T., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin.

Stempfer, L., Goetz, T., Voracek, M., van Tilburg, W. A. P., Tran, U. S., & Pekrun, R. (2025). *Boredom and performance: A meta-analysis across contexts*. Manuscript submitted for publication.

Sweller, J. (2023). The development of cognitive load theory: Replication crises and incorporation of other theories can lead to theory expansion. *Educational Psychology Review*, *35*, 95. https://doi.org/10.1007/s10648-023-09817-2

Trafimow, D. (2016). The attenuation of correlation coefficients: A statistical literacy issue. *Teaching Statistics, 38*(1), 25–28. https://doi.org/10.1111/test.12087

Tze, V. M. C., Daniels, L. M. & Klassen, R. M. (2016). Evaluating the relationship between boredom and academic outcomes: A meta-analysis. *Educational Psychological Review, 28*, 119–144. https://doi.org/10.1007/s10648-015-9301-y

UNESCO Institute for Statistics. (2012). *International standard classification of education: ISCED 2011*. UNESCO.

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*, 486–492. https://doi.org/10.1080/17405629.2012.686740

Velasco, F., & Jorda, R. (2020). Portrait of boredom among athletes and its implications in sports management: A multi-method approach. *Frontiers in Psychology, 11*, 831. https://doi.org/10.3389/fpsyg.2020.00831

Vodanovich, S. J. (2003). On the possible benefits of boredom: A neglected area in personality research. *Psychology and Education-An Interdisciplinary Journal, 40*, 28–33.

Wechsler, D. (2006). *Wechsler-Intelligenztest für Erwachsene: WIE; Manual; Übersetzung und Adaption der WAIS-III*. Harcourt Test Services.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale--Third Edition (WAIS-III)*. The Psychological Corporation.

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/10.1037/rev0000097

Zeidner, M. (1998). *Test anxiety: The state of the art*. Plenum.

**Table 1**

*Study 1: Measurement Invariance Across Difficult and Easy Part of the Test*

| Model | $\chi^2$ | df | CFI | TLI | RMSEA | SRMR | BIC |
|---|---|---|---|---|---|---|---|
| | | | *First-order factor model* | | | | |
| Configural invariance | 334.84 | 207 | .940 | 0.921 | 0.054 | 0.051 | 9621.80 |
| Metric invariance | 342.04 | 215 | .941 | 0.924 | 0.053 | 0.054 | 9599.74 |
| Scalar invariance | 351.19 | 223 | .940 | 0.926 | 0.053 | 0.054 | 9563.76 |
| | | | *Second-order factor model* | | | | |
| Configural invariance | 259.64 | 234 | .941 | 0.931 | 0.051 | 0.060 | 9537.27 |
| Metric invariance | 362.07 | 237 | .942 | 0.932 | 0.050 | 0.060 | 9525.04 |
| Sclar invariance | 374.09 | 248 | .941 | 0.935 | 0.049 | 0.060 | 9475.25 |

*Note*. $N = 208$. Residual covariances between the same items in the difficult and easy parts of the test were included (Little, 2013); five additional residual covariances were included (for a detailed description, see online supplemental material [S1]).

**Table 2**

*Study 1: Latent Means, Standard Deviation, and Correlation Coefficients*

| Component of Boredom | *M* | *SD* | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Difficult Part | | | | | | |
| 1. Affective | 1.60 | 0.91 | | | | | | | | |
| 2. Cognitive | 1.57 | 0.78 | .88 | | | | | | | |
| 3. Motivational | 1.45 | 0.78 | .88 | .98 | | | | | | |
| 4. Physiological | 1.37 | 0.71 | .80 | .93 | .94 | | | | | |
| | | | | Easy Part | | | | | | |
| 5. Affective | 1.53 | 0.82 | .33 | .35 | .43 | .33 | | | | |
| 6. Cognitive | 1.55 | 0.69 | .26 | .30 | .40 | .32 | .91 | | | |
| 7. Motivational | 1.42 | 0.67 | .32 | .36 | .42 | .36 | .93 | .88 | | |
| 8. Physiological | 1.31 | 0.59 | .19 | .27 | .30 | .25 | .87 | .95 | .94 | -- |
| Difficult - Easy Part | Δ*M* | *SE* | *p* | *d* | | | | | | |
| Affective | 0.07 | 0.07 | .316 | 0.08 | | | | | | |
| Cognitive | 0.02 | 0.06 | .761 | 0.03 | | | | | | |
| Motivational | 0.03 | 0.06 | .612 | 0.04 | | | | | | |
| Physiological | 0.06 | 0.06 | .312 | 0.09 | | | | | | |

*Note*. $N = 208$. To ensure the same metrics for the items and thus an equivalent interpretation of the mean scores across items, effect coding (Little et al., 2006) was used for scale setting of latent variables and model identification. The effect coding method allows interpreting latent

means on a non-arbitrary metric of the measured indicators, that is, from 1 (*not at all true*) to 5 (*completely true*), without standardizing the factors (fixed factor method) or relying on single reference indicators for each factor (marker variable method). Δ*M* = Unstandardized latent mean difference, *SE* = Standard error; *d* = latent Cohen's *d*. All latent correlation coefficients were statistically significant at α = .05.

**Table 3**

*Study 2: Descriptive Statistics*

| Variable | Difficult Part | | | Easy Part | | |
|---|---|---|---|---|---|---|
| | *M* | *Md* | *SD* | *M* | *Md* | *SD* |
| Boredom | 3.56 | 4.00 | 1.25 | 4.02 | 4.00 | 0.96 |
| Reaction time deterioration | 2.59 | -1.56 | 47.60 | 16.43 | 11.62 | 54.46 |
| Working memory performance | 6.55 | 7.00 | 1.29 | 6.95 | 7.00 | 0.99 |

*Note.* $N = 132$. Test boredom was assessed using a five-point response scale ranging from 1 (*completely disagree*) to 5 (*completely agree*). Reaction time deterioration (i.e., the difference in reaction time after minus before the manipulation) ranged from -131.63 to 371.29 when working on the difficult manipulation trials and from -342.44 to 226.32 when working on the easy manipulation trials. Working memory performance ranged from 3 to 8 when working on the difficult manipulation trials and from 4 to 8 when working on the easy manipulation trials.

**Table 4**

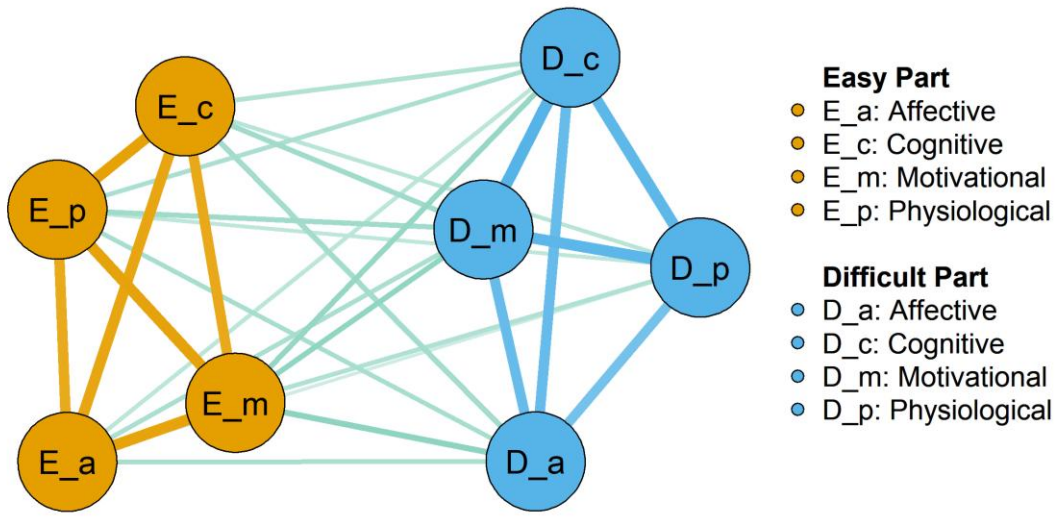
*Summary Statistics for Linear Mixed-Models*

| Boredom | β [95% CI] | *t* | *p* | $R^2_m$ |
|---|---|---|---|---|
| | Working memory performance | | | |
| Boredom – Difficult Part | -0.18* [-0.34, -0.01] | -2.03 | .044 | .03 |
| Boredom – Easy Part | -0.08 [-0.25, 0.09] | -0.93 | .353 | .01 |
| | Reaction time deterioration | | | |
| Boredom - Difficult Part | 0.19* [0.03, 0.36] | 2.30 | .023 | .04 |
| Boredom - Easy Part | 0.07 [-0.10, 0.24] | 0.83 | .407 | .01 |

*Note.* $N = 132$. * $p < .05$. Each row corresponds to one regression model. $R^2_m$ represents marginal *R*-squared, i.e., the amount of variance in the dependent variable explained by the fixed-effects portion of the model (Nakagawa & Schielzeth, 2013). Reaction time was assessed in ms, reaction time deterioration is the difference of reaction time post minus pre manipulation.

**Figure 1**

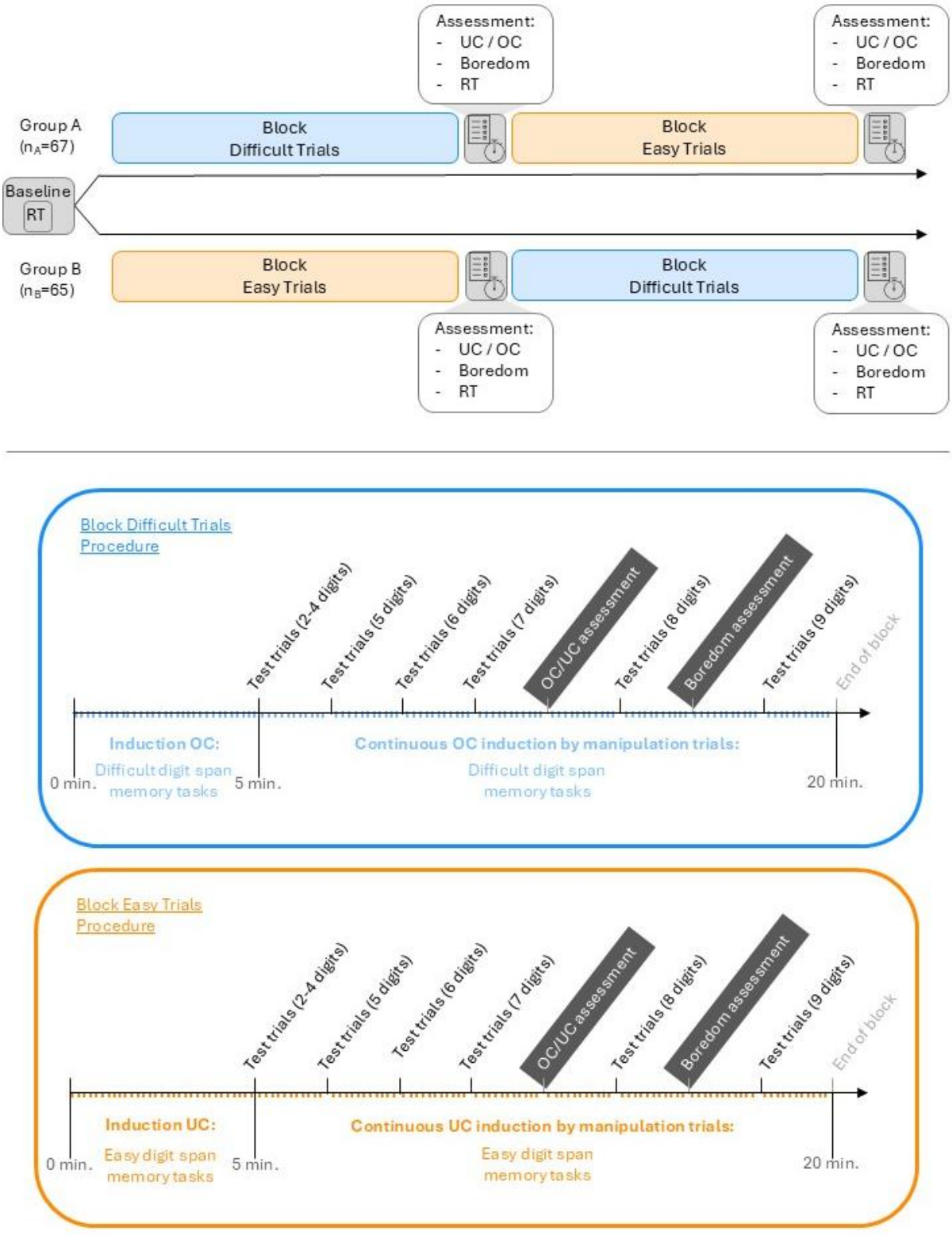
*Network Visualisation of the Correlation Matrix*




*Note*. The correlation matrix is visualized as a network, where orange nodes represent components of boredom in the easy part of the test (i.e., underchallenge boredom) and blue nodes represent components of boredom in the difficult part (i.e., overchallenge boredom). The strength of the correlations is indicated by the thickness of the lines (i.e., stronger correlations are represented by thicker lines) and the distance between nodes (i.e., nodes with stronger correlations are placed closer together).

**Figure 2**

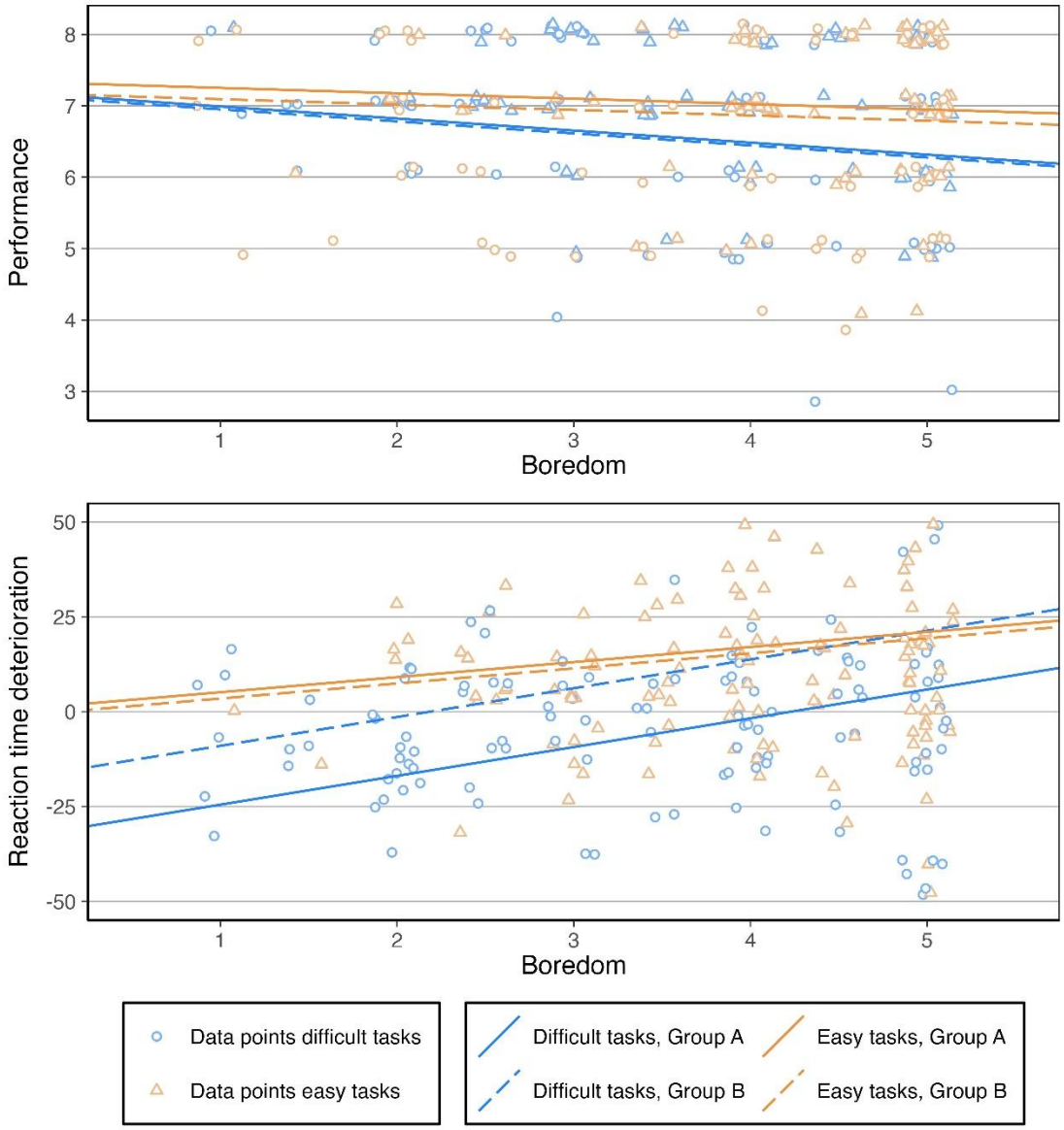
*Procedure of Study 2 and of Over- and Underchallenging Experimental Blocks.*



*Note.* The upper part of the figure shows the procedure on the experimental level, the lower part of the figure shows the detailed procedure within the difficult and easy blocks. Top: upper and lower black arrows represent the timelines for Group A and B, including the difficult (blue boxes) and easy (orange boxes) experimental blocks and the post-block assessments (grey boxes) of over- and underchallenge, boredom and reaction time. Bottom: "Block Difficult Trials Procedure" and "Block Easy Trials Procedure" show the detailed timeline of the difficult and easy blocks. Manipulation trials inducing over- and underchallenge were continuously presented (blue and orange short vertical lines). Test trials were interspersed to assess working memory performance, as well as items assessing feelings of under- or overchallenge and boredom (black arrow). OC = overchallenge, UC = underchallenge, RT = reaction time.

**Figure 3**

*Scatter Plots for Hypothesis 3*




*Note.* Each panel presents a scatter plot containing two regression models: one for data points corresponding to working on difficult tasks and another for data points corresponding to working on easy tasks. The data points are slightly jittered to make them distinguishable. The models are displayed without *z*-standardization of the variables to facilitate interpretation on their native

scales. Each model includes two regression lines with identical slopes but varying intercepts, corresponding to the random-effects group variable (i.e., Groups A and B). The y-axis in the top panel (i.e., performance) indicates each participants' number of correctly solved items in the Digit Span subtest from the WAIS-III. To enhance readability, the y-axis in the bottom panel (i.e., reaction time deterioration) was truncated to the interval ranging from -50 to 50 ms, so values outside this range are not displayed in the plot (9.47 percent of values are above the upper threshold and 2.27 percent are below the lower threshold). For the boredom assessment, participants responded to the single item "*I am bored*" using a 5-point rating scale ranging from 1 (*completely disagree*) to 5 (*completely agree*).