Test Boredom: Exploring a Neglected Emotion

Thomas Goetz[1], Maik Bieleke[1], Takuya Yanagida[1], Maike Krannich[2], Anna-Lena Roos[3],
Anne C. Frenzel[4], Anastasiya A. Lipnevich[5], and Reinhard Pekrun[6,7,4],

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

[2]Department of Psychology, University of Zurich

[3]Institute for Research and Development of Collaborative Processes, School of Applied Psychology, University of Applied Sciences and Arts Northwestern Switzerland

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

[5]Queens College and the Graduate Center, The City University of New York

[6]Department of Psychology, University of Essex

[7]Institute for Positive Psychology and Education, Australian Catholic University, North Sydney

Date of acceptance: 31 March 2023

Citation for this article:

Goetz, T., Bieleke, M., Yanagida, T., Krannich, M., Roos, A.-L., Frenzel, A. C., Lipnevich, A. A., & Pekrun, R. (in press). Test boredom: Exploring a neglected emotion. *Journal of Educational Psychology*.

Author Note

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

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

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

Maike Krannich: https://orcid.org/0000-0001-9239-3283

Anna-Lena Roos: https://orcid.org/0000-0002-7853-0330

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

Anastasiya A. Lipnevich: https://orcid.org/0000-0003-0190-8689

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

We have no conflicts of interest to disclose. This research was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft) awarded to Reinhard Pekrun (PE 320/11-1). The data used for this research and the findings have not previously been disseminated. The data, research materials, and analysis code are available at [link will be provided here].

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 emotion of boredom has sparked considerable interest in research on teaching and learning, but boredom during tests and exams has not yet been examined. Based on the control-value theory of achievement emotions, we hypothesized that students may experience significant levels of boredom during testing (“test boredom”; H1), and that test boredom may be significantly related to theoretically hypothesized antecedents (control and value appraisals; H2) and outcomes (performance; H3). We further hypothesized that test boredom was more detrimental when students felt overchallenged during the test than when they felt underchallenged (‘abundance hypothesis’; H4). We tested these hypotheses in two studies (Study 1: *N*=208 8th graders; 54% female; Study 2: *N*=1,612 5th-10th graders, 47% female) using both trait and state measures of test boredom in mathematics and their proposed antecedents and outcomes. In support of H1, participants reported statistically significant levels of boredom during tests. Further, the relations of test boredom with its control and value antecedents (i.e., being over- or underchallenged, facets of value) were in line with our assumptions (H2). In support of H3, test boredom was significantly negatively related to academic achievement (grades). In line with H4, test scores were negatively related to boredom due to being overchallenged but unrelated, or even positively related, to boredom due to being underchallenged. Directions for future research on test boredom as well as practical implications are outlined.

Keywords: boredom, test, achievement, mathematics, control-value theory

Educational Impact and Implications Statement

Our research shows that boredom occurs during achievement tests, and that the level of test boredom can be quite high. Primary causes of test boredom seem to be over- or underchallenge as well as perceived low importance of the test. Furthermore, test boredom appears to have negative effects on academic outcomes, particularly boredom that results from being overchallenged. Test boredom could be mitigated by designing tests such that over- or underchallenge are reduced, and by increasing the perceived intrinsic value of tests.

Test Boredom: Exploring a Neglected Emotion

The last 15 years have seen a strong increase in studies on boredom in the context of learning and achievement (Goetz et al., 2019). A crucial reason for this increasing interest is the accumulating empirical evidence on its negative effects on learning and achievement outcomes, including students' motivation, learning behavior, grades, and career aspirations (e.g., Pekrun et al., 2014; for meta-analyses, see Camacho-Morles et al., 2021; Tze et al., 2016). Due to these consistent negative effects, research on the antecedents of boredom (e.g., being over- or underchallenged; Daschmann et al., 2011) and on how to cope with it (e.g., by trying to enhance the perceived value of the situation; e.g., Nett et al., 2010) has been initiated. This research typically focuses on boredom experienced in class (e.g., in high schools and universities), during individual learning situations (e.g., when preparing for an exam), and while doing homework (Goetz et al., 2019). To this end, various measures of boredom have been developed and published (see Bieleke et al., 2021, and the review by Vodanovich & Watt, 2016).

Considering the high level of attention to academic boredom, it is intriguing that no single study exists with an explicit focus on boredom experienced in test situations, despite the high prevalence of tests and exams in any academic context. A key reason for why test boredom has been neglected might be that it is counter-intuitive to think of tests to ever be boring. This intuition is in line with the propositions of Pekrun's (2006, 2018, 2021) control-value theory (CVT) of achievement emotions. First, tests are typically seen as inherently high in value (Pekrun et al., 2004) which, according to the CVT, should lead to reduced levels of boredom. Second, tests, if well-designed, should include tasks with a level of difficulty appropriate to the ability level of the individuals being tested (Wainer, 2000). According to the CVT, having an adequate level of control should also preclude boredom (Pekrun et al., 2023; see also Westgate & Wilson, 2018).

However, upon second view, one realizes that some tests may in fact have rather low value for certain students. This might particularly be true for low-stakes testing which has proliferated in recent years. Thus, it can be assumed that the core antecedents of boredom, namely, low value and inadequate levels of control, can also be present during tests (Asseburg & Frey, 2013). In this study, we drew upon these theoretical assumptions and investigated how strongly boredom was experienced during a low-stakes test situation and whether it was related in theoretically plausible ways to its assumed antecedents. Further, to show the potential practical importance of test boredom, we investigated its negative relations with academic achievement (i.e., test scores and grades) as proposed by CVT. We examined these relations using both trait and state assessments to capture both habitual (i.e., trait-like) and real-time (i.e., state) experiences of test boredom and their links to corresponding trait and state variables. Ultimately, we wanted to open a new field of research into test boredom by offering initial evidence of theoretical and practical relevance of this construct.

Test Boredom – Definition

To conceptualize boredom, we use the component process model of emotions (Scherer, 2000; Scherer & Moors, 2018), which suggests that individuals’ emotions are best understood in terms of their underlying processes. From this perspective, boredom can be defined as a unique emotional process consisting of four components: affective (unpleasant, aversive feeling), cognitive (altered perceptions of time, mind wandering), motivational (desire to withdraw from the current situation), and physiological/expressive (low arousal, yawning, looking tired; Goetz et al., 2019; Pekrun et al., 2010, 2014). The term ‘academic boredom’ refers to boredom experienced in learning and achievement situations (Pekrun et al., 2002). According to the specific learning context to which boredom is related, academic boredom can be either class-related, learning-related (including homework), or test-related. Thus, test boredom is a subtype of academic boredom.

Similarly to other types of boredom, test boredom can be conceptualized as a trait or as a state. This distinction is in line with research on test anxiety, which has traditionally distinguished between trait and state test anxiety (Zeidner, 1998), as well as with previous research on academic boredom. For example, in the Achievement Emotions Questionnaire (AEQ), class- and learning-related boredom can be captured as trait or state constructs (Pekrun et al., 2011). Consistently with the differentiation of trait and state boredom in the AEQ, trait test boredom is defined as habitual boredom in test situations, that is, boredom that recurs across test situations and over time. State test boredom, on the other hand, is a current experience of boredom during a given test. Regarding the relations of test boredom to other constructs, it makes sense to analyze relations between trait test boredom and other trait constructs as well as relations between state test boredom and other state constructs (cf., Brunswik, 1952; see also Geiser et al., 2017). Based on the relative universality assumptions of the CVT (Pekrun et al, 2006, 2018, 2021), similar structural relations with antecedents and outcomes can be assumed for trait and state test boredom.

Apart from the specifics of testing situations, it can be assumed that, from a phenomenological perspective, test boredom is quite similar to the boredom experienced in other school situations (i.e., class- and learning-related boredom), with its unique nature stemming from the context of testing. There is a lack of empirical studies investigating if test boredom can be empirically distinguished from classroom- and learning-related boredom. However, because previous research has shown that other academic emotions (e.g., enjoyment, pride, anger, anxiety) can be clearly delineated in terms of the situation in which they are experienced (e.g., Pekrun et al., 2011), this can also apply to boredom.

An important issue in defining test boredom is what a "test" actually is. Although there are widely varying definitions of the term "test" in different fields of research, the Cambridge Dictionary defines "test" as "a way of discovering, by questions or practical activities, what someone knows, or what someone or something can do or is like." An important and commonly used differentiation of tests is based on the direct personal consequences associated with test scores (Barry et al. 2010). High-stakes test scores have important personal consequences (e.g., achievement, admissions, and placement tests), while low-stakes test scores have little to no personal consequences (e.g., only average country test scores are reported; e.g., in the PISA studies; OECD, 2019).

However, beyond these formal definitions, it is important to note that whether a test is actually experienced as a low- or high-stakes test depends on individuals' judgment. For example, even tests that have no consequences may be very important to some students with high achievement motivation. Conversely, even objectively very important tests can be rated as unimportant by individual students because they do not see – or do not want to see – their relevance.

Another commonly used distinction is whether the assessments are formative or summative. Formative assessments collect data to improve student learning, whereas summative assessments use data to assess how much a student knows or has retained at the end of a learning sequence (American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education [AERA, APA & NCME], 2014).

In conceptualizing "test boredom," we refer to all types of tests, that is, low-stakes and high stakes tests, as well as both formative and summative assessments. This usage of the term is consistent with the use of the term "test" in more than 50 years of research on "test anxiety" (Mandler & Sarason, 1952), in which "tests" have also been defined broadly (e.g., von der Embse et al., 2018; Zeidner, 1998). In sum, we define "test boredom" as follows: Test boredom is the experience of boredom in situations that are labeled and/or experienced as tests.

Occurrence and Antecedents of Test Boredom

Occurrence

In a sample of sixth-graders, Goetz et al. (2007) empirically identified levels of test boredom, although this construct was not in the center of the study. State test boredom was assessed during a low-stakes mathematics achievement test with a single-item measure. Mean levels on two assessments during the test were *M* = 1.98 and 2.11 (*SD* = 1.25/1.36), respectively, on a scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). There is a further study by Raccanello et al. (2019), in which elementary students' trait test boredom in mathematics was assessed via a 4-item scale (adapted from the Achievement Emotions Questionnaire – Elementary School, AEQ-ES, Lichtenfeld et al., 2012). In this study, boredom was also not the focus of the investigation. The mean of this scale was *M* = 1.96 (*SD* = 1.26), with an answer format ranging from 1 (*not at all*) to 5 (*very much*). Although these findings provide initial evidence on the occurrence of test boredom, attesting to its manifestation during tests, these results are limited in scope (e.g., silent about its antecedents and effects).

Antecedents

CVT is a key theory that can explain possible antecedents of test boredom (and other academic emotions; Pekrun, 2006, 2018, 2021; Pekrun et al., 2023). This theory posits that individuals' perceptions of their personal control and value concerning achievement activities and outcomes represent the most important psychosocial antecedents of boredom in achievement settings. Based on the relative universality assumptions of the CVT, the structural relations between boredom and its antecedents in test situations (i.e., test boredom) should generally be similar to those of boredom in other academic settings (i.e., class- and learning-related boredom).

Nevertheless, test boredom has unique antecedents, namely features of the test. In other words, the relations between boredom and its antecedents can be assumed to be universal, with specific antecedents sometimes being quite different and consequently leading to different levels of boredom (i.e. relative universality). Thus, test boredom may differ in magnitude from other types of boredom due to the specifics of the situations (i.e. tests) in which it is experienced.

Perceived Control. Perceived control refers to individuals' perceived causal influence over actions and outcomes (Skinner, 1996). CVT suggests that the relation between test boredom and perceived control is curvilinear, with higher levels of boredom experienced when perceived control is either very low or very high (Pekrun et al., 2023). This is consistent with traditional approaches to boredom, in which its occurrence is attributed to a lack of fit between person and environment (Csikszentmihalyi, 1975/2000, 1990; Westgate & Wilson, 2018). Here, the experience of test boredom (and other types of boredom) differs from other emotions, which are assumed to have linear rather than curvilinear relations with perceived control (Pekrun & Goetz, in press).

The proposed link between levels of control and boredom has found partial support in studies on learning- and class-related boredom. Rather than the predicted curvilinear relation, perceived control was commonly found to negatively relate to boredom (e.g., Forsblom et al., 2021; Pekrun et al., 2010, 2014, 2023; see also Goetz & Hall, 2020). This could be due to the fact that tasks in schools and universities are designed to present challenges that facilitate learning. As such, typical tasks are not extremely easy to solve, so that a very high level of control rarely occurs (e.g., Dicintio & Gee, 1999; Goetz et al., 2006, 2012). However, a recent experimental study showed that boredom in fact occurred in situations characterized by very high as well as very low perceived control (Struk et al., 2021).

Such non-optimal (i.e., very high or very low) levels of control may occur when there is a lack of fit between task demands and individuals' task-related abilities. It is important to note that there may be various indicators for such non-optimal challenge. For example, when task demands exceed students' ability, low perceived control, overchallenge, and low task-related self-efficacy (see Marsh et al., 2019) may be identified. On the other hand, when one's abilities exceed task demands, high perceived control, underchallenge, and high self-efficacy may be reported. A more objective indicator would be, for example, the difference between the difficulty of a given task and estimates for a person's ability. Such a difference could be calculated in tests that are scaled using Rasch modeling (Rasch, 1980). The difference should also be related to the constructs described above (i.e., perceived control, overchallenge, underchallenge, self-efficacy). Thus, various indicators of very low and very high levels of perceived control during tests can be used to assess antecedents of test boredom.

Perceived Value. Perceived value concerns the relevance of actions and outcomes for an individual (Pekrun, 2006). CVT posits a negative relation between perceived value and test boredom. In this respect, the experience of test boredom (and other types of boredom) differs from other emotions that are assumed to have a positive relation with perceived value (Pekrun & Goetz, in press). It is important to note that different facets of value can be distinguished, including intrinsic value (e.g., interest) and extrinsic value (e.g., grades), professional utility (e.g., career aspirations), and general utility for life (e.g., using math competences in daily life; Gaspard et al., 2015). Test boredom can be assumed to relate negatively to all facets of value. In line with this assumption, empirical studies have consistently reported negative correlations of learning- and class-related boredom with different types of subjective value (e.g., Goetz et al., 2006; Pekrun et al., 2010, 2011). However, these studies have mainly examined a single value facet, which does not allow for a systematic comparison of potentially variable relations between boredom and different types of values. To date, the extent to which different value facets differ in their relation to boredom is largely an open question (Pekrun & Goetz, in press). In our study, we focus on the traditional distinction between intrinsic and extrinsic value. Intrinsic value implies that the task is an end in itself (e.g., enjoyment of working on the task; Gaspard et al., 2015) and is therefore related to the constructs of intrinsic motivation (Ryan & Deci, 2009) and individual interest (Pintrich, 2003). In contrast, extrinsic value is instrumental in nature (e.g., related to achieving good grades or a professional position) and is closely related to extrinsic motivation (Ryan & Deci, 2009). For test boredom, it can be assumed that high-stakes and low-stakes tests will have different effects on the subjective experience of extrinsic value, with extrinsic value likely to be higher in high-stakes tests and, consequently, boredom being lower during these tests (Barry et al. 2010).

On the basis of propositions of the CVT and in light of empirical evidence for academic boredom beyond testing situations (i.e., learning- and class-related boredom), strong arguments for the occurrence of test boredom can be derived: (1) For diagnostic reasons, tasks within a test typically cover a variety of difficulty levels. Thus, during tests a number of situations may occur, in which students would experience non-optimal levels of control (Wainer, 2000). These situations may give rise to the experience of test boredom. (2) It is plausible that students may perceive many tests as having low intrinsic (i.e., lack of interest in the topic) and/or extrinsic value, which provides another route to the experience of test boredom (Pekrun et al., 2023; Westgate & Wilson, 2018).

Effects of Test Boredom on Achievement

Assumptions Based on Control-Value Theory

The CVT (Pekrun, 2006) explains possible effects of academic emotions on achievement outcomes. Following the relative universality assumptions of CVT, relations with outcomes should be similar for test boredom and boredom in other academic situations (Pekrun & Goetz, in press). Test boredom can be assumed to deplete cognitive resources due to mind wandering, to reduce motivation to work on tasks and exert effort, to lead to use of superficial strategies (e.g., no deep thinking), and to undermine flexible adaptation of strategy use to the specific demands of the test, all of which should reduce test performance.

Existing studies in fact suggested that higher levels of boredom corresponded with poorer achievement (e.g., Camacho-Morles et al., 2021; Daniels et al., 2009; Goetz et al., 2010; Pekrun et al., 2010, 2011, 2014). Moreover, longitudinal studies indicated that boredom and achievement were linked by reciprocal effects over time, with boredom having consistently negative effects on later performance which, in turn, contributed to subsequent higher levels of boredom (e.g., Pekrun et al., 2014, 2017).

In their meta-analysis that included 29 studies involving 19,025 students, Tze et al. (2016) found that boredom had a consistent negative relation with academic outcomes ($\bar{r}$ = -.24). In a subsequent meta-analysis of 66 studies (Camacho-Morles et al., 2021; total N = 28,410), the disattenuated correlation corrected for measurement error was $\rho = -.25$. Observed correlations between boredom and academic performance of around r = -.25 are on a similar level as correlations between other positive and negative emotions and performance (Goetz & Hall, 2020). Most studies examined test anxiety and found that typical correlations with achievement outcomes were between r = -.20 and -.25 (Goetz & Hall, 2020). In sum, the correlations between boredom and achievement are on a similar level as those of other academic emotions. They are sizable relative to typical effect sizes in the educational and psychological literature (Gignac & Szodorai, 2016).

There exists only one study that has examined relations between test boredom and achievement. Raccanello et al. (2019) investigated relations between trait test boredom and achievement (grades) in the language domain (native language) and mathematics in elementary school students in Italy. No significant relations between test boredom and grades were found in the language domain but significant negative relations in mathematics were revealed (*r* = -.26; *p* < .001).

In general, test boredom seems to be a promising construct to examine relations between boredom and achievement, because a performance measure to which test boredom relates is directly available. Performance measures of boredom in the classroom and in learning (e.g., subsequent performance outcomes) tend to be less directly related to the situation in which boredom occurs.

Abundance Hypothesis

With respect to the effects of test boredom on test performance, it may be important to consider whether boredom results from over- or underchallenge. Over- and underchallenge are two types of non-optimal challenge (i.e., a lack of fit between a person's ability and task demands; see Csikszentmihalyi, 1975/2000; 1990; Pekrun, 2006, 2018, 2021). Both over- and underchallenge have been shown to be associated with higher levels of boredom in the classroom (Krannich et al., 2019). Thus, it can be assumed that test boredom also arises from these qualitatively different types of non-optimal challenge.

In principle, test boredom should be expected to have a negative impact on mediators of boredom-achievement relations as noted earlier (e.g., reduced cognitive resources, low motivation; Pekrun, 2006), regardless of whether boredom results from over- or underchallenge. However, when working on easy tasks (i.e., being underchallenged), the negative effects of boredom are likely to be relatively small because even significantly reduced resources may still be sufficient to solve the task. In other words, resources may be abundant to simultaneously process the emotion and perform the task. In contrast, a reduction in resources due to boredom during difficult tasks (i.e., being overchallenged) should have stronger adverse effects on achievement outcomes. For difficult tasks, all resources would need to be allotted to solve the task, but are only partially available because they are consumed by boredom, thus reducing performance. In situations of severe overchallenge, almost all cognitive resources are likely to be devoted to boredom processing (and, depending on the situation, to other emotions, such as anxiety), and the student may even stop working on the tasks because he or she sees no chance of solving them anyway. In this way, boredom differs from anxiety, which can also occur in situations of being overchallenged, but is usually associated with high value (Pekrun, 2006) and therefore is more likely to keep one engaged in the task. Based on these considerations, we hypothesized that test boredom would be more detrimental when students feel overchallenged during the test than when they feel underchallenged (‘abundance hypothesis’, H4). To our knowledge, the abundance hypothesis for boredom has not yet been proposed or tested.

An important implication in case of the empirical support for the abundance hypothesis would be that the potential strength of the relations between test boredom and performance would be underestimated if boredom due to overchallenge and underchallenge were not analyzed separately. In other words, potentially strong negative effects of test boredom on performance due to overchallenge would not be detected if the antecedents of overchallenge and underchallenge were not separated in the analyses.

Aims and Hypotheses of the Present Research

To our knowledge, there is no research on the occurrence of test boredom, its antecedents, and its effects. In the current research, we aimed to fill this gap. Based on key propositions of the CVT (Pekrun, 2006), test boredom should occur because many test situations should give rise to the antecedents as outlined in this theory, namely non-optimal levels of control and low levels of value. Further, from a theoretical perspective and in line with earlier findings (Raccanello et al., 2019), test boredom should have negative effects on achievement outcomes. We also tested the assumption that test boredom would be more harmful when learners were overchallenged during a test than when they were underchallenged, as they should largely have sufficient resources for task completion in the case of underchallenge but not in the case of overchallenge (abundance hypothesis).

We conducted two studies testing these hypotheses. As boredom in education has been shown to be domain-specific (Goetz et al., 2007), in both studies we focused on one domain, namely, mathematics. We chose mathematics because it is a core school subject and is often studied in the context of STEM education research (i.e., science, technology, engineering, and mathematics; e.g., Li et al., 2020). Further, the perceived value of this domain is typically rather high (Goetz et al., 2014; Haag & Goetz, 2012), presumably resulting in a relatively low level of test boredom compared to other domains. By choosing to investigate test boredom in mathematics, we opted for a rather conservative test of the hypothesis that test boredom occurs. However, based on the relative universality assumptions of the CVT (Pekrun, 2006, 2018, 2021), structural relations between test boredom and its antecedents and effects should be quite similar across academic domains.

Study 1 focused on the occurrence, antecedents, and effects of trait and state test boredom as experienced during a low-stakes test. Trait (i.e., habitual) test boredom was assessed one to three weeks prior to state boredom. State boredom (i.e., real-time boredom) was assessed several times during a difficult and an easy part of a math achievement test inducing over- and underchallenge, respectively. Study 2 differed from Study 1 in the following respects: First, in this study, we focused more specifically on the occurrence and effects of state test boredom. Second, to improve the generalizability of results, we analyzed data from a larger sample. Third, to improve ecological validity we used a valid standardized math test aligned with the course curriculum, during which state test boredom was assessed several times. Fourth, to vary the operationalization of non-optimal challenge we used a different indicator of over- and underchallenge in Study 2 than in Study 1. Finally, to further increase the generalizability of our results, we used a different statistical approach to test the abundance hypothesis, namely the latent moderated structural equations (LMS) method.

Across the two studies and based on the theoretical propositions of the CVT, we aimed to test the following hypotheses:

Hypothesis 1: Students report levels of test boredom that are statistically significantly different from not being bored at all.

Hypothesis 2: Test boredom shows significant relations with core antecedents: positive relations with non-optimal control and negative relations with both intrinsic and extrinsic value.

Hypothesis 3: Test boredom shows negative relations with core achievement indicators, including achievement test scores and grades.

Hypothesis 4: Test boredom has a stronger negative effect on test performance when students feel overchallenged during the test than when they feel underchallenged (‘abundance hypothesis’).

Transparency and Openness

In line with the openness and transparency standards of the Journal of Educational Psychology (Kendeou, 2021), we describe the sample and procedure and report all data exclusions in detail. All data, measures, and analysis codes are available at [link will be provided here]. Data were analysed using Mplus 8.6 (Muthén & Muthén, 1998-2017). Study 1 was not preregistered. The analysis of Study 2 consists of a secondary data analysis of the PALMA study, which was not preregistered.

Study 1

Study 1 explored the intensity of (1) mathematics trait test boredom and (2) state test boredom during a low-stakes mathematics achievement test. The achievement test consisted of two sections with easy and difficult tasks, respectively. With this test design we aimed to induce boredom due to non-optimal levels of control (i.e., being under- or overchallenged). Relations of both trait and state boredom to its proposed antecedents (control, value) and effects (achievement) were analyzed.

Method

Participants

The sample consisted of 208 students (54% female; mean age = 13.73 years, *SD* = 0.44, Min = 12.65, Max = 15.55) from nine 8th grade math classes. These classes came from four different schools in the high-achieving track of the three-track German secondary school system (i.e., Gymnasium; approximately 40% of the total student cohort attend this track; Federal Statistical Office [Statistisches Bundesamt], 2020). The reason for focusing on one grade level and one school track was that this allowed us to use the same math test for all students in our sample.

Procedure

The study was part of a larger project (Goetz et al., 2017) investigating students' emotions in testing situations. The study was conducted in compliance with ethical standards described in the WMA Declaration of Helsinki. It has been approved by the Institutional Review Board of the first author's institution, with all study procedures have been deemed appropriate. For the sake of conciseness, we focus on those procedures that pertain to the present research questions.

Trait-Assessment, Assessment of Grades and Demographic Data. In each classroom, the study started with an assessment of trait variables, achievement outcomes, and demographic data during a regular math class. We assessed trait test boredom related to mathematics tests as well as trait antecedents of trait test boredom (i.e., trait non-optimal levels of control and trait value during mathematics tests). Achievement outcomes were assessed as self-reported mathematics grades. We used a paper-and-pencil questionnaire to gauge these variables. One class (*n* = 23) did not participate in the trait-assessment so our sample size was 185 students in the analyses involving these data. Students were informed that they would participate in a second assessment, which would mainly be a mathematics achievement test.

State-Assessment, Mathematics Test. One to three weeks after the trait-assessment, participants worked on the mathematics achievement test (paper-and-pencil version) during their regular math classes. To make the test subjectively relevant and encourage students to perform well, the test was described as a preparatory test (i.e., practice test) for the upcoming state-wide comparison tests (VERA-8 [VERgleichsArbeiten], grade level 8; Graf et al., 2016; for a detailed description of the test see below). The task material also stated that it was a test. Additionally, we awarded a prize of 250 Euros to the class with the best average test performance. Our test was a low-stakes test. Students received no feedback on their test score and their score was not counted towards their grades. As tests typically comprise tasks of different difficulty levels, students worked on one part with several relatively easy tasks and one part with several relatively difficult tasks. By splitting the test into a block of difficult tasks and a block of easy tasks, we aimed to elicit a different suboptimal level of control in each block (i.e., being over- or underchallenged). We fully counterbalanced within classrooms whether students started with the easy or with the difficult part.

State test boredom was assessed five times, using each a single-item rating scale and a multi-item scale: (a) once before each part of the test (two concurrent assessments, measuring boredom as experienced in this moment), (b) once after each part (two assessments related to the preceding part, that is, two retrospective reports of boredom as experienced while working on the math tasks), and (c) once after the test (one concurrent assessment, measuring state test boredom as experienced in this moment). *State perceived value* (intrinsic and extrinsic; each assessed with a single item) was also assessed five times: (a) once before each part (two concurrent assessments, measuring value as perceived in this moment), (b) once after each part (two assessments related to the preceding part, that is, two retrospective reports of value as perceived while working on the math tasks), and (c) once after the test (retrospective assessment, value as perceived with respect to the whole math test, i.e., both parts). *State non-optimal control* (levels of being over- or underchallenged) was assessed three times: (a) once after each part (two retrospective assessments, measuring being over-/underchallenged as perceived in this part), and (b) once after the test (retrospective assessment, being over-/underchallenged as perceived with respect to the whole math test, i.e., both parts). All these assessments were embedded in the test booklets (i.e., they also had a paper-and-pencil format).

Students were allotted 45 minutes for working on the math test. Additional five minutes were given for completing the self-report questions about boredom and its antecedents integrated into the test. Thus, the total administration time was 50 minutes.

Missing Data

A total of 3.21% of data were missing, stemming from 89 incomplete records. The percentage of missing values across the 102 variables ranged from 0.00% to 15.38%. Full information maximum likelihood (FIML) was used to deal with the missing data (see Enders, 2010).

Measures

Our strategy for constructing and selecting self-report measures of boredom, non-optimal control, and value was guided by the following considerations. (1) We aimed to assess boredom both in the trait and state assessment by using a multi-item scale reflecting the different components of boredom. (2) In addition to using the multi-item scales, we aimed to assess test boredom with a single item in both the trait and state assessments. The reason for using a single item additionally to the multi-item scale was that the mean level of the single item (e.g., “How strongly do you typically experience boredom during math exams?”) is much easier to interpret than a score aggregating answers from a multi-item scale. (3) Given the extensive assessment of state boredom, we decided to use a single item for all other self-report assessments both in the trait and state assessment in order to limit administration time (Gogol et al., 2014). (4) To make trait and state assessments as comparable as possible, we used parallel versions of trait and state items and scales.

Test Boredom - Trait. The wording of the single item was “How strongly do you typically experience boredom during math exams?”. Participants responded using a five-point rating scale ranging from 1 (*not at all*) to 5 (*very strongly*). Response alternatives 2, 3, and 4 were not specified. The wording of the item and the answer format were based on the study by Krannich et al. (2019).

The multi-item scale measuring trait test boredom (Test Boredom Scale-Trait, TBS-Trait) was constructed by modifying items from the class- and learning-related boredom scales of the Academic Emotions Questionnaire (AEQ; Pekrun et al., 2011). Similar to the AEQ scales, the TBS-Trait comprised four sub-scales each representing a different component of boredom. In the TBS-Trait each component was assessed with 3 items, including the affective component (e.g., “I’m bored during math exams”), the cognitive component (e.g., “I’m so bored during math exams that I find myself daydreaming”), the motivational component (e.g., “I’m so bored that I would prefer not to start the math exams at all”), and the physiological/expressive component (e.g., “I’m so bored that I get tired”). Answers were provided on a 5-point response scale ranging from 1 (*not at all true*), 2 (*slightly true*), 3 (*partly true*), 4 (*mostly true*), to 5 (*completely true*). Reliability was α = .86 for the overall score comprising all four components. An overview of all test boredom trait items is provided in Appendix A (Table A1).

Test Boredom - State. The wording of the single-item was “How strongly do you experience boredom at the moment?” (concurrent) and “How strongly did you experience boredom while working on the math tasks?” (retrospective). Participants responded using a 5-point rating scale ranging from 1 (*not at all*) to 5 (*very strongly*). Response alternatives 2, 3, and 4 were not specified. The wording of the item and the response format were based on a study by Goetz et al. (2007).

The multi-item state test boredom scale (Test Boredom Scale-State; TBS-State) was also based on the AEQ (Pekrun et al. 2011). The wording was parallel to the TBS-Trait. It comprised four sub-scales representing the different components of boredom with 3 items each, namely the affective component (e.g., concurrent: “I’m bored”, retrospective: “I was bored”), the cognitive component (e.g., concurrent: “I’m so bored that I find myself daydreaming”, retrospective: “I was so bored that I found myself daydreaming”), the motivational component (e.g., concurrent: “I’m so bored that I would prefer not to start the math tasks at all”, retrospective: “I was so bored that I would have preferred not to start at all with the math tasks”), and the physiological/expressive component (e.g., concurrent: “I’m so bored that I am tired”, retrospective: “I was so bored that I was tired”). Answers were provided on a 5-point response scale ranging from 1 (*not at all true*), 2 (*slightly true*), 3 (*partly true*), 4 (*mostly true*), to 5 (*completely true*). Across the five assessments, coefficient α ranged from .83 to .94 for the overall score comprising all four components. An overview of all state test boredom items is provided in Appendix A (Table A2).

Non-Optimal Control – Trait and State. We measured students’ non-optimal experiences of control in terms of perceived over- and underchallenge using items developed by Krannich et al. (2020). The items in the trait assessment were “During math exams I feel overchallenged” and “During math exams I feel underchallenged”. In the state assessment the items were "I am feeling overchallenged [underchallenged]" (concurrent) and "I felt overchallenged [underchallenged]" (retrospective). For both the trait and state assessment, participants responded using a 5-point rating scale ranging from 1 (*not at all true*) to 5 (*completely true*).

Perceived Value – Trait and State. Previous studies have shown that boredom might be differentially related to different types of value (e.g., Goetz et al., 2006). Hence, we focused on two traditionally assessed value types, namely intrinsic and extrinsic value (see Gaspard et al, 2015). We adapted two items for the trait and state assessment of intrinsic and extrinsic value, respectively, which were each based on an item of the corresponding scales of the Project for the Analysis of Learning and Achievement in Mathematics (PALMA; Pekrun et al., 2007). The trait items were "Math is very important to me regardless of the grade I get" (intrinsic value) and "It is very important for me to get a good grade in math" (extrinsic value). The state items for concurrent assessments were "The math tasks are important to me regardless of the result" (intrinsic value) and "In this math tasks it is important to me to achieve a good result" (extrinsic value). The state items for retrospective assessments were "The math tasks were important to me regardless of the result" (intrinsic value) and "In this math tasks it was important to me to achieve a good result" (extrinsic value). Answers were provided on 5-point rating scale ranging from 1 (*not at all true*) to 5 (*completely true*).

Mathematics Test/Test Achievement Measure. A mathematics test was developed to match the study design (i.e., test sections of varying difficulty). The test represented the performance measure in the study. The math tasks were adapted from the database of a nationwide written mathematics test (VERA 8 [VERgleichsArbeiten], grade level 8; see Graf et al., 2016) taken by students in the 8th grade of the German school system as a standardized achievement test (developed by the Institute for Educational Quality Improvement; IQB, Berlin, Germany). The tasks covered four different content areas (i.e., numbers, measurement, space and form, functional relationships) and are classified by the IQB as easy or difficult based on solution frequencies in independent nationwide representative studies. There were multiple-choice tasks as well as tasks requesting short open answers (e.g. calculations, writing down the solution). Relying on these tasks allowed us to create a relatively authentic and ecologically valid test situation that nevertheless, unlike an actual exam, made it possible to experimentally vary the difficulty of the tasks in full accordance with ethical considerations (i.e., there was no disadvantage from taking the exam because the result did not count toward students' grades). A Grade 8 mathematics teacher was consulted to select easy and difficult tasks in line with the regular curricula of the four participating schools. This resulted in a pool of 22 tasks for the easy part and 10 tasks for the difficult part of the test. We chose fewer difficult than easy tasks as they take more time to work on. Results of the item analysis showed low item-total correlations for four easy and three hard items, which were subsequently excluded from the math score. Thus, we used 20 easy and 7 difficult tasks. Coefficient α for the math score was .74.

Academic Achievement. Academic achievement was operationalized as students' last midterm grade in mathematics, which is typically based on scores for written exams combined with scores for course-specific oral exams in German schools. Grades range from 1 (*very good*) to 6 (*insufficient*). For the ease of the interpretation, we inverted grade scores so that higher numbers indicated better performance.

Analytic Strategy

Hypothesis 1: Occurrence of Test Boredom. To test H1, we ran one-sample *t*-tests using Bonferroni correction to test whether the mean value was different from 1. We did this for each single item assessing trait and state test boredom (i.e., "*not at all*" on the Likert scale). For the overall scale scores, confirmatory factor analysis (CFA) was conducted to estimate separate hierarchical measurement models for trait and state test boredom. Each of the two models (i.e., state and trait) included the four test boredom components (i.e., affective, cognitive, motivational, physiological) as primary factors, and overall test boredom as a secondary factor. This is consistent with our definition of test boredom as a construct that is composed of four components. Based on the CFA models, we tested whether the latent means for trait and state boredom were different from 1 (i.e., *"not at all true"* on the Likert scale). Means different from 1 indicate that test boredom did in fact occur as reported by students.

Hypotheses 2 – 4: Antecedents (H2) and Effects (H3, H4) of Test Boredom. To test H2 and H3, we again estimated hierarchical measurement models for trait and state test boredom using confirmatory factor analysis (CFA), with boredom as a secondary factor. For state test boredom, we conducted a multilevel CFA to take the hierarchical data structure into account (i.e., the nestedness of state measures of boredom within students). Latent correlations with other variables were based on this multilevel CFA.

We investigated correlations among trait and state test boredom and their proposed antecedents (H2; being over- and underchallenged, intrinsic and extrinsic value) and outcomes (H3; math score, academic achievement). To test the abundance hypothesis (H4), we investigated the relations between state test boredom and test scores (i.e., the results of the math test) separately for the two different parts of the test (i.e., difficult vs. easy part), which were designed to induce overchallenge in the difficult part and underchallenge in the easy part.

For the analyses of the state data, multilevel models were estimated, with state test boredom and antecedent variables at Level 1, and persons at Level 2. An exception is the analyses testing H4 that used scores related to the two parts of the test. As only one assessment of state test boredom was available for each of the two parts, we did not use multilevel analysis for testing H4.

All analyses (trait and state) were run on the between-person level based on latent variables. We did not run within-person analyses due to the low number of assessments within students for the antecedent and outcome variables of test boredom (e.g. only two state assessments for being over- and underchallenged – one assessment after each part of the test).

Models were estimated with Mplus 8.6 (Muthén & Muthén, 1998-2017) using the robust maximum likelihood estimator (MLR). Cluster-robust standard errors were used to take the non-independence of observations due to the hierarchical data structure (i.e., students nested in classrooms) into account. Model fit of each of the measurement models was evaluated using the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). We considered typical cutoff scores reflecting good fit to the data, that is, CFI and TLI close to or higher than .90, RMSEA < .08, and SRMR < .08 (see Brown, 2015). All analyses were conducted based on a statistical significance level of α = .05. The analysis scripts are accessible via OSF (https://osf.io/ftr4g/?view_only=beac4ca87987492294aeac4a1bb96c86).

Results

Descriptive Statistics and Measurement Models

Means and standard deviations of all manifest antecedent and outcome variables, as well as their intercorrelations are presented in Table 1. Means and standard deviations of the boredom measures are shown in Table 2 (see next section for details and statistical tests).

Confirmatory factor analysis (CFA) showed a good fit for the second-order factor measurement model for *trait test boredom* using the four components of boredom as primary factors ($\chi^2(50) = 79.04$, $p = .006$, CFI = .946, TLI = 0.929, RMSEA = 0.057, and SRMR = 0.050). Multilevel CFA also showed a good fit for the *state test boredom* second-order factor measurement model ($\chi^2(111) = 249.53$, $p < .001$, CFI = .962, TLI = 0.955, RMSEA = 0.035,

$\text{SRMR}_{\text{Within}}$ = 0.044 and $\text{SRMR}_{\text{Between}}$ = 0.062).

The single-item measure for *trait test boredom* showed a high positive correlation with the overall trait test boredom scale (r = .66), indicating that the single item was substantially associated with the multi-item scale. In line with this finding, for the *state test boredom* single item also had a high positive correlation with the overall state test boredom scale (r = .87). To examine the relations between trait and state test boredom, we additionally conducted a CFA that included all trait and state measures. The model fit was as follows: χ^2 (344) = 820.65, CFI = .921, TLI = 0.907 RMSEA = 0.037, $\text{SRMR}_{\text{Within}}$ = 0.051, $\text{SRMR}_{\text{Between}}$ = 0.065. The correlation between trait and state boredom was r = .50 (p < .001) for the multi-item scales and r = .21 (p = .002) for the single items (the reduced strength of the latter correlation may be due to low reliability of single-item assessments; Gogol et al., 2014).

Hypothesis 1: Occurrence of Test Boredom

For *trait test boredom*, the means of the single item and the overall score were statistically different from 1 (with 1 on the Likert scale indicating no boredom experience; *p*s < .001). Effect sizes were 0.61 for the single item (Cohen's *d* for one-sample t-tests; Cohen, 1988, p. 46) and 0.89 (*latent d*; Hancock, 2001) for the overall score of the scale (Table 2).

Results of the one-sample *t*-tests revealed that *state test boredom* measured with single-item or overall scores were all also not equal to 1 throughout all parts of the test (adjusted *p*-values using Bonferroni correction for multiple testing were all < .001), with effect sizes ranging from d = 0.65 to 1.04. The mean score across all state single items on test boredom was M = 1.91 (SD = 1.14) and the mean score across all state scale scores on boredom was M = 1.53 (SD = 0.67). For a graphical illustration, mean values and violin plots for all trait and state boredom measures are shown in Figure 1. Mean values were relatively low. Nevertheless, the violin plots show that the scores were distributed across a wide range, with some students even reaching the highest possible score.

Hypothesis 2: Antecedents of Test Boredom

Latent correlations between the overall (i.e., multi-item) trait and state boredom scores and antecedents are presented in Table 3. For state boredom, the coefficients represent latent between-person correlations at Level 2 derived from the multilevel CFA model. For both the trait and state assessment, the boredom scores showed positive correlations with both overchallenge and underchallenge. In addition, for the trait and state assessments, the boredom score was negatively related to both intrinsic and extrinsic value. Thus, supporting H2, all correlations for the trait and state assessments were significant and in the expected directions.

Hypotheses 3 and 4: Relations of Test Boredom with Achievement

Correlation coefficients among the trait and state multi-item boredom scores and achievement outcomes are presented in Table 4.

No significant relations between boredom and the score in the math test were found. However, both for the trait and state assessments, the overall boredom score showed negative correlations with math grades. Thus, with respect to H3, all significant correlations both for the trait and state assessments were in the expected directions.

Table 4 also shows the results for the abundance hypothesis testing (H4). Correlations between state boredom experiences during the easy part of the test as well as during the difficult part of the test with corresponding test achievement (i.e., achievement in the easy and difficult part) are shown separately.

In terms of being over- and underchallenged, students reported state levels for the easy and difficult parts of the test. For the easy part, the mean level of being underchallenged was M = 2.17 (SD = 1.12) and of being overchallenged M = 2.03 (SD = 0.97). For the difficult part, the mean level of being underchallenged was M = 1.45 (SD = 0.70) and of being overchallenged M =

2.96 (*SD* = 1.09). In the easy part of the test, underchallenge scores were significantly higher ($t(204) = 8.83$, $p < .001$, Cohen's $d = 0.61$) and overchallenge scores were significantly lower ($t(201) = -11.76$, $p < .001$, Cohen's $d = -0.83$) than in the difficult part of the test, suggesting that students actually experienced the easy part as less challenging than the difficult part.

In line with the abundance hypothesis (H4), we found a significant negative correlation between state boredom and the test score for the difficult part of the test ($r = -.22$). In contrast, the correlation for the easy part of the test was not significant ($r = .09$). As there was no overlap in the confidence intervals of the two correlations, they were significantly different.

Discussion

In line with our assumptions (H1), we found that both trait and state test boredom occur at a statistically significant level. Also in line with our assumptions (H2), both trait and state test boredom were related to their proposed antecedents, namely non-optimal control (over- or underchallenge), intrinsic value, and extrinsic value. Both being over- and underchallenged showed significant positive relations with test boredom. In addition, largely in line with our assumptions (H3), test boredom showed significant relations with academic achievement. Both trait and state test boredom showed negative relations with students' math grades.

Finally, we also found support for our abundance hypothesis (H4): Boredom was negatively related to test scores in the difficult part of the test, and not significantly related to the scores in the easy part of the test. Thus, when students are underchallenged, test boredom seems to be merely a side effect of working on tasks, without affecting test performance – likely because students have sufficient resources, motivation, and strategies to succeed on easy tasks even when being bored. However, when students feel overchallenged, boredom can be expected to have a negative effect on performance because it is likely to consume resources that would actually be needed to complete the task.

However, Study 1 also had limitations. The sample was relatively small, and we used a non-standardized mathematics test that was divided into an easy and a difficult part, which is not what happens in natural testing situations. Awarding a prize of 250 Euros to the class with the best average test performance to make the test subjectively relevant and encourage students to perform well also does not reflect a typical test situation in school. Furthermore, the abundance hypothesis was investigated for the first time in this study. To test the generalizability of the findings, a conceptual replication using a different approach to measuring over- and underchallenge is needed. Study 2 addresses these issues.

Study 2

Study 2 used a dataset that is based on a large sample and a classic standardized mathematics test. We also used a different indicator of being over- versus underchallenged, namely, students' self-efficacy expectations to be able to solve math problems (i.e., anticipatory challenge).

This study further explored the occurrence of mathematics state test boredom (H1) as well as the relations of state test boredom with test performance (H3), including the abundance hypothesis (H4). As an indicator of being over- vs. underchallenged during the test, students' self-efficacy expectations were assessed, which reflect anticipatory challenge. According to the abundance hypothesis, test boredom should be negatively related to test performance at low but not high levels of self-efficacy. As such, we assumed an interaction effect of test boredom and self-efficacy on test performance.

Method

Participants

The sample consisted of 1,612 students (grades 5-10; 46.84% female; mean age = 13.75 years, *SD* = 1.86) from 70 classrooms in 19 different schools in the state of Bavaria, Germany.

The sample comprised students from a wide range of socioeconomic backgrounds, including both rural and urban areas, and from all three school tracks of the public school system in this state.

Procedure

The study is a secondary analysis of an existing data set from a cross-sectional study (grade levels 5 to 10) that was part of the PALMA project (*Project for the Analysis of Learning and Achievement in Mathematics*; see, e.g., Murayama et al., 2013; Marsh et al., 2019; Pekrun et al., 2007; Pekrun et al., 2019), namely the PALMA Pilot Study 2. Findings for the present dataset have been published by Pekrun et al. (2019). However, findings from this study for the state data (i.e., the assessments during the mathematics test) and for boredom (trait and state) have not yet been published. The studies of the PALMA project received Institutional Review Board approval from the Bavarian State Ministry for Education, Science, and the Arts (reference III/5-S4200/4–6/68 908). Stratified sampling in the state of Bavaria was provided by the Data Processing and Research Center of the International Association for the Evaluation of Educational Achievement (IEA-DPC, Hamburg, Germany). Schools were recruited so that the resulting student sample was representative in terms of students' living in urban versus rural areas, socioeconomic status of parents, and school type within the three-tier school system in Bavaria. All instruments in this study were administered by the DPC's trained external test administrators in students' classrooms. Parental consent was obtained, and students' responses were kept confidential.

Students worked on a low-stakes mathematics achievement test (paper-and-pencil version) during their regular math classes. The mathematics tasks were verbally explicitly referred to as a test, and the term "test" was also used in the task material. The results of the test did not count toward students' grades, so it was a low-stakes test. State boredom was assessed at the beginning of the test (i.e., before starting to work on the tasks; current experience of boredom), after part 1 and part 2 of the test (also current experiences), and after part 3 of the test (retrospective judgement of boredom during the test), resulting in four assessments of state test boredom. As a measure of over- versus underchallenge, self-efficacy was assessed once directly before the first task on the math test. Students were allotted 90 minutes for working on the math test and the state assessments.

Missing Data

A total of 5.22% of data were missing, stemming from 467 incomplete records. The percentage of missing values across the nine variables ranged from 0.56% to 16.19%. Full information maximum likelihood (FIML) was used to deal with the missing data (see Enders, 2010).

Measures

State Test Boredom. In the three current assessments (once before and after part 1 and 2 of the test), students were asked “How do you feel at this moment”. Boredom was assessed with the single item “I am bored”. After part 3 of the test, students were asked “How did you feel when you worked on the math tasks”, and boredom was assessed with the single item “I was bored”. Participants responded on a 5-point rating scale ranging from 1 (*not at all true*), 2 (*slightly true*), 3 (*partly true*), 4 (*mostly true*), to 5 (*completely true*).

Self-efficacy. Self-efficacy was assessed using the approach proposed by Pajares and Graham (1999), which is aligned with Bandura's originally definition of task-related self-efficacy (Marsh et al., 2019). Students were offered the following instructions: “Imagine that you were asked to solve the following mathematics tasks. For each task, please indicate how confident you are that you can solve it correctly. So, you don’t have to solve the following three tasks; only estimate whether you think you *could* solve them.” Subsequently, three tasks of different difficulty (easy, medium, difficult) were shown. The tasks were adapted to fit the competency levels of participants from different grade levels and school tracks. The selection of the tasks was based on pilot studies (Goetz, 2004). After each of the three tasks students were asked: “How confident are you that you could solve this task?”. Students used an 8-point Likert scale ranging from 1 (*not confident at all*) to 8 (*completely confident*) to rate their confidence. The reliability of the three-item scale was α = .71. An example of the three tasks used for grade five students can be found in the Supplementary Materials (S1). In all grades and school tracks, students were required to factually complete the three tasks at a later time during the test. Thus, they were ecologically valid with respect to the content of the test.

Mathematics Test. The PALMA Mathematical Achievement Test (Murayama et al., 2013; Pekrun et al., 2007) was used to measure students’ current achievement. The PALMA test is a standardized test assessing competencies in arithmetic, algebra, and geometry across a wide range of ability. The test included both multiple-choice items and short-answer items (e.g., calculations, writing down the answer; see also supplementary material [S1] for sample self-efficacy assessment items). The reliability of the test was α = .87.

Analytic Strategy

Hypothesis 1: Occurrence of Test Boredom. To test H1, we ran a series of one-sample *t*-tests using Bonferroni correction to test if mean state test boredom scores were different from 1 (i.e., “*not at all true*” on the Likert scale). Means different from 1 indicate the occurrence of test boredom as reported by students.

Hypotheses 3 and 4: Relations of Test Boredom with Achievement. We investigated correlations between state test boredom and test achievement (H3). To test the abundance hypothesis (H4), we examined the relations between test boredom and test achievement as a function of self-efficacy. As noted, the hypothesis implies that boredom should show stronger negative effects on achievement for students with low self-efficacy (i.e., students for whom the test can be assumed to be overchallenging), than for students with high self-efficacy (i.e., students for whom the test can be assumed to be underchallenging). To test this hypothesis, we probed the latent interaction of boredom and self-efficacy using the latent moderated structural equations (LMS) method (Klein & Moosbrugger, 2000). The analysis was based on a CFA measurement model for self-efficacy and the mean centered boredom scores. Test scores were used as the outcome variable. The model was estimated with Mplus 8.6 (Muthén & Muthén, 1998-2017) using the robust maximum likelihood estimator (MLR). Cluster-robust standard errors were used to consider the non-independence of observations due to the hierarchical data structure (i.e., students nested in classrooms). The model was saturated. The analysis scripts are accessible via OSF (https://osf.io/ftr4g/?view_only=beac4ca87987492294aeac4a1bb96c86).

Results

Hypothesis 1: Occurrence of Test Boredom

Descriptive statistics (*M*, *SD*) for state test boredom, one-sample *t*-tests and the corresponding Cohen's *d* are presented in Table 5. Results of the one-sample *t*-tests showed that all state test boredom scores were different from 1, with effect sizes ranging from *d* = 0.69 to 0.76. The distributions of state test boredom scores are shown in Figure 2. Despite the relatively low mean values, the boredom scores were distributed across a wide range of values, with even the highest possible values reported by the participants.

Hypotheses 3 and 4: Relations of Test Boredom with Achievement

Correlations among state test boredom and math test scores are presented in Table 6. We found no statistically significant correlation between state test boredom and the scores on the math test, which is not in line with our hypothesis (H 3).

Table 7 shows the results for the abundance hypothesis test (H4). The effect of the latent interaction of state test boredom and self-efficacy on test scores was positive and significant (β = 0.38). Thus, in support of the abundance hypothesis, the strength of the effect of test boredom on the test score differs depending on the level of self-efficacy.

Figure 3 depicts the Johnson-Neyman plot for the interaction. The plot shows that the slope for boredom is significantly negative when the self-efficacy score is lower than 4.49 and becomes significantly positive when the self-efficacy score is higher than 5.96. Within the self-efficacy score interval from 4.49 to 5.96, in which the mean self-efficacy score was located (M = 5.58; SD = 1.34), the slope is not significant. Our findings are in line with the abundance hypothesis: Test boredom shows a negative effect on the test score for students with low self-efficacy, that is, for students for whom the test can be assumed to be overchallenging. For students with high self-efficacy, that is, for whom the test can be assumed to be underchallenging, we found less negative and even positive effects of test boredom on the test score.

Discussion

In a large sample (N = 1,613 5th to 8th graders), we found significant levels of boredom during a standardized low-stakes math test (H1). Test boredom was not related to test achievement, which is not in line with H3. However, in line with the abundance hypothesis (H4), test boredom was significantly negatively related with test achievement for students with low mathematics self-efficacy (i.e., for students who were likely to feel overchallenged). A plausible explanation is that overchallenged students need all their cognitive resources to complete complex or difficult tasks, and that boredom due to overchallenge consumes cognitive resources, such that the remaining resources are not sufficient to successfully complete the tasks. Although we assumed that boredom would have had less of a negative effect on performance for underchallenged students (i.e., students with high self-efficacy) because these students had sufficient resources available, we actually found significantly positive associations between boredom and test performance for these students. The reason for those positive correlations may be that perceptions of underchallenge (i.e., high levels of self-efficacy) may strengthen students' confidence and motivation (see Krannich et al., 2019), which can, in turn, enhance their test achievement (see below for this point). Without including self-efficacy as a moderator of the relations between test boredom and test performance, the relations between both variables were not significant. The reason for this could be the opposing effects of test boredom on test performance, which were dependent on the level of self-efficacy as found in our study.

General Discussion

Although research on academic boredom has proliferated in the past fifteen years, research on boredom during tests is largely lacking. We aimed to close this chasm. The main goal of our research was to investigate the occurrence of test boredom and its links with important antecedents and outcomes. Based on the CVT, we hypothesized that students experienced significant levels of boredom during testing (H1), and that test boredom was significantly related to theoretically hypothesized control-value antecedents (H2) and performance outcomes (H3). In addition, we proposed the abundance hypothesis (H4) which stated that test boredom was more detrimental when students felt overchallenged during the test compared to when they felt underchallenged.

Occurrence of Test Boredom (H1)

The results on the occurrence of test boredom were consistent across the two studies and supported H1. We found that test boredom occurred on a significant level both measured as a trait (Study 1) and as a state (Studies 1, 2). Importantly, reports on test boredom on the trait level indicated that test boredom was not an experience specific to the test situation we created in our study (i.e., as a state), but was also prevalent in other testing situations.

To judge levels of test boredom, mean scores on single items can be used. The mean score across all state single items on test boredom was *M* = 1.91/ 1.84 in Studies 1 and 2, respectively, on a Likert scale ranging from 1 to 5. This result is in line with findings of a study by Goetz et al. (2007), in which state test boredom was assessed with a single item twice during a low-stakes mathematics achievement test and yielded means of *M* = 1.98 and 2.11 using a similar response scale as the present research. Thus, the evidence on the occurrence of test boredom is consistent across these studies. As compared to other negative emotions during low-stakes tests, the level of boredom found in our study was relatively high. For example, in the study by Goetz et al. (2007) the values for the two assessments (each single items) during a low-stakes math test were *M* = 1.44/1.57 for anger and *M* = 1.32/1.31 for anxiety. Roos et al. (2021) found levels of anxiety during a low-stakes math test of *M* = 1.24 to *M* = 1.60 (median: 1.47; single items, retrospective assessments after each of the six parts of the test, 6-point Likert ranging from 0 = *"no anxiety at all"* to 5 *"very strong anxiety"*).

The score of the trait single item (Study 1) was *M* = 1.48 (*SD* = 0.78). Thus, the mean score for the trait assessment was below the mean score for the state assessment. It is important to note that these scores can be directly compared due to the use of fully parallel items. If state levels are seen as "real" due to being directly measured in the situation of interest, it might be that trait-like assessments underestimate students' levels of boredom during tests. One main reason for underestimating levels of a construct in trait assessments is subjective beliefs (e.g., Goetz et al., 2013; Robinson & Clore, 2002). In the case of test boredom, students might feel that taking a test cannot be boring, which could lead to their underestimation of levels of test boredom in the trait assessment. In addition, it is important to note that our trait assessment of test boredom was related to math exams, which are usually graded and therefore are of great personal importance to students. However, in our studies we assessed state test boredom as experienced in low-stakes tests. Given the likely relatively low extrinsic value of such tests, the level of state test boredom may have been higher than if we had measured state boredom during high-stakes tests.

We have focused on boredom during mathematics tests. It is important to note that the perceived value of achievement in this domain is typically rather high as compared to other domains (Goetz et al., 2014; Haag & Goetz, 2012). Given that value reduces boredom, test boredom might be relatively rare in mathematics and more frequent in other domains. As such, the current estimates of test boredom may be conservative given that they were derived from assessments during math tests. In other words, if test boredom can be found in mathematics, it seems likely that it should also be experienced in other domains.

Antecedents of Test Boredom (H2) – Study 1

Non-Optimal Levels of Control

Both for the trait and state assessments, the findings are in line with our hypotheses on the relations between boredom and non-optimal levels of control. The trait and state boredom scores were each positively correlated with both perceived over- and underchallenge during the test. These results are in line with findings by Krannich et al. (2019) who examined trait boredom as experienced during typical lessons (high school). In all three academic domains investigated in their study (mathematics, German, French), class-related boredom was positively related to both being over- and underchallenged. Our study extends this finding to testing situations.

Intrinsic and Extrinsic Value

For both the trait and state assessments, the hypothesis about negative relations between perceived value and test boredom was supported. Students’ trait and state boredom scores were negatively correlated with both intrinsic and extrinsic value during the test. Our results are in line with previous studies examining boredom experienced in other academic settings, which also show negative correlations between boredom and facets of perceived value (e.g., Forsblom et al., 2021; Goetz et al., 2006; Pekrun et al., 2010, 2011).

Low levels of extrinsic value and, consequently, high levels of boredom might be of particular relevance in low-stakes testing. In recent years low-stakes testing was used more frequently. This trend may continue due to enhanced demands for accountability and evidence-based policy making, which typically rely on standardized low-stakes tests (a growing number of countries participate in low-stakes large-scale assessments; see, e.g., OECD, 2017). Prime examples are international student assessments such as the OECD Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), the Programme for the International Assessment of Adult Competencies (PIAAC), and the Early Grade Reading Assessment (EGRA). There are also high numbers of low-stakes tests that are often not labeled as such. Examples are homework assignments, preparation tests (e.g., for the Test of English as a Foreign Language (TOEFL), voluntary intelligence tests, clinical development tests for children, self-assessments (e.g., "quick quizzes" in self-help books), and different formative assessment techniques, such as clicker questions (e.g., with Kahoot) and two-stage assessments (e.g., receiving feedback on an essay which will then be graded in a second step). Test boredom may play an important role in all of these types of low-stakes assessments.

Test Boredom and Achievement Outcomes (H3, H4) – Studies 1 and 2

Math Test

Consistent with our abundance hypothesis (H4), we found that test boredom was differentially related to test achievement depending on boredom due to being over- versus underchallenged. Test boredom was negatively related to achievement on the math test when students worked on difficult tasks. In contrast, boredom and achievement were unrelated (Study 1) or even positively related (Study 2) when students work on easy tasks. Boredom during easy, underchallenging tasks may have less or even no effect on achievement because students have sufficient cognitive and motivational resources to complete the tasks anyway. However, when working on difficult tasks, students may be overchallenged, and some of their cognitive resources, which would be needed to successfully complete the task, would be consumed by experiencing boredom. In both of our studies, we found evidence to support this assumption. These findings are also in line with theory (e.g., Eysenck, 1992; Eysenck et al., 2007) and empirical findings (Ashcraft, 2002) on test anxiety, showing that anxiety is more detrimental for achievement when learners are working on complex and attention-demanding tasks.

An intriguing result of Study 2 was that boredom was positively related with test achievement for students with high levels of self-efficacy. The results of Study 1 also point to this pattern; we found a positive, though not significant, correlation between test boredom and test performance in the underchallenge situation in Study 1. This positive relation was unexpected but makes sense, as being underchallenged in mathematics can be assumed to be associated with a positive math self-concept. In fact, in their study with Swiss eleventh graders, Krannich et al. (2019) found positive correlations between underchallenge and academic self-concept in the domains of English, French, and mathematics, whereas the reported correlations between overchallenge and self-concept were negative. Boredom due to being underchallenged could be interpreted by students as an indicator of high competence (feeling as information; Schwarz & Clore, 1983), which may strengthen their self-confidence and thus contribute to the beneficial effects of self-concept on achievement (Marsh et al., 2018; Niepel et al., 2021). Such plausible mechanisms might be investigated in future studies.

Our results suggest that the strength of the relation between state test boredom and achievement would be underestimated if boredom due to over- or underchallenge were not considered separately in the analysis. Relatively strong negative effects of boredom due to overchallenge on achievement scores would not be detected. In fact, in Study 2, probably due to the opposing effects of test boredom due to over- and underchallenge on test performance, we found no significant overall relation between test boredom and test performance (i.e., when not accounting for different levels of challenge).

We found no significant relation between trait test boredom and achievement on the math test (Study 1). This result suggests that test boredom may be situation specific and, consequently, that generalized trait assessments may be relatively weak predictors of achievement on a specific single test.

Math Grades

Consistent with our hypotheses, we found significant negative correlations with students' math grades in Study 1. The relation between math grades and trait/state test boredom was r = -.22/-.29. This result is in line with the findings of the meta-analyses of Tze et al. (2016) and Camacho-Morles et al. (2021), in which mean correlations of $\bar{r}$ = -.24 and ρ = -.25 between boredom and academic outcomes were reported.

Although the present correlations between test boredom and academic achievement (i.e., test scores and grades) were not very high, it is important to note that students work on numerous tests during their academic career, which may entail strong cumulative effects over longer periods of time. This assumption is supported by evidence on reciprocal relations between boredom and academic achievement (e.g., Pekrun et al., 2014, 2017), which can result in vicious cycles of boredom and poor achievement. Thus, our results contribute to existing findings on boredom and achievement showing that the relations between both constructs as demonstrated in previous studies can be extended to test boredom.

Presumably because of the opposing effects of test boredom due to over- or underchallenge on test achievement, we found no significant relations between boredom and overall test performance in either study when not differentiating levels of challenge. However, this begs the question of why test boredom related negatively to math grades without such differentiation (Study 1). For trait test boredom, one explanation may be that math exams in school are, on average, more likely to be over- than underchallenging, leading to lower grades according to the abundance hypothesis. This interpretation is supported by the results from Study 1, in which students reported much higher levels of over- than underchallenge. It is also supported by a study by Krannich et al. (2019) which found significantly higher levels of overchallenge than underchallenge in mathematics classes. The negative relation between state test boredom and grades could be explained in a similar way, as students also reported higher levels of overchallenge compared to underchallenge when taking our test. Thus, our test may have reflected the average math test performance level in school which tends to be overchallenging for students. The high prevalence of overchallenging situations at school may also be the reason for the overall negative relations between boredom and achievement found in the meta-analyses cited earlier.

Integrating Test Boredom into Research on Academic Boredom/Academic Emotions

In summary, our results show that test boredom occurs at significant levels. Furthermore, like other types of boredom (i.e., class- and learning-related boredom; Pekrun et al., 2010), test boredom has clear links with theoretically hypothesized antecedents and effects according to the findings. Thus, it is reasonable to include test boredom in the domain of academic boredom. Our results suggest that test boredom is quite similar to other types of boredom in terms of its component structure and its relations to antecedent and outcome variables. This does not mean, however, that it is not important to evaluate it as a separate construct. On the contrary, test boredom deserves specific attention, given that testing situations are very common and that test scores seem to be influenced by boredom.

Whether the abundance hypothesis, which has been confirmed for test boredom in the present research, also holds for other types of boredom is an open research question. Test boredom could differ more or less from other types of boredom in its effects on performance. In general, considering test boredom may broaden the perspective on boredom in academia, but also outside of school (e.g., in sports, arts, business; see Bieleke, Wolff, & Martarelli, in press).

From a broader perspective, future studies could consider test boredom along with other test emotions to identify similarities and differences (e.g., test-related anxiety, anger, hopelessness, joy, and pride; Pekrun et al., 2011). This could also be done at the component level of test emotions (e.g., Lange & Zickfeld, 2021). Based on the results of our research, it can be hypothesized that test boredom shows similar relations with other test-related emotions as class- and learning-related boredom show with other emotions during classes and learning.

Limitations

Some limitations of the present study should be noted and can be used to derive directions for future research. First, concerning the assessment of test boredom and its appraisals antecedents, we relied on self-report data, which may have resulted in common method bias (Podsakoff et al., 2003). Although we used a real-time assessment method for the state assessment, it was still a self-report. To control for possible biases, future studies may add more objective assessments of boredom or at least of its components (e.g., physiological assessments of reduced arousal; see Pekrun, 2023; Roos et al., 2021).

Second, because we focused exclusively on test boredom, we cannot draw conclusions about how test boredom differs in its magnitude, component structure, antecedents, and effects from other types of academic boredom, such as class- and learning-related boredom. Future studies could analyze the structure of different types of academic boredom (i.e., test-related, class-related, and learning-related boredom) to explore to what extent test boredom differs from other types of boredom. This should also be done for different types of testing situations, some of which are relatively similar to class or learning situations (i.e., low-stakes tests as addressed in the present research). In this regard, the scales developed in our study (TBS-Trait, TBS-State) could be used in combination with the class- and learning-related boredom scales of the AEQ (Pekrun et al., 2011).

Third, although our research used different indicators of non-optimal challenge (i.e., subjective experiences of over- and underchallenge, self-efficacy), which is a strength of this research, future studies could analyze how these different assessments might differ in terms of predicting effects of test boredom. Future studies could also consider other theories on the antecedents and effects of boredom (e.g., Eastwood et al., 2012) and include related variables in their studies (e.g., attention problems, creativity).

Fourth, in our study we could not directly assess whether optimal challenge was associated with very low or even no boredom experiences (Study 1 did not include medium difficulty tasks, whereas Study 2 included them but didn't measure boredom related to different task difficulty). Future studies could address this issue.

Fifth, our approach does not allow for conclusions on the causal ordering of variables. Future studies in this field may combine assessment of short-term dynamics with developments over a longer time periods (e.g., by using measurement-burst designs; Sliwinski, 2008) in order to model growth processes and their causal antecedents and effects. In this context, future studies should also examine within-person relations between test boredom and its antecedents and outcomes. Multilevel structural equation modeling could be used for simultaneous analyses of between- and within-person relations.

Finally, we have focused on a single academic domain, namely, mathematics. As mentioned earlier, we decided to focus on this domain because it typically has high subjective value and thus it permitted us to test the occurrence of test boredom in a conservative way. Furthermore, mathematics is a core area of STEM subjects, and test boredom and its impact on performance outcomes can have a strong influence on educational and career choices, as well as on motivation for lifelong learning in these subjects (Wigfield, Battle, Keller, & Eccles, 2002). However, future studies may also focus on other academic domains, such as languages, history, arts, and sports.

Implications for Research and Practice

Our study has several implications for research on boredom in academic settings and for educational practice. First, our results suggest that promoting students' competence beliefs (e.g. through appropriate types of feedback; Goetz et al., 2018) and increasing their perceptions of the value of tests may reduce their experiences of test boredom. However, it is important to note that enhancing extrinsic value can increase other negative emotions, such as anxiety, anger, and hopelessness (Pekrun, 2006). Thus, including tasks with high intrinsic value may be helpful to reduce boredom without giving rise to other negative emotions. To help educators reduce their students' test boredom, future studies may build upon our work by exploring additional ways to reduce or avoid boredom in testing situations. A challenge for such future studies may be to find ways to avoid including tasks that are too easy or too difficult without compromising the diagnostic properties of the test. For example, computerized adaptive (tailored) testing (CAT; e.g., Asseburg & Frey, 2013; Wainer, 2000) may be helpful to reduce situations of non-optimal challenge during tests. In CAT, items are individually selected depending on the test takers' previously shown responses. Thus, having given a wrong answer prompts the selection of an easier item to be presented next, and vice versa.

Second, in order to understand the cognitive mechanisms generating the effects on performance as explained by the abundance hypothesis, future studies could refer to cognitive load theory (Sweller, 2011). Such studies could incorporate measures of cognitive load (e.g., intrinsic and extrinsic cognitive load) in addition to measures of non-optimal challenge and test boredom.

Third, our multi-item test boredom scales (i.e., TBS-Trait, TBS-State) could easily be adapted to investigate the role of boredom in academic domains other than mathematics (e.g., Goetz et al., 2007). Although single-item measures of test boredom are likely to be the best choice for studying test boredom in the vast majority of cases, multi-item scales can be useful when the research question relates to components of test boredom, for example (for related research on components of test anxiety, see Roos et al., 2021, 2022).

Fourth, future studies could examine boredom in different testing situations (i.e., low-stakes vs. high-stakes testing). In high-stakes tests, assessing boredom and other constructs while students are working on a test could be problematic, as boredom assessments could compromise test outcomes for some students. However, test boredom assessments could be administered immediately after the test. For high-stakes tests in particular, it might be helpful to also include an assessment of anxiety to analyze the relations between boredom and test anxiety, as well as examine possible joint effects of both constructs on achievement outcomes.

Fifth, test boredom may be assessed above and beyond academic contexts, for example, at work, in sports, and in the performing arts. For instance, it is plausible that sport activities and competitions can be characterized by individuals' non-optimal experiences of control in a way similar to test situations at school, potentially giving rise to test boredom and impairing performance (e.g., Velasco & Jorda, 2020). In line with this argument, there have been recent calls to investigate boredom in the context of physical activity and sports as well as initial evidence for its relevance (Wolff et al., 2021).

Sixth, our newly formulated abundance hypothesis may be further investigated in future studies. Our finding that being over- versus underchallenged may moderate effects of boredom should be taken into account when designing studies on boredom. Also, meta-analyses of the relations between boredom and achievement may consider over- and underchallenge as moderators of this relation.

Finally, our research on test boredom completes the picture on the overall negative relations of academic boredom with achievement outcomes (see Camacho-Morles et al., 2021; Tze et al., 2016). Educators, parents, and students should be informed about these findings, especially in light of the empirically unfounded but frequently communicated argument that boredom in school has its good sides (see Vodanovich, 2003). Boredom, especially related to tests, is often viewed as a nonexistent or "silent" emotion (Pekrun et al., 2010). Our research has shown that it is anything but “silent” in terms of its occurrence and effects, so we invite researchers and practitioners to be mindful of it when designing their studies and instructional activities.

Appendix A: Test Boredom Scales

Table A1

Test Boredom Scale – Trait (TBS-Trait)

Nr.	English	German
1 (a)	I'm bored in math exams.	In Mathearbeiten bin ich gelangweilt.
2 (a)	In math exams everything seems monotonous and dull to me due to boredom.	In Mathearbeiten erscheint mir vor Langeweile alles eintönig und grau.
3 (a)	I'm bored to death in math exams.	In Mathearbeiten langweile ich mich zu Tode.
4 (c)	I'm so bored during math exams that I find myself daydreaming.	In Mathearbeiten bin ich so gelangweilt, dass ich mich beim Tagträumen ertappe.
5 (c)	I find my mind wandering in math exams.	In Mathearbeiten bin ich mit den Gedanken woanders.
6 (c)	I can't concentrate in math exams because I'm so bored.	In Mathearbeiten kann ich mich nicht konzentrieren, weil ich so gelangweilt bin.
7 (m)	I'm so bored that I would prefer not to start the math exams at all.	In Mathearbeiten würde ich vor lauter Langeweile am liebsten gar nicht erst anfangen.
8 (m)	In math exams I frequently look at my watch because time does not pass.	In Mathearbeiten schaue ich ständig auf die Uhr, weil die Zeit nicht vergeht.
9 (m)	In math exams I would like to leave the classroom out of boredom.	In Mathearbeiten würde ich aus Langeweile das Klassenzimmer am liebsten verlassen.
10 (p)	I start yawning in math exams because I'm so bored.	In Mathearbeiten muss ich vor Langeweile gähnen.
11 (p)	I get so bored in math exams that I get tired.	In Mathearbeiten langweile ich mich so, dass ich ganz matt werde.
12 (p)	I get so bored I have problems staying alert in math exams.	In Mathearbeiten kann ich mich vor Langeweile kaum noch wachhalten.

Note. (a) Affective, (c) cognitive, (m) motivational, (p) physiological component of boredom.

Table A2

Test Boredom Scale – State (TBS-State)

Nr.	English	German
Concurrent Assessment		
1 (a)	I'm bored.	Ich bin gelangweilt.
2 (a)	Everything seems monotonous and dull to me due to boredom.	Vor Langeweile erscheint mir alles eintönig und grau.
3 (a)	I'm bored to death.	Ich langweile mich zu Tode.
4 (c)	I'm so bored that I find myself daydreaming.	Ich bin so gelangweilt, dass ich mich beim Tagträumen ertappe.
5 (c)	My mind is wandering.	Ich bin mit den Gedanken woanders.
6 (c)	I can't concentrate because I'm so bored.	Ich kann mich nicht konzentrieren, weil ich so gelangweilt bin.
7 (m)	I'm so bored that I would prefer not to the math exams at all.	Vor lauter Langeweile würde ich am liebsten gar nicht erst mit den Matheaufgaben anfangen.
8 (m)	I frequently look at my watch because time does not pass.	Ich schaue ständig auf die Uhr, weil die Zeit nicht vergeht.
9 (m)	I would like to leave the classroom out of boredom.	Aus Langeweile würde ich das Klassenzimmer am liebsten verlassen.
10 (p)	I'm yawning because I'm so bored.	Vor Langeweile muss ich gähnen.
11 (p)	I'm so bored that I am tired.	Ich langweile mich so, dass ich ganz matt werde.
12 (p)	I am so bored I have problems staying alert	Vor Langeweile kann ich mich kaum noch wachhalten.
Retrospective Assessment		
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
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.

References

American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for educational & psychological testing*. Washington, DC: Author.

Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science, 11*, 181–185. https://doi.org/10.1111/1467-8721.00196

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.

Barry, C., L., Horst, S. J., Finney, S. J., & Brown, A. R. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342-363. doi: 10.1080/15305058.2010.508569

Bieleke, M., Ripper, L., Schüler, J., & Wolff, W. (2021). Boredom is the root of all evil - or is it? A psychometric network approach to individual differences in behavioral responses to boredom. *PsyArXiv*. https://doi.org/10.31234/osf.io/mje7v

Bieleke, M. Wolff, W., & Martarelli, C. (in press). *Handbook of Boredom Research*. New York: Routledge.

Brunswik, E. (1952). *The conceptual framework of psychology*. Oxford, England: University Chicago Press.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Publishers.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper and Row.

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

Dicintio, M. J., & Gee, S. (1999). Control is the key: Unlocking the motivation of at-risks students. *Psychology in the Schools, 36*, 231-7. https://doi.org/10.1002/(SICI)1520-6807(199905)36:3%3C231::AID-PITS6%3E3.0.CO;2-%23

Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press.

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

Eysenck, M.W., & Calvo, M.G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition and Emotion, 6,* 409–434. https://doi.org/10.1080/02699939208409696

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion, 7*, 336-353. https://doi.org/10.1037/1528-3542.7.2.336.

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]. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Schulen/Publikationen/Downloads-Schulen/schnellmeldung-schueler-5211003208004.htm

Forsblom, L., Pekrun, R., Loderer, K., & Peixoto, F. (2021). Cognitive appraisals, achievement emotions, and students' math achievement: A longitudinal analysis. *Journal of Educational Psychology.* Advance online publication. https://doi.org/10.1037/edu0000671

Gaspard, H., Dicke, A.-L., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology, 107*, 663-677. https://doi.org/10.1037/edu0000003

Geiser, C., Goetz, T., Preckel, F., & Freund, P. A. (2017; Eds.). States and traits - theories, models, and assessment. Special Issue. *European Journal of Psychological Assessment*, 33.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74-78. http://dx.doi.org/10.1016/j.paid.2016.06.069

Goetz, T. (2004). *Emotionales Erleben und selbstreguliertes Lernen bei Schülern im Fach Mathematik*. Utz.

Goetz, T., & Frenzel, A. C. (2006). Phänomenologie schulischer Langeweile [Phenomenology of boredom at school]. *Zeitschrift für Entwicklungspsychologie und Pädagogische*

Psychologie, 38(4), 149-153. https://doi.org/10.1026/0049-8637.38.4.149

Goetz, T., Frenzel, A. C., Pekrun, R., Hall, N. C., & Lüdtke, O. (2007). Between- and within-domain relations of students' academic emotions. *Journal of Educational Psychology, 99*(4), 715-733. https://doi.org/10.1037/0022-0663.99.4.715

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., Frenzel, A. C., & Pekrun, R. (2007). Regulation von Langeweile im Unterricht. Was Schuelerinnen und Schueler bei der 'Windstille der Seele' (nicht) tun [Regulation of boredom in class. What students (do not) do when experiencing the 'Windless Calm of the Soul']. *Unterrichtswissenschaft*, *35*(4), 312–333. https://doi.org/10.25656/01:5499

Goetz, T., & Hall, N. C. (2020). Emotion and achievement in the classroom. In J. Hattie and E. M. Anderman (Eds.), *Visible Learning Guide to Student Achievement* (pp. 145-152). Routledge.

Goetz, T., Krannich, M., Gogol, K., & Roos, A.-L. (2017). *Students' creativity, emotions and physiology*. Application to the Committee of Research (AFF) for Performance-based Funding. University of Konstanz, Germany.

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., Lipnevich, A. A., Krannich, M., & Gogol, K. (2018). Performance feedback and emotions. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge Handbook of Instructional Feedback* (pp. 554-574). Cambridge: Cambridge University Press.

Goetz, T., Nett, U. E., Martiny, S. E., Hall, N. C., Pekrun, R., Dettmers, & Trautwein, U. (2012). Students' emotions during homework: Structures, selfconcept antecedents, and achievement outcomes. *Learning and Individual Differences, 22*(2), 225-34. https://doi.org/10.1016/j.lindif.2011.04.006

Goetz, T., Pekrun, R., Hall, N. C., & Haag, L. (2006). Academic emotions from a social-cognitive perspective: Antecedents and domain specificity of students' affect in the context of Latin instruction. *British Journal of Educational Psychology, 76*(2), 289-308. https://doi.org/10.1348/000709905X42860

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,* 3-16. https://doi.org/10.1016/j.lindif.2006.12.002

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

Graf, T., Harych, P., Wendt,, W., Emmrich, R., & Brunner, M. (2016). Wie gut können VERA-8-Testergebnisse den schulischen Erfolg am Ende der Sekundarstufe I vorhersagen? [How well can state-wide proficiency tests predict school success at the end of secondary school]. *Zeitschrift für Pädagogische Psychologie / German Journal of Educational Psychology, 30*(4), 201–211. https://doi.org/10.1024/1010-0652/a000182

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and mimic approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66,* 373–388. doi:10.1007/BF02294440

Hintze, J. L. & Nelson, R. D. (1998) Violin plots: A box plot-density trace synergism. *The*

American Statistician, 52, 181-184. https://doi.org/10.1080/00031305.1998.10480559

Kendeou, P. (2021). Enhancing research excellence through diversity and transparency. *Journal of Educational Psychology, 113*(1), 1-2. http://dx.doi.org/10.1037/edu0000652

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65**,** 457-474. https://doi.org/10.1007/BF02296338

Krannich, M. Goetz, T., & Lipnevich, A. (2016, April). *The effects of boredom due to being over- or underchallenged on students' occupational choice intentions*. [Conference presentation]. Annual Meeting of the American Educational Research Association, Washington, DC.

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. doi:10.1016/j.lindif.2018.10.004

Krannich, M., Goetz, T., Roos, A.-L., & Lipnevich, A. A. (2020). *Boredom fosters creativity in specific cases: Examining the boredom-creativity link by including students' level of over- or underchallenge* [Paper submitted for publication].

Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods, 21*, 369-387. https://doi.org/10.1037/met0000093

Li, Y., Wang, K., Xiao, Y. et al. (2020). Research and trends in STEM education: a systematic review of journal publications. *International Journal of STEM Education, 7*(11). doi:10.1186/s40594-020-00207-6

Lichtenfeld, S., Pekrun, R., Stupnisky, R. H., Reiss, K., & Murayama, K. (2012). Measuring students’ emotions in the early years: The Achievement Emotions Questionnaire-Elementary School (AEQ-ES). *Learning and Individual Differences, 22*(2), 190–201. https://doi.org/10.1016/j.lindif.2011.04.009

Mandler, G. and Sarason, S.B. (1952). A study of anxiety and learning. *Journal of Abnormal and Social Psychology, 47*, 166-173. doi:10.1037/h0062855

Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018). An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over 6 years. *Developmental Psychology, 54*(2), 263–280. https://doi.org/10.1037/dev0000393

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology, 111*, 331–353. https://doi.org/10.1037/edu0000281

Murayama, K., Pekrun, R., Lichtenfeld, S., & vom Hofe, R. (2013). Predicting long-term growth in students’ mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development, 84*, 1475–1490. http://dx.doi.org/10.1111/cdev.12036

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

Niepel, C., Marsh, H. W., Guo, J., Pekrun, R., & Möller, J. (2021). Revealing dynamic relations between mathematics self-concept and perceived achievement from lesson to lesson: An experience-sampling study. *Journal of Educational Psychology.* Advance online publication. http://dx.doi.org/10.1037/edu0000716

Organization for Economic Cooperation and Development [OECD] (2017). *PISA 2015 results (Volume 3): Students' well-being.* Paris, France: Author.

Organization for Economic Cooperation and Development [OECD] (2018). *Education 2030. The Future of Education and Skills*. OECD Publishing. https://pisa2021-maths.oecd.org/#Twenty-First-Century-Skills.

Organization for Economic Cooperation and Development [OECD] (2019). PISA 2018 Mathematics Framework, in *PISA 2018 Assessment and Analytical Framework*, OECD Publishing, Paris, https://doi.org/10.1787/13c8a22c-en.

Pajares, F., & Graham, L. (1999): Self-Efficacy, Motivation Constructs, and Mathematics Performance of Entering Middle School Students. *Contemporary Educational Psychology, 24*, 124-139. https://doi.org/10.1006/ceps.1998.0991

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. (2023). Mind and body in students' and teachers' engagement: New evidence, challenges, and guidelines for future research. *British Journal of Educational Psychology.*

Advance online publication. https://doi.org/10.1111/bjep.12575

Pekrun, R., & Goetz, T. (in press). How universal are academic emotions? A control-value theory perspective. In G. Hagenauer, R. Lazarides, & H. Järvenoja, *Motivation and emotion in learning and teaching across educational contexts: Theoretical and methodological perspectives and empirical insights*. Earli book series "New Perspective on Learning and Instruction".

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., Perry, R. P., Kramer, K., Hochstadt, M., & Molfenter, S. (2004). Beyond Test Anxiety: Development and Validation of the Test Emotions Questionnaire (TEQ). *Anxiety, Stress and Coping*, *17*(3), 287-316. doi:10.1080/10615800412331303847

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., 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., Hofe, R. vom, Blum, W., Frenzel, A. C., Goetz, T., & Wartha, S. (2007). Development of mathematical competencies in adolescence: The PALMA longitudinal study. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report of the DFG Priority Programme* (S. 17-37). Waxmann.

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (2017). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child Development, 88*, 1653-1670. https://doi.org/10.1111/cdev.12704

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

Pekrun, R., Murayama, K., Marsh, H. W., Goetz, T., & Frenzel, A. (2019). Happy fish in little ponds: Testing a reference group model of achievement and emotion. *Journal of Personality and Social Psychology, 117*(1), 166-185. https://doi.org/10.1037/pspp0000230

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*, 667–686. http://dx.doi.org/10.1037/0022-0663 .95.4.667

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *The Journal of applied psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Raccanello, D., Brondino, M., Moè, A., Stupnisky, R., & Lichtenfeld, S. (2019) Enjoyment, boredom, anxiety in elementary schools in two domains: Relations with achievement. *The Journal of Experimental Education, 87*(3), 449-469, https://doi.org/10.1080/00220973.2018.1448747

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin, 128*, 934–960. https://doi.org/10.1037/0033-2909.128.6.934

Roos, A.-L., Goetz, T., Krannich, M., Donker, M., Bieleke, M., Caltabiano, A., & Mainhard, T. (2022). Control, anxiety and test performance: Self-reported and physiological indicators of anxiety as mediators. *British Journal of Educational Psychology*, 00, 00–00. https://doi.org/10.1111/bjep.12536

Roos, A.-L., Goetz, T., Krannich, M., Jarrell, A., Donker, M., & Mainhard, T. (2021) Test anxiety components: An intra-individual approach testing their control antecedents and effects on performance. *Anxiety, Stress, & Coping, 34*(3), 279-298. https://doi.org/10.1080/10615806.2020.1850700

Ryan, R. M., & Deci, E. I. (2009). Promoting self-determined school engagement: Motivation, learning, and well-being. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 171–195). New York, NY: Routledge.

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

Schwarz, N., & Clore, G. L. (1983): Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45,* 513-523. https://doi.org/10.1037/0022-3514.45.3.513

Sliwinski, M. J. (2008). Measurement-burst designs for social health research. *Social and Personality Psychology Compass, 2*(1), 245-261. https://doi.org/10.1111/j.1751-9004.2007.00043.x

Skinner, E. A. (1996). A guide to constructs of control. *Journal of Personality and Social Psychology, 71,* 549-570. https://doi.org/10.1037/0022-3514.71.3.549

Sticca, F., Goetz, T., Bieg, M., Hall., N. C., Eberle, F., & Haag, L. (2017). Examining the accuracy of students' self-reported academic grades from a correlational and a discrepancy perspective: Evidence from a longitudinal study. *PLoS ONE 12*(11): e0187367. https://doi.org/10.1371/journal.pone.0187367

Struk, A. A., Scholer, A. A., & Danckert, J. (2021) Perceptions of control influence feelings of boredom. *Frontiers in Psychology, 12*, 687623. https://doi.org/10.3389/fpsyg.2021.687623

Sweller, J. (2011). Cognitive load theory. In J. Mestre & B. Ross (Eds.), *The psychology of learning and motivation: Cognition in education,* Vol. 55, pp. 37–76. Oxford: Academic Press.

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

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.

Vodanovich, S. J., & Watt, J. D. (2016) Self-Report Measures of Boredom: An Updated Review of the Literature*, The Journal of Psychology, 150*(2), 196-228, https://doi.org/10.1080/00223980.2015.1074531

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A Primer* (2nd Edition). Lawrence Erlbaum Associates.

Westgate, E. C., & Wilson, T. D. (2018). Boring thoughts and bored minds: The MAC model of boredom and cognitive engagement. *Psychological Review, 125,* 689–713. http://dx.doi.org/10.1037/rev0000097

Wigfield, A., Battle, A., Keller, L. B., & Eccles, J. S. (2002). Sex differences in motivation, self-concept, career aspiration, and career choice: Implications for cognitive development. In A. McGillicuddy-De Lisi & R. de Lisi (Eds.), *Biology, society, and behaviour. The development of sex differences in cognition* (pp. 93–124). Ablex.

Wolff, W., Bieleke, M., Martarelli, C. S., & Danckert J. (2021) A primer on the role of boredom in self-controlled sports and exercise behavior. *Frontiers in Psychology, 12*, 637839. https://doi.org/10.3389/fpsyg.2021.637839

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

Table 1

Correlations Among Antecedent and Outcome Variables – Study 1

Variable	1.	2.	3.	4.	5.	6.
Trait[a]						
1. Overchallenge						
2. Underchallenge	-.14					
3. Intrinsic Value	**-.27**	**.23**				
4. Extrinsic Value	-.01	.07	**.24**			
5. Math Score	**-.30**	**.26**	.10	.00		
6. Academic Achievement (grades)	**-.44**	**.24**	**.20**	**.20**	**.44**	
M (*SD*)	2.34 (0.97)	1.71 (0.81)	3.10 (1.13)	4.02 (0.86)	45.53 (13.66)	3.40 (0.91)
State[b]						
1. Overchallenge						
2. Underchallenge	-.05					
3. Intrinsic Value	-.21	**.20**				
4. Extrinsic Value	-.19	.07	**.56**			
5. Math Score	**-.30**	**.44**	**.18**	**.04**		
6. Academic Achievement (grades)	**-.22**	**.23**	**.14**	**.14**	**.46**	
M (*SD)*	2.50 (0.70)	1.88 (0.56)	2.50 (0.861)	2.96 (1.01)	46.66 (14.51)	3.43 (0.92)

Note. [a] Single-level modeling. [b] Multilevel modeling (measures within persons for overchallenge, underchallenge, intrinsic value, extrinsic value). *N* = 180 students for Trait boredom (due to one whole class not participating in the trait-assessment and missing data from 5 students) and *N* = 208 students for State boredom. For the assessment of challenge and value participants responded using a 5-point rating scale ranging from 1 (*not at all true*) to 5 (*completely true*). Grades ranged from 1 (*very good*) to 6 (*insufficient*). For the ease of the interpretation, we inverted grade scores so that higher numbers indicated better performance. **Bold** coefficients: *p* < .05.

Table 2

Means, Standard Deviations, and Cohen’s d for Test Boredom Measures – Study 1

	Single Item			Scale - Overall Score		
	M	*SD*	*d*	*M*	*SD*	*Latent d*
Trait Boredom	1.48	0.78	0.61	1.44	0.50	0.89
State Boredom						
Before easy part (current)	2.02	1.06	0.96	1.49	0.48	1.04
After easy part (retrospective)	1.62	0.95	0.65	1.32	0.48	0.67
Before difficult part (current)	1.80	1.10	0.73	1.47	0.63	0.75
After difficult part (retrospective)	1.99	1.26	0.78	1.63	0.84	0.75
End of study (current)	2.10	1.33	0.83	1.73	0.90	0.81

Note. *N* = 180 for trait test boredom (due to one whole class not participating in the trait-assessment and missing data from 5 students) and *N* = 208 for state test boredom. Participants responded using a 5-point rating scale ranging from 1 (*not at all*) to 5 (*very strongly*). Model fit for trait boredom scale: χ^2 (50) = 79.04, CFI = .946, TLI = 0.929, RMSEA = 0.057, SRMR = 0.050; model fit for state boredom scale: χ^2(1520) = 2856.14, CFI = .849, TLI = 0.824, RMSEA = 0.065, SRMR = 0.066.

Table 3

Correlations between Multi-Item Boredom Measures, Control, and Value – Study 1

	Non-optimal control		Value	
	Trait			
Variable	Over-challenge	Under-challenge	Intrinsic value	Extrinsic value
Trait test boredom	**.17** [.07, .31]	**.27** [.09, .49]	**-.18** [-.33, -.03]	**-.23** [-.36, -.08]
	State			
	Over-challenge	Under-challenge	Intrinsic value	Extrinsic value
State test boredom	**.44** [.16, .67]	**.25** [.01, .42]	**-.17** [-.31, -.04]	**-.29** [-.41, -.14]

Note. Analyses of trait and state data are based on multi-level modeling (state: measures nested within persons). For both the trait and the state data between-person correlations are shown. *N* = 180 students for trait boredom (due to one whole class not participating in the trait-assessment and missing data from 5 students) and *N* = 208 students for state boredom. Model fit for the trait assessment: $\chi^2(92) = 142.61$, CFI = .937, TLI = 0.918, RMSEA = 0.055, SRMR = 0.055; model fit for state assessment: $\chi^2(202) = 429.51$, CFI = .953, TLI = 0.944, RMSEA = 0.033, $\text{SRMR}_{\text{Within}}$ = 0.057, $\text{SRMR}_{\text{Between}}$ = 0.065. **Bold** coefficients: $p < .05$. 95% confidence intervals are shown in brackets.

Table 4

Correlations between Multi-Item Boredom Measures and Achievement – Study 1

Variable	Math Test	Math grades
Trait test boredom [a]	.01[-.11,.13]	**-.22** [-.40, -.03]
State test boredom [b]	-.08 [-.23, .07]	**-.29** [-.48, -.08]
State test boredom – easy part [c]	.09 [-.08, .25]	
State test boredom – difficult part [c]	**-.22** [-.30, -.14]	

Note. All coefficients are between-person correlations. [a] Analyses are based on single-level modeling. [b] Analyses are based on multilevel modeling (measures nested within persons); the coefficients are Level 2 correlations. [c] Analyses are based on single-level modeling; there was only one state assessment of boredom for each part of the test. Correlations between state test boredom and the corresponding math score in each part (easy vs. difficult part) are shown. **Bold** coefficients: $p < .05$. 95% confidence intervals are shown in brackets. Model fit for Trait test boredom/Math Test: $\chi^2(59) = 77.33$, CFI = .971, TLI = 0.961, RMSEA = 0.039 , SRMR = 0.051; model fit for Trait test boredom/Math grades: $\chi^2(59) = 79.72$, CFI = .967, TLI = 0.956, RMSEA = 0.044 , SRMR = 0.051; model fit for State test boredom/Math Test: χ^2 (121) = 272.60, CFI = .961, TLI = 0.954, RMSEA = 0.035 , $\text{SRMR}_{\text{Within}}$ = 0.043, $\text{SRMR}_{\text{Between}}$ = 0.059; model fit for State test boredom/Math grades: $\chi^2(121) = 266.16$, CFI = .962, TLI = 0.955, RMSEA = 0.034 , $\text{SRMR}_{\text{Within}}$ = 0.043, $\text{SRMR}_{\text{between}}$ = 0.059; model fit for State test boredom – easy part/Math Test: $\chi^2(61) = 112.23$, CFI = .944, TLI = 0.929, RMSEA = 0.064 , SRMR = 0.044; model fit for State test boredom – difficult part/Math Test: $\chi^2(61) = 133.99$, CFI = .923, TLI = 0.902, RMSEA = 0.076, SRMR = 0.037.

Table 5

Means and Standard Deviations: State Test Boredom – Study 2

Boredom score	*n*	*M*	*95% CI*	*SD*	*p* [a]	Cohen's *d*
Beginning of the test	1,586	1.74	[1.68, 1.79]	1.06	< .001	0.69
After Part 1	1,556	1.77	[1.71, 1.82]	1.12	< .001	0.69
After Part 2	1,351	1.95	[1.88, 2.02]	1.28	< .001	0.75
After Part 3	1,574	1.90	[1.84, 1.96]	1.19	< .001	0.76

Note. Participants responded using a 5-point rating scale ranging from 1 (*not at all*) to 5 (*very strongly*). [a] Adjusted *p*-values using Bonferroni correction for multiple testing.

Table 6

Latent Correlations Between Study Variables – Study 2

Variable	1.	2.	3.
1. Test boredom			
2. Self-efficacy	-.02 [-.10, .07]		
3. Test scores	.03 [-.04, .10]	**.33** [.26, .40]	
M (*SD*)	1.81 (0.89)	5.58 (1.34)	0.00 (0.99)

Note. For test boredom, *M* and *SD* across the four single-item ratings of state test boredom are shown (answer format: 5-point rating scale from 1 = *not at all true* to 5 = *completely true*). For self-efficacy *M* and *SD* are reported for the three-item scale (answer format: 8-point Likert scale ranging from 1 = *not confident at all* to 8 = *completely confident*). The test score is based on a standardized mathematics test. χ^2 (18) = 87.96, CFI = .974, TLI = 0.960, RMSEA = 0.049, SRMR = 0.020. **Bold** coefficients: *p* < .05. 95% confidence intervals are shown in brackets.

Table 7

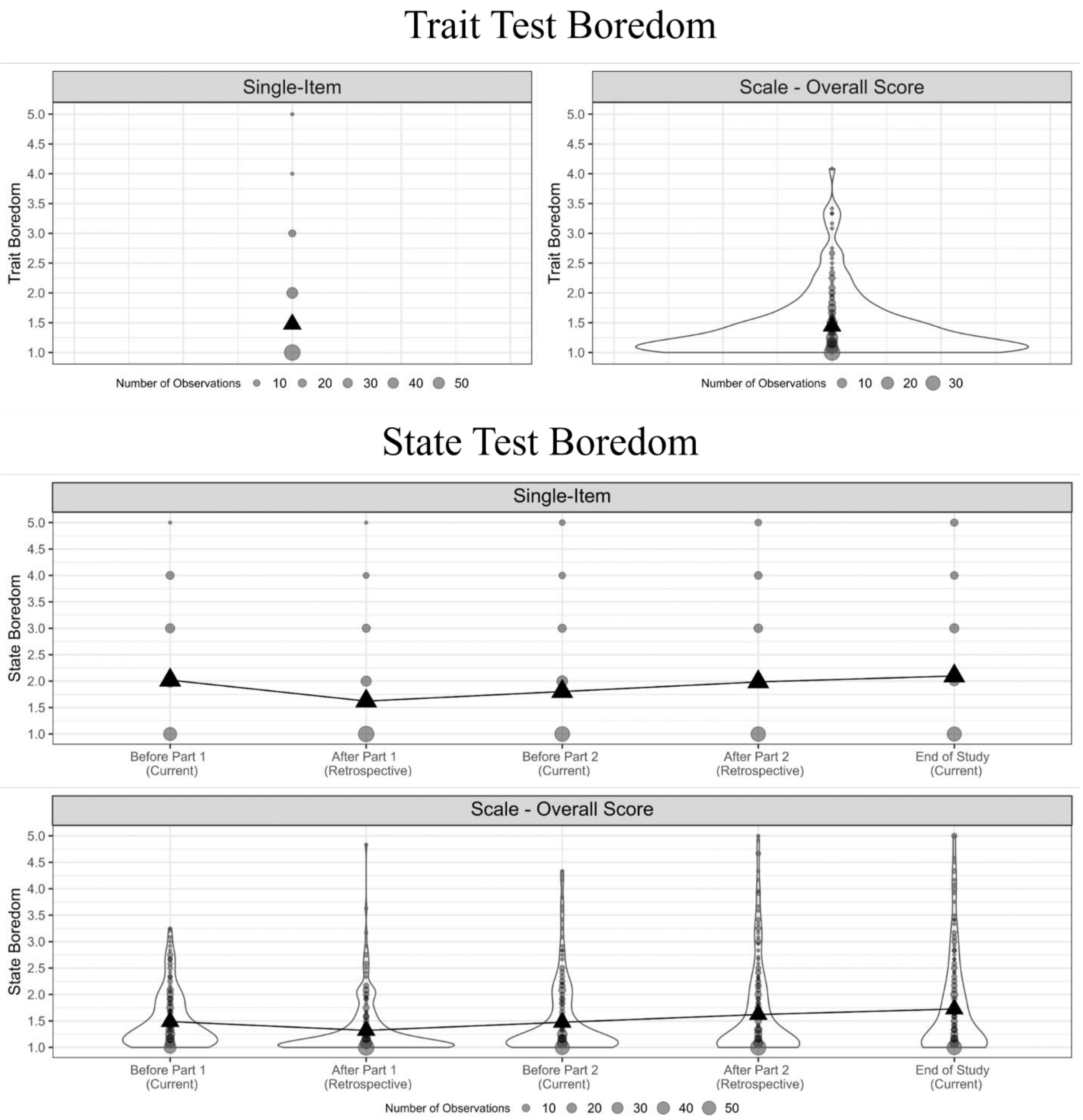
Math Test Scores Predicted by Test Boredom and Self-Efficacy – Study 2

	Test score		
	Unstandardized Effect	95% CI	Standardized Effect
Boredom	012	[-0.09, 0.33]	0.04
Self-efficacy	**0.33**	[0.25, 0.38]	**0.32**
Boredom x Self-efficacy	**0.38**	[0.19, 0.58]	**0.13**

Note. Fixed factor variance approach was used for model identification of the measurement model for the predictor self-efficacy, i.e., the conditional effect of boredom is tested at the average level of the predictor self-efficacy; predictor boredom was centered at the mean. **Bold** coefficients: *p* < .05. 95% confidence intervals are shown in brackets.

Figure 1

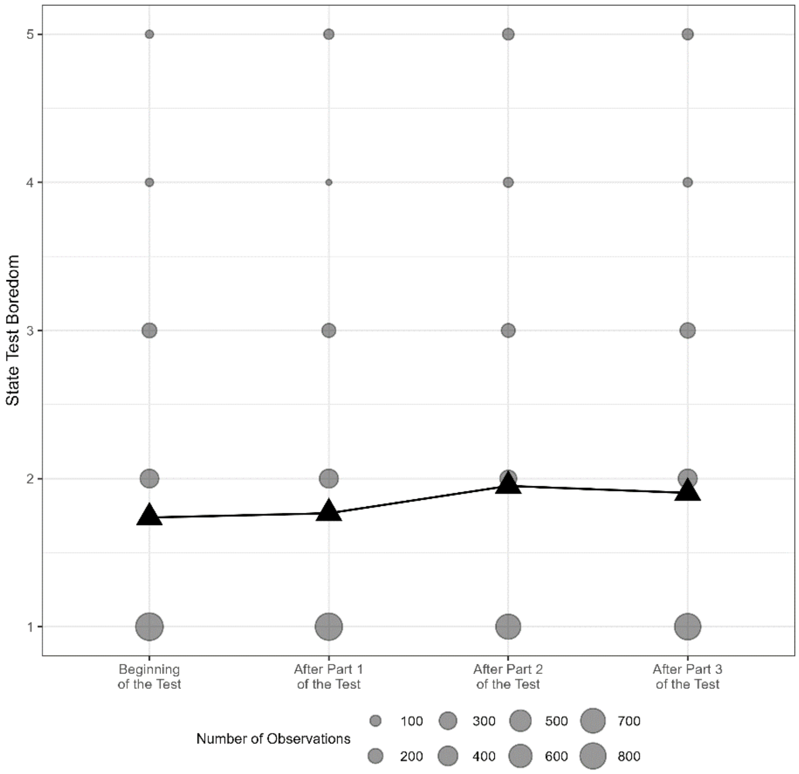
Mean Values and Violin Plots for Trait and State Test Boredom – Study 1



Note. Circles represent individual values, where sizes of the circles are relative to the number of observations. Filled triangles represent mean values. Violin plots show a rotated density plot on each side smoothed by a kernel density estimator (Hintze, 1998).

Figure 2

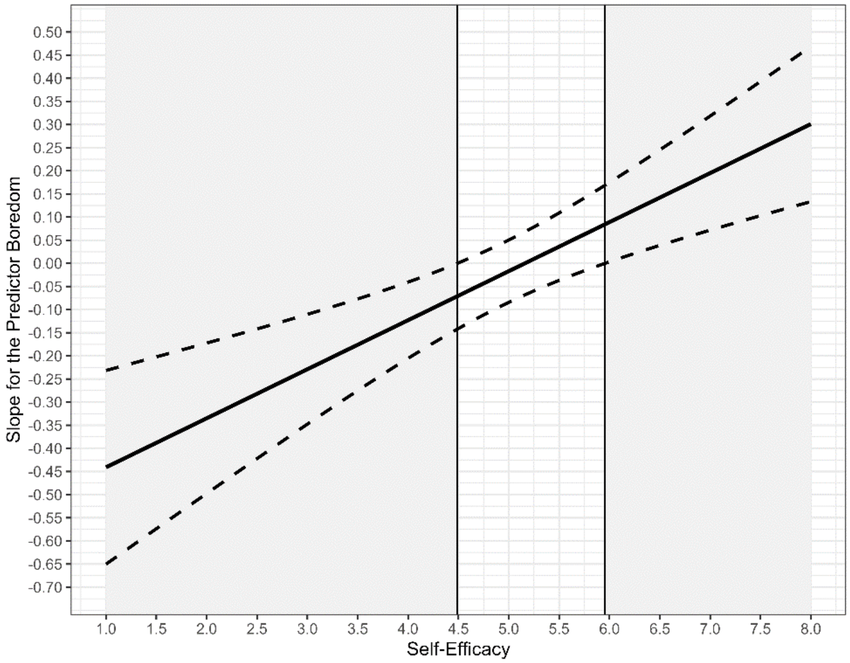
Mean Values for State Test Boredom – Study 2



Note. Circles represent individual values, where sizes of the circles are relative to the number of observations. Filled triangles represent mean values.

Figure 3

Johnson-Neyman Plot: Slope of the Effect of Boredom on Test Achievement as a Function of Self-Efficacy – Study 2



Note. The dashed lines represent the lower and upper bounds of the 95% confidence bands. The dark gray areas represent the regions of statistical significance for effects of boredom at α = .05. The mean of the self-efficacy score is *M* = 5.58 (*SD* = 1.34, Min = 2.37, Max = 7.89). The slope for low self-efficacy (- 1 *SD*) can be seen at self-efficacy = 4.24, and the slope for high self-efficacy (+ 1 *SD*) can be seen at self-efficacy = 6.92.