Enjoyable for some, stressful for others? Physiological and subjective indicators of achievement emotions during adaptive versus fixed-item testing

Enjoyable for Some, Stressful for Others? Physiological and Subjective Indicators of Achievement Emotions during Adaptive versus Fixed-Item Testing

Miriam Wünsch[a], Anne C. Frenzel[a], Reinhard Pekrun[b, c, a], and Luning Sun[d]

[a] Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 München, Germany

[b] Department of Psychology, University of Essex, Wivenhoe Park, University Of Essex, Valley Rd, Colchester CO4 3SQ, UK

[c] Institute for Positive Psychology and Education, Australian Catholic University, PO Box 968 North Sydney NSW 2059, Australia

[d] The Psychometrics Centre, University of Cambridge, Cambridge Judge Business School, Trumpington Street, Cambridge CB2 1AG, UK

Author Note

Miriam Wünsch https://orcid.org/0009-0009-5019-9576

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

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

Luning Sun https://orcid.org/0000-0002-2470-4278

We thank Andreas Frey for valuable feedback on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Miriam Wünsch, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 München. Email: Miriam.Wuensch@psy.lmu.de

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Funding Sources

Department of Psychology, Ludwig-Maximilians-Universität München

Physiological and Subjective Indicators of Achievement Emotions during Adaptive versus Fixed-Item Testing

The use of computerized adaptive tests is on the rise. Adaptive tests are used more frequently than ever in educational assessments, including high-stakes exams such as the Standardized Aptitude Test in the USA (College Board, 2022) as well as large-scale international assessments like the Programme for International Student Assessment (PISA; OECD, 2022). While there is ample and consistent evidence that the use of adaptive testing yields increased efficiency and accuracy compared to classical fixed-item testing (Frey & Ehmke, 2007), its effects on test-takers' emotional experiences remain underexplored. Given that emotions have been shown to substantially influence test performance (Pekrun et al., 2023) and school-related well-being (Obermeier et al., 2022), it is crucial to ensure that the increased use of adaptive testing does not negatively impact test-takers' emotional situation.

The present study contributes to this goal by employing a within-person, repeated-measures design to explore test-takers' emotional experiences during adaptive testing compared to classical fixed-item testing, as well as potential interactions of test type with ability and perceived test difficulty. We add to previous research by investigating not only self-reported emotions, but also psychophysiological states as indicated by test-takers' skin conductance

response. Doing so allows for a differentiated picture of the effects of adaptive testing on test-takers' affective state.

Principles of Adaptive Testing

In classical fixed-item testing, each individual is presented with the same set of items of varying difficulty. In adaptive testing, instead, an underlying algorithm continuously estimates participants' ability and selects the next item accordingly (Thompson & Weiss, 2011; Weiss, 2004). Usually, test difficulty is set to 50%, as this maximizes test efficiency and precision (Wise, 2014). This procedure implies that items are selected in a way that the test-taker has a 50% probability of solving them correctly based on the algorithm's current estimate of their ability (Thompson & Weiss, 2011). Following this procedure, adaptive testing is more efficient than fixed-item testing, with tests requiring only 53-57% of items to achieve the same precision of ability estimates as a fixed-item test (Frey & Ehmke, 2007). In practice, adaptivity of a test usually implies that solving an item correctly is followed by the presentation of a more difficult item, while an easier item follows incorrect answers. As a result, individuals with higher ability face relatively difficult items, and individuals with lower ability encounter relatively easy items. In contrast, for fixed-item testing, all participants get the same set of items with identical difficulties. Accordingly, items on a fixed-item test should be easier to solve for individuals with high ability than for those with low ability. As different individuals encounter different sets of items in adaptive testing, item-response-theory is used to obtain a final ability estimate, as opposed to the typical use of sum scores in fixed-item testing (Frey & Ehmke, 2007).

While adaptive testing outperforms fixed-item testing in terms of efficiency, it is crucial to ensure that these advantages do not happen at the expense of test-takers' affective and motivational experiences. The present study compares adaptive and fixed-item testing in terms of test-takers' achievement emotions, specifically psychophysiological arousal and subjective affective experience, driven by differential experiences of test difficulty.

Achievement Emotions and Their Link with Test Difficulty

The emotions test-takers experience during testing are likely achievement emotions, which are defined as emotions that occur in situations "judged according to competence-based standards of quality" (Pekrun et al., 2023, p. 146). They are "multicomponent processes, with components loosely coupled" (Pekrun et al., 2023, p. 146), which comprise the subjective affective experience as well as motivational tendencies, expressive behavior, cognitive appraisals, and psychophysiological processes (Pekrun, 2006).

Effects of tests on test-takers' emotions can be explained using Pekrun's (2006, 2021, 2024) control-value theory of achievement emotions. This theory posits that subjective control, that is, an individual's perceived causal influence over actions and outcomes, and subjective value, that is, perceived intrinsic and extrinsic value of the activity or outcome, interact in predicting achievement emotions. According to the theory, positive achievement emotions are prompted by high levels of subjective control and value. Negative achievement emotions are triggered by a lack of control, combined with high value (except for boredom, which should be reduced by high value; Pekrun et al., 2023).

We propose that the use of an adaptive testing format would not have any systematic effect on the perceived value of a given test (e.g., the value of a college entrance exam should be similarly high for an individual, independent of whether it is an adaptive or fixed-item test). Control, in contrast, may differ between an adaptive and a fixed-item test, based on individuals'

perceptions of difficulty while taking the respective test. In fixed-item testing, control perceptions should vary considerably between individuals, depending on their ability: Test-takers with low ability should experience a fixed-item test as relatively hard and, therefore, less controllable, whereas test-takers with high ability should experience the same test as relatively easy, and thus more controllable. In adaptive testing, instead, the difficulty of each item is adapted to the person's ability, supposedly leading to perceptions of difficulty and corresponding experiences of control being similar for all persons (Betz & Weiss, 1976). We propose that differences in perceived control due to different perceptions of difficulty in adaptive and fixed-item tests drive effects of test type on affective states, with differential effects on the psychophysiological and subjective components of achievement emotions.

Psychophysiological Arousal

Psychophysiological arousal refers to activation in the autonomic nervous system, which is divided into the sympathetic and parasympathetic branches (Kreibig & Gendolla, 2014). The sympathetic nervous system is activated in situations that stimulate the individual. Activation of this system is related to stress (Weissman & Mendes, 2021), attention (e.g., Zhang et al., 2021), and cognitive load (e.g., Nourbakhsh et al., 2017; Vanneste et al., 2021). Activation of the parasympathetic nervous system is associated with relaxation and recovery (Weissman & Mendes, 2021). Whereas parameters such as heart and respiration rates are influenced by both branches, the eccrine sweat glands are only innervated by the sympathetic nervous system, making them a good indicator of sympathetic arousal without parasympathetic influences (Braithwaite et al., 2015; Ishikawa, 2023; Kreibig & Gendolla, 2014). Eccrine sweat glands are the root of electrodermal activity (EDA), as changes in sweat production cause fluctuations in the skin's conductance and electric potentials (Christopoulos et al., 2019).

EDA is commonly captured by attaching two electrodes to the skin, usually the palm of the hand, and measuring the level of conductance between them when applying a constant current (Boucsein et al., 2012). The resulting signal can be divided into a tonic and a phasic component: The tonic component, termed skin conductance level (SCL), changes rather slowly over time. The phasic component – the skin conductance response (SCR) – responds to stimuli more quickly, with a latency of around one to four or five milliseconds. Due to SCR showing faster changes in response to stimuli compared to SCL, the present study focuses on SCR. SCR is visible as sudden increases (i.e., peaks) in skin conductance (Boucsein et al., 2012; Christopoulos et al., 2019).

Psychophysiology and Achievement Emotions

Both SCR and SCL have been shown to positively relate to the intensity of positively and negatively valenced emotions, likely triggered by emotion-inherent action tendencies that prepare the body physiologically for approach or avoidance behaviors (Kreibig, 2010). Within the educational-psychological literature, scattered studies have reported on the relationships between EDA and self-reported emotions. Although we still lack a clear understanding of their association (Horvers et al., 2021), it appears that psychophysiological responses are related, yet not equivalent, to subjective emotional experiences. Hence, the inclusion of SCR as a measure of sympathetic physiological arousal adds a new layer of understanding to the question of how adaptive testing influences test-takers' emotional experiences.

SCR in Adaptive and Fixed-Item Testing

As outlined earlier, perceptions of difficulty might determine differences in emotions between adaptive and fixed-item testing. Relative to fixed-item testing, taking an adaptive test can either constitute a gain or a loss in terms of the perceived ease of solving problems and related perceptions of control. We propose that in both cases, the adaptive test will be accompanied by a higher SCR than the fixed-item test due to the better fit between test difficulty and the individual's ability level. When test items are too difficult for an individual, they report less effort and more boredom (Asseburg & Frey, 2013). Thus, individuals for whom the fixed-item test is clearly too difficult and some items are unsolvable (i.e., individuals with lower ability) may disengage from the task and show weaker arousal, indicated by a decrease in SCR. In contrast, on the adaptive test, which better matches their ability level, these individuals might stay engaged with the task, showing higher levels of SCR.

Instead, individuals for whom the fixed-item test is relatively easy (i.e., individuals with higher ability) would encounter more difficult items in the adaptive test. Although harder, the items on the adaptive test are still solvable for these individuals, given that item difficulty adapts to their ability level. Thus, they would not disengage in the adaptive test but instead experience a moderate level of perceived control, which might also increase their arousal compared to the fixed-item test. Taken together, since SCR indicates physiological arousal irrespective of valence, we hypothesize that adaptive testing leads to higher SCR relative to fixed-item testing.

Subjective Experiences of Achievement Emotions

Prior Findings

In addition to psychophysiological arousal, we also investigated the subjective emotional experience in adaptive versus fixed-item testing. In early work in this field, Betz and Weiss (1976) hypothesized an overall positive effect of adaptive testing on test-takers' affective experiences, due to the higher fit between test-takers' ability and test difficulty. They posited that this would lead to high-ability individuals being less bored and low-ability individuals being less stressed and frustrated in adaptive than in fixed-item testing. Intriguingly, Betz and Weiss did not appear to consider the possibility of inverse effects, particularly increased stress and frustration in high-ability individuals. In fact, their initial claim did not hold up in their empirical investigation, as they found a main effect of test type with increased anxiety in adaptive testing for all participants (Betz & Weiss, 1976).

Since these findings were published, considerable advances have been made regarding – now computerized – adaptive testing. One might argue that adaptive testing should have generally beneficial effects on the emotional experience during test taking today. However, more recent research further challenges the notion of a uniformly positive effect of adaptive testing on subjective affective experiences: Two meta-analyses by Akhtar et al. (2022) and Frey et al. (2024) comparing adaptive and fixed-item testing concluded that there were no significant differences in self-reported emotions between the two test types. While Akhtar et al. (2022) focused on the experience of test anxiety, Frey et al. (2024) investigated negative and positive emotions, noting that more studies on distinct emotions are needed for a more in-depth understanding. Both meta-analyses found test difficulty to be a central factor. When the adaptive test was set to an average success rate higher than 50% in individual studies, participants reported less test anxiety (Akhtar et al., 2022) and generally less intense negative emotions (Frey et al., 2024) than in fixed-item testing. However, in these meta-analyses, difficulty was only assessed with regard to the adaptive test. Not coded was how this feature differed between the adaptive and fixed-item test, that is, whether the adaptive test was more or less difficult than the fixed-item test. With this information missing, it is possible that it is not the adaptivity of an

easier adaptive test that drives the more positive emotional experience compared with a fixed-item test, but simply the fact that the adaptive test happened to be easier, and therefore more controllable, than the fixed-item test. This is in line with control-value theory, according to which different perceptions of difficulty of the two tests should determine which of them is accompanied by more intense positive or negative emotional experiences. Therefore, we propose that asking for a main effect of adaptive versus fixed-item testing on subjective emotional experience is too simplified. Instead, the difference should depend on a combination of features of both tests as well as the test-taker.

Considering Ability and Difficulty Perceptions

To fully grasp the potential effects of adaptive versus fixed-item testing on subjective emotional experiences, we propose that it is essential to consider test-takers' ability and their perceptions of the relative difficulty of the two tests. In fixed-item testing, people with higher ability should perceive items as less difficult than people with lower ability, and hence have a more positive emotional experience (see Goetz et al., 2007, for supporting evidence). In contrast, in adaptive testing, perceptions of difficulty would be independent of test-takers' ability, as item difficulty is adapted to the individual's ability (Betz & Weiss, 1976). These assumptions are supported by Akhtar and Kovacs' (2024) findings showing a significantly positive correlation between individuals' ability and their perception of performing well for a fixed-item test. In contrast, this correlation was non-significant if the test was adaptive. As such, when contrasting adaptive versus fixed-item testing within individuals, test type and personal ability should interact in their effects on test difficulty: For low-ability individuals, the adaptive test would be easier than the fixed-item test; for medium-ability individuals, both tests should be similar in difficulty; and for high-ability individuals, the adaptive test should be harder than the fixed-item test. This central role of relative difficulty, which depends on the interaction of the pre-defined difficulties of the two tests with the individual's ability, might explain the lack of consistency in findings from studies directly comparing adaptive and fixed-item testing: Findings on the effects of test type might take different directions depending on test difficulties and ability levels in the sample. In line with this reasoning, a few studies have considered ability. The results were not consistent: Some studies found that ability was a significant moderator of the relation between test type and value, effort, perceived probability of success, and feelings of satisfaction (e.g., Betz & Weiss, 1976; Ortner et al., 2013, 2014). Others found no such moderating effect for anxiety as an outcome (e.g., Betz & Weiss, 1976; Ling et al., 2017).

Importantly, these deliberations rest on the assumption that a person's objective ability and resulting differences in test difficulties translate into subjective perceptions of difficulty and control. However, this assumption may not always be correct, especially for adaptive tests. As reported by Ortner et al. (2013), metacognitions may not accurately represent test performance in adaptive testing, as the number of items solved correctly is not an indicator of the final ability estimate. Furthermore, individuals usually do not receive feedback on their performance, which may lead to some holding overly optimistic or pessimistic views of their performance and the controllability of both tests, irrespective of objective difficulty level. As such, individuals might subjectively perceive one test as more difficult than the other, although based on their ability and the resulting objective item difficulties, the opposite would be the case.

Consequently, a possible reason explaining the lack of consistent empirical support for ability as a moderator for the effects of adaptive versus fixed-item testing on affective outcomes might be a discrepancy between objective and subjective difficulty. This is supported by Powell's (1994) finding that not actual, but only perceived performance determined test-takers' preferences for a certain test type. Similarly, perceived performance mediated the relationship

between objective performance and metacognitions of difficulty, effort, and satisfaction (Ortner et al., 2013), as well as between objective performance and motivation (notably, however, not between perceived performance and anxiety; Tonidandel et al., 2002). Taken together, it might not be the relative objective difficulty of an adaptive and a fixed-item test that determines emotional experience. Even more important might be the relative perceived difficulty of the two tests. In line with control-value theory, we therefore sought to test the assumption that whichever test the individual perceives as easier, hence more controllable, will elicit a more favorable emotional experience.

The Present Study

While the superiority of adaptive over fixed-item testing in terms of psychometric efficiency seems undisputed (Frey & Ehmke, 2007), potential effects of adaptive testing on test-takers' affective states are underexplored. The present study aims to contribute to this literature in three ways: First, by complementing the classical mode of inquiry through self-report by a psychophysiological measure; second, by systematically considering the interaction between test type and person characteristics for the subjective emotional experiences; and third, by applying a within-person instead of between-person experimental design.

Self-report is prone to response sets and memory biases, and it covers only the subjective aspect of the emotion process. Nevertheless, to the best of our knowledge, no study on adaptive testing has used psychophysiological measures to date. Based on other studies using skin conductance measures as indicators of achievement emotions, we chose SCR as the psychophysiological outcome of interest, indicating sympathetic arousal (e.g., Kiuru et al., 2022; Roos et al., 2023). We assumed that an adaptive test would be accompanied by higher levels of SCR than a fixed-item test.

Furthermore, previous studies typically considered adaptivity as the only difference between adaptive and fixed-item tests, or at best included either test takers' ability or characteristics of only the adaptive test in their investigations. The existing research neglected the possibility that differences between the tests on features other than adaptivity, such as their difficulties, likely influence emotional responses via control perceptions. Therefore, in the present study, we considered participants' ability and their perceptions of the relative difficulty of the two types of tests as possible moderators of effects of test type on achievement emotions.

Another possible reason for the lack of consistency in empirical findings is the predominant use of between-group experimental designs. These designs are susceptible to a priori-group differences, specifically when samples are small and hence, sampling errors are large (see also Pekrun, 2023). Furthermore, they are not well suited to capture the within-person processes that generate emotions. Therefore, the present study used a within-person experimental design to investigate differences in affective states during a computerized adaptive test (CAT) versus a computerized fixed-item test (FIT). The within-person design ensured that person characteristics were held constant across the two conditions.

Succinctly stated, we tested the following hypotheses:

Hypothesis 1. Participants show higher psychophysiological arousal as indicated by SCR in the CAT compared to the FIT, as EDA is independent of emotional valence and the CAT should generally elicit stronger emotional arousal compared to the FIT.

Hypothesis 2. For self-reported emotions, we expected no main effects of test type due to the following disordinal interactions with ability and relative perceived difficulty.

2a. Ability moderates the effect of test type on self-reported emotions: Higher-ability individuals experience more negative and less positive emotions in the CAT than the FIT (because for them, the CAT should be harder/less controllable). Lower-ability individuals experience more intense positive and less intense negative emotions in the CAT than the FIT (because for them, the FIT should be harder/less controllable).

2b. The effect of test type on self-reported emotions depends on relative perceived difficulty, with more positive and less negative emotions in the test that is perceived as easier by the individual.

At first sight, it may seem counterintuitive to assume a main effect of test type on SCR, but interactions of test type with ability and relative perceived difficulty for self-reported emotions. However, although related, the physiological and subjective components of achievement emotions are not identical. Higher general arousal during the adaptive test does not conflict with specific emotions being experienced at a higher level in the fixed-item test, depending on ability and relative perceived difficulty. The combination of the two hypotheses demonstrates how different features of the emotional experience can be integrated to provide a more comprehensive picture of individuals' emotions during testing.

Methods

Ethics Statement and Data Transparency

The research reported herein was conducted in accordance with the APA ethical standards and has received a formal waiver of ethical approval by the ethics committee of the Faculty of Psychology and Educational Sciences at the University of Munich. Participation in the study was voluntary, and no identifiers that could link individual participants to their results were obtained. All participants provided informed consent. Data and analysis code are available on OSF (https://osf.io/ys7qx/?view_only=e895327efb9a4cab9ce704d1eed2d2ee).

Participants

The study was conducted at a large, research-oriented university in southern Germany. Participants were recruited via university mailing lists and social media postings. Of the $N = 89$ participants, 60 identified as female, 26 as male, and 2 as diverse. For one participant, demographic information was missing. Age ranged between 18 and 77 years with a mean of 26.57 years ($SD = 8.72$)[1]. Of the sample, 84% were students in different undergraduate and graduate degree programs. The remaining 16% were working or retired, with the majority also holding a university degree.

Based on an a-priori specified validation protocol, 19 participants were excluded due to being non-responders, that is, not showing reactions to external stimuli in their SCR (for information on non-responders, see e.g., Ikezawa et al., 2012; Venables & Mitchell, 1996). Furthermore, EDA has not been recorded for two participants due to technical issues, resulting in a sample of 68 persons for the analysis testing H1. For the analyses of self-reported emotions,

[1] Based on the broad age range, we checked the robustness of our findings in a reduced sample excluding six age-related outliers (defined as values of more than 1.5 interquartile ranges above or below the mean age). All effects related to Hypotheses 1, 2a, and 2b had the same significance and direction as in the full sample. As such, we report results from the full sample.

two participants had to be excluded due to an error in generating participant codes, resulting in a sample of 87 participants for analyses of self-reported emotions testing H2a and H2b.

Procedure

After participants had arrived at the lab and filled in a consent form, the experimenter attached two electrodes and a wristband to their non-dominant hand, which they placed on a foam block on the table. They were asked to follow the instructions on the screen and move as little as possible to avoid movement artifacts in the skin conductance recording. First, they underwent the validation protocol for the EDA recording, in which they were instructed to hold their breath and bite their tongue for ten seconds each. They then received the test instructions with the note to only use the hand without the electrodes. Aiming to increase the perceived value of the test, participants were told that the items measured numerical reasoning ability as one component of intelligence and that they would receive feedback on their performance at the end.

They were informed that there would be two separate tests, but not what the difference entailed. The order of presentation (adaptive first vs. fixed-item first) was counterbalanced across participants. Each test consisted of 12 numerical reasoning items, split into three blocks of four items. Items were not timed. Participants were instructed to enter "X" to proceed if they could not find a correct solution. After the first test, there was a break where participants were instructed to take a breath and relax. They could end the break and continue whenever they wanted by clicking a button on the screen. After the second test, they received information on how many of the total 24 items they had solved correctly. Since there was no time limit on the items and the break, time spent on the two tests, including the self-report ratings and the break, varied between participants. The mean duration was 41.8 minutes (SD = 12.9). Upon completion, participants were debriefed on the different test types and the purpose of the study and received either twelve euros or participant credits for their participation.

Measures

Numerical Reasoning Tests

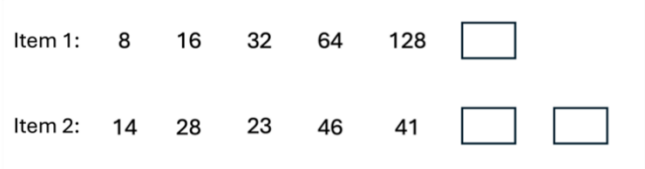
Both the CAT and the FIT consisted of twelve items assessing numerical reasoning ability. The items were rows of numbers, in which participants had to identify a pattern and complete the rows with one or two numbers accordingly (see examples in Figure 1). The items were presented on a computer screen. Due to counter-balancing the test order, 45 participants started with the CAT, and 44 started with the FIT. Both tests were based on the 49 numerical reasoning items generated by Loe et al. (2018). Of those, twelve items of varying difficulty were selected for the FIT in a way that the Rasch-scaled difficulty estimate would be above zero for half of the items and below zero for the other half, with a mean of 0.165. The goal of this selection process was to create an average success rate of 50% in the FIT for our sample, which primarily comprised university students with presumably above-average cognitive abilities. The twelve items of the FIT were presented with increasing difficulty. The remaining 37 items constituted the item pool for the CAT. The first item of the CAT had a Rasch-scaled difficulty of 0.11, and the following eleven items were selected using the maximum Fisher information criterion based on Bayes modal ability estimates with test difficulty set to 50%. Both tests were run on the Concerto Platform (The Psychometrics Centre, n.d.).

Skin Conductance Response

Participants' skin conductance was measured using two Shimmer EDA electrodes attached to the palm of the non-dominant hand connected to a Biopac BioNomadix wristband. The signal was transmitted to a Biopac MP160 receiver and Biopac Bionomadix 2CH GSR/EDA Amplifier and recorded in the software iMotions (iMotions, 2022) with a frequency of 500 Hz.

Figure 1

Exemplary Numerical Reasoning Items



Note. These items are examples of the type of items, not actual items used in the present study. Solutions are: "256" for Item 1, and "82; 77" for Item 2.

The experimenters continually recorded potential reasons for response artifacts during the experiment, such as participants talking or loud noises. In case of such events and corresponding visible artifacts in the signal, the time periods containing these artifacts were manually removed. Furthermore, the time periods in which participants filled in the self-report questionnaires were removed, so that the skin conductance signal only contained periods when participants were working on the numerical reasoning task. Using the Peak Detection Algorithm implemented in the iMotions software (see Table 1 for settings), a Peaks Per Minute (PPM) value was calculated for each participant on the CAT and the FIT, respectively.

Self-Reported Achievement Emotions and Relative Perceived Difficulty

After each item block, participants were instructed to fill in a pen-and-paper questionnaire placed in front of them, assessing their subjective affective state and perception of difficulty. To assess affective state, participants were presented with single-item statements for each emotion: "I am enjoying this/I feel proud/I feel angry/I feel bored/I feel stressed/I feel frustrated/I feel tense and nervous" and asked to indicate their endorsement ("Please choose the option that describes best how you are currently feeling") on a five-point Likert scale (1 = *completely disagree* to 5 = *completely agree*). A total score for each emotion was obtained by averaging the ratings across all three time points per test. In case of only one missing value per test, the mean score of the remaining two time points was used.

Table 1

Settings for Peak Detection Algorithm in iMotions Software

Phasic Filter Length	8000 ms
Lowpass Filter Cutoff Frequency	5 Hz
Peak Onset Threshold	0.01 microSiemens
Peak Offset Threshold	0 microSiemens
Peak Amplitude Threshold	0.01 microSiemens
Minimum Peak Duration	500 ms

Note. We used the default settings implemented in the iMotions software, with the exception of Peak Amplitude Threshold, which we set manually based on recommendations by Boucsein (2012).

Participants' emotion ratings varied across blocks, likely due to the increasing item difficulty in the FIT. Nevertheless, the ratings were highly consistent in each of the two tests, as indicated by high Cronbach's Alphas across the three time points per test (CAT: .93/.83./.85/.87/.87/.93; FIT: .91/.85/.77/.88/.82/.92 for joy/pride/anger/boredom/frustration/anxiety, respectively).

At each of the three self-report time points per test, participants were further asked to complete the statement "To me, the tasks are...," with the five response options: "very easy" (1), "rather easy" (2), "neither easy nor hard" (3), "rather hard" (4), "very hard" (5). From all three time points per test, a mean score of perceived difficulty was calculated for the CAT and FIT, respectively (Cronbach's Alpha .77 for CAT and .68 for FIT). Based on these mean scores, a relative perceived difficulty score was calculated for each individual by subtracting the perceived difficulty of the FIT from the perceived difficulty of the CAT. Hence, values of relative perceived difficulty below zero imply that the participant found the CAT to be easier than the FIT, and values above zero that the participant perceived the FIT to be easier than the CAT.

Ability Estimate

To obtain ability estimates, participants' performance on all 24 items of the two tests was considered. Given that item difficulties were available for all items, a Rasch model could be applied for estimating each participant's ability score, resulting from which items they had solved correctly across both tests. We used the thetaEst function of the catR package (Magis & Raîche, 2012) to obtain these ability estimates.

Analyses

We accounted for the within-person design by estimating multilevel linear regression models with random intercepts. To test Hypothesis 1 on physiological arousal, we estimated a multilevel linear regression model with PPM as the outcome and test type (CAT vs. FIT) as well as time (first vs. second test) as predictors. To test Hypotheses 2a and 2b, we specified the same multilevel linear regression models, one for each of the discrete emotion scores as outcomes. For Hypothesis 2a, we additionally included ability as well as a term for the cross-level interaction between ability and test type as predictors. For Hypothesis 2b, we added relative perceived difficulty and its interaction with test type as predictors. Our primary focus was on the interaction effects, presumably showing different directions of the effect of test type depending on ability and relative perceived difficulty, respectively. To determine statistical significance, we used $\alpha = .05$.

Results

Descriptive Statistics and Manipulation Check

Descriptive information for all variables, separately for each test type, can be found in Table 2. On average, participants showed five to six peaks per minute in their SCR while performing the tests. Further, while the items for enjoyment, pride, anger, frustration, and anxiety were endorsed, on average, just below the mid-point of the scale, endorsement was lowest for the boredom items, with a total average towards the lower end of the self-report scale. Participants' Rasch ability estimates based on their performance on all 24 items ranged from -2.08 to 3.24, with a mean of 0.17 (*SD* = 1.09), which is slightly above the population average of 0 based on the item calibration described in Loe et al. (2018). It is worth noting that the average Rasch-scaled item difficulty was -.12 lower on the CAT than on the FIT. This difference, although small in size, was significant, as indicated by a paired-sample t-test, $t(88) = -2.07$, $p = 0.041$. In line with this finding, test scores (i.e., the number of correctly solved items) were slightly higher on the CAT compared to the FIT, $t(88) = 2.10$, $p = .039$. This difference needs to be considered when interpreting the effects of test type on physiological and self-reported outcome variables. Finally, the average item endorsement for the difficulty judgement was just above the mid-point of the scale for both tests. The measure of relative perceived difficulty of the CAT and the FIT varied quite symmetrically around zero, ranging between -1.33 and 1.33 with a mean of 0.05 (*SD* = 0.61).

To gain a better understanding of the relative difficulty perceptions, we split the sample into three subgroups: $n = 33$ participants who perceived the CAT as easier than the FIT, $n = 35$ who perceived the FIT as easier than the CAT, and $n = 19$ for whom the two tests had the same mean perceived difficulty. Descriptively (see Table 3), the average ability level was highest in the group that found the FIT easier and lowest in the group that found the CAT easier. However, the differences in ability scores between the three groups were not statistically significant, as indicated by a one-way ANOVA, $F(2, 84) = 1.01$, $p = .368$. This finding supports the notion that a person's objective ability does not necessarily directly translate into their relative subjective experiences of difficulty in the different tests. Regarding objective difficulty, paired-sample *t*-tests within each subgroup showed no significant differences between objective difficulty on the CAT and FIT within the group that found the FIT easier, $t(34) = 1.02$, $p = .314$, and in the group that perceived the same level of difficulty, $t(18) = -0.27$, $p = .792$. In the group that perceived the CAT as easier, however, the objective mean difficulty level was indeed significantly lower on the CAT than on the FIT, $t(32) = -2.72$, $p = .011$.

Table 2

Descriptive Statistics per Test Type

	CAT		FIT	
	M (*SD*)	Min - Max	*M* (*SD*)	Min - Max
PPM (*n* = 68)	5.61 (3.15)	0.03 - 12.61	5.23 (2.79)	0.00 - 10.53
Joy (*n* = 87)	2.86 (1.22)	1.00 - 5.00	2.84 (1.16)	1.00 - 5.00
Pride (*n* = 87)	2.25 (1.01)	1.00 - 5.00	2.20 (0.95)	1.00 - 4.67
Anger (*n* = 87)	2.25 (1.00)	1.00 - 5.00	2.11 (0.88)	1.00 - 4.33
Boredom (*n* = 87)	1.72 (0.85)	1.00 - 4.67	1.65 (0.88)	1.00 - 5.00
Frustration (*n* = 87)	2.77 (1.04)	1.00 - 5.00	2.72 (1.03)	1.00 - 5.00
Anxiety (*n* = 87)	2.67 (1.06)	1.00 - 5.00	2.63 (1.10)	1.00 - 5.00
Obj. Difficulty (*n* = 89)	0.04 (0.58)	-1.52 - 1.54	0.16 (0.00)	0.16 - 0.16
Test Score (*n* = 89)	6.54 (2.19)	2.00 - 12.00	6.12 (2.58)	1.00 - 12.00
Perc. Difficulty (*n* = 87)	3.66 (0.72)	1.67 - 5.00	3.61 (0.65)	1.67 - 5.00

Note. Descriptive statistics per test type without considering test order. Sample size varies between variables, since data for some variables needed to be excluded for some participants (total sample *N* = 89; 68 included for physiological measure, 87 for self-report measures, full sample for information related to test difficulty and score; see section “Participants”). Objective difficulty was obtained by averaging the difficulty estimates of the 12 items per test. Test score is the number of correctly solved items out of the 12 items per test.

Paired-sample *t*-tests comparing scores (i.e., the number of correct responses) on the two tests within each group further revealed that scores were significantly higher on the CAT than on the FIT for the group that perceived the CAT as easier, $t(32) = 2.43$, $p = .021$, as well as in the group that perceived the tests as similar in difficulty, $t(18) = 3.01$, $p = .007$. In contrast, in the group

that perceived the FIT as easier, scores did not differ significantly between the two tests, $t(34) = -0.61$, $p = .549$. Altogether, relative perceived difficulty does not seem to consistently follow from either an individual's ability, the actual difficulty of the two tests, or participants' test scores, suggesting that it is a highly specific individual appraisal.

Table 3

Descriptive Statistics within Subgroups of Relative Perceived Difficulty

	Average item difficulty on CAT *M* (*SD*)	Ability *M* (*SD*)	Test score *M* (*SD*)	
			CAT	FIT
CAT perceived easier (*n* = 33)	-0.11 (0.59)	-0.07 (1.08)	6.24 (2.26)	5.52 (2.54)
Same perceived difficulty (*n* = 19)	0.13 (0.64)	0.18 (1.11)	6.79 (1.99)	5.63 (2.52)
FIT perceived easier (*n* = 35)	0.08 (0.50)	0.28 (0.98)	6.49 (2.15)	6.69 (2.37)

Next, we explored whether the CAT adaptivity algorithm built into the Concerto Platform (The Psychometrics Centre, n.d.) was indeed adaptive in terms of matching items to the individuals' abilities. To this end, we obtained a correlation between the average objective, Rasch-scaled item difficulty in the CAT, and participants' ability estimate. We expected a strong relationship between the two variables, as the CAT should present more difficult items to examinees with higher levels of ability. Indeed, this correlation was $r = .86$, indicating that the CAT adapted the difficulty of the items to the examinee's ability level. Furthermore, as expected, there was a high correlation between ability and test score on the FIT ($r = .89$), showing that participants with higher ability solved more items than those with lower ability. However, counter to our expectation, ability and test score were also strongly related on the CAT ($r = .93$), implying that participants with higher ability still solved more items correctly than participants with lower ability, despite the adaptivity of the test. Thus, in our CAT, item difficulty was indeed adaptive to test-takers' ability level, but still, individuals with higher ability levels solved considerably more items correctly than individuals with lower ability levels.

Effects of Test Type on Psychophysiological Arousal (H1)

In Hypothesis 1, we expected higher physiological arousal during the CAT compared to the FIT. We used a multilevel linear regression model with test type and time as predictors and a random intercept per participant. The results support our one-sided hypothesis, with an average

of 0.30 PPM less in the FIT compared to the CAT ($p = .013$), controlling for time and allowing for random intercepts. Time itself was also a significant predictor of psychophysiological arousal, with 1.26 PPM less in the second compared to the first test ($p < .001$). The fixed-effects intercept was 6.20 PPM ($p < .001$).

Effects of Test Type and Ability on Self-Reported Achievement Emotions (H2a)

Hypothesis 2a posited an interaction between test type and ability in predicting discrete achievement emotions. Table 4 shows the results from the random-intercept multilevel linear regression models with test type, time, ability, and the ability*test type interaction as predictors of test-takers' emotion scores. As expected, there were no significant main effects of test type on any of the self-reported emotions. Contrary to expectations, the interaction between ability and test type was also not significant for any of the emotions (Table 4). Independently of test type, time (first vs. second test) had a significant effect. Reported joy and anxiety scores were significantly lower, and boredom significantly higher during the second test. Likewise, independent of test type, ability significantly affected the emotions; participants with higher ability reported significantly more joy and pride, and less frustration and anger. For anxiety and boredom, no relationship with ability could be detected.

Effects of Test Type and Relative Perceived Difficulty on Self-Reported Achievement Emotions (H2b)

Hypothesis 2b posited an interaction between test type and relative perceived difficulty in predicting discrete achievement emotions. We used multilevel linear regression with a random intercept for all six self-reported emotion variables. Predictors were test type, relative perceived difficulty, time, and the interaction between test type and relative perceived difficulty. The results are displayed in Table 5.

Table 4

Results of Multilevel Linear Regression Analysis for H2a

	Joy	Pride	Anxiety	Frustra-tion	Boredom	Anger
Intercept	2.90***	2.29***	2.76***	2.75***	1.57***	2.26***
Test type: FIT	-0.02	-0.06	-0.02	-0.03	-0.06	-0.14
Time: Second test	-0.21**	-0.14	-0.19**	0.10	0.31***	0.05
Ability	0.49***	0.25*	0.01	-0.23*	-0.08	-0.27**
Test type*ability	0.01	0.10	-0.11	-0.13	-0.04	-0.05

Note. $N = 87$. Coefficients are unstandardized regression coefficients. Effects shown pertain to reference categories "CAT" and "First Test." *$p < .05$. ** $p < .01$. *** $p < .001$.

There were no significant main effects of test type or relative perceived difficulty on the emotions (with the exception of frustration, which was higher the more the FIT was perceived easier than the CAT). Time significantly predicted joy, pride, and anxiety (decreasing over time) as well as boredom (increasing over time), whereas frustration and anger remained stable over time. Most importantly, as expected, there were significant interactions between test type and relative perceived difficulty on all emotions except boredom. These interactions are visualized in Figure 2. In line with our hypotheses, the findings indicate that the test perceived as easier was accompanied by more positive and less negative emotions, relative to the other test. These differences were more pronounced with higher differences in perceived difficulty. That is, when perceived difficulty differed only slightly between the two tests, the emotional experience was more similar across the tests than when one test was perceived as much easier or harder than the other.

Table 5

Results of Multilevel Linear Regression Analysis for H2b

	Joy	Pride	Anxiety	Frustra-tion	Boredom	Anger
Intercept	2.98***	2.33***	2.75***	2.70***	1.55***	2.21***
Test type: FIT	-0.04	-0.07	-0.03	-0.02	-0.06	-0.13
Time: Second test	-0.22**	-0.15*	-0.19**	0.11	0.32***	0.06
Relative perceived difficulty	-0.29	-0.12	0.24	0.39*	0.20	0.20
Test Type* Relative perceived difficulty	0.57***	0.54***	-0.24*	-0.66***	-0.21	-0.43***

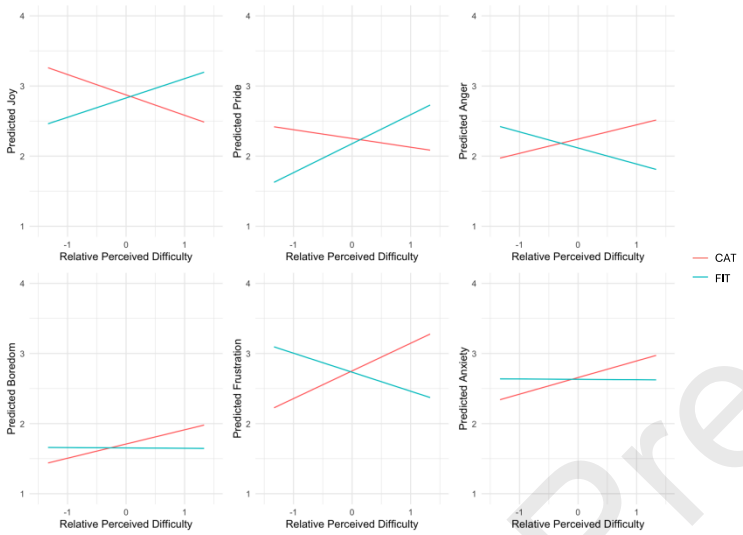
Note. $N = 87$. Coefficients are unstandardized regression coefficients. Effects shown pertain to reference categories "CAT" and "First Test." Relative Perceived Difficulty represents the

difference between difficulty on the CAT and the FIT, with values < 0 indicating that the CAT was perceived as easier and > 0 that the FIT was perceived as easier.

*p <.05. ** p <.01. *** p <.001.

Figure 2

Interactions between Relative Perceived Difficulty and Test Type in Predicting Self-Reported Achievement Emotions



Note. Relative Perceived Difficulty < 0 indicates that the CAT was perceived as easier and > 0 that the FIT was perceived as easier. For boredom, the interaction effect was not significant.

Discussion

Driven by the increasing use of adaptive testing in educational assessment, the present study investigated the question of whether and how adaptive testing influences test-takers' emotional experiences compared to classical fixed-item testing. As achievement emotions considerably impact performance (Pekrun et al., 2023) and school-related well-being (Obermeier et al., 2022), emotional experiences in adaptive testing need to be investigated to ensure that this more efficient way of testing is not accompanied by undesirable emotional effects on test-takers. To assess test-takers' affective states, we assessed their psychophysiological arousal (specifically, their SCR), alongside their self-reported discrete achievement emotions (joy, pride, anger, anxiety, frustration, and boredom). Regardless of whether the adaptive or non-adaptive test was administered first, we observed a significant decrease in physiological arousal and self-reported joy, pride, and anxiety, as well as a significant increase in self-reported boredom from the first to the second test. These findings suggest that over the course of a testing situation, physiological arousal and emotional activation decreased, while the experience of boredom as a deactivating emotion increased. To maintain participant engagement throughout a testing situation, it may therefore be advisable to keep tests as short as possible.

Regarding the effects of adaptive compared to non-adaptive testing, in line with expectations, a key finding was that test-takers were more strongly physiologically aroused while working on the adaptive test than during the fixed-item test. However, counter to our

expectations, test-takers with low ability did not benefit emotionally from receiving easier tasks in adaptive testing, nor did test-takers with high ability suffer emotionally from adaptive testing presenting them with objectively harder tasks. Yet, we did find support for our hypothesis that the subjective perception of the relative difficulty of the two tests impacted participants' emotional experiences: Whichever test was perceived as easier was accompanied by more joy and pride, and less frustration, anxiety, and anger. Hence, the key message of the present contribution is that while adaptive tests appear to elicit stronger arousal in participants, the subjective emotional experience seems to be driven by subjective perceptions of difficulty independent of the presence or absence of adaptivity. These results alleviate concerns regarding possible adverse effects of adaptive testing on test-takers' affective state.

Effects of Adaptive versus Fixed-Item Testing on Psychophysiological Arousal

To the best of our knowledge, this is the first study to investigate the effect of adaptive testing on test-takers' psychophysiological responses, thereby including an objective indicator of emotional arousal (Pekrun et al., 2023). Specifically, we measured test-takers' SCR as an indicator of sympathetic arousal, which is a process that prepares the individual for action (Rosebrock et al., 2017). In line with our expectations, this valence-independent level of psychophysiological arousal was higher during the adaptive test compared to the fixed-item test. We propose that this main effect of test type on psychophysiological arousal is driven by differential mechanisms depending on the individual's ability level: For individuals with lower ability, the fixed-item test might have exceeded their ability level to an extent that caused them to "switch off" (for similar findings, see Asseburg & Frey, 2013), resulting in generally lower arousal during this test. In contrast, the adaptive test offered them solvable items throughout, keeping them engaged and therefore leading to higher levels of arousal. For individuals with higher ability, the items on both tests were solvable, presumably supporting a certain engagement during both tests. The adaptive test challenged these participants more, thus resulting in higher arousal.

Furthermore, average item difficulty was slightly lower, and test scores were accordingly slightly higher in the CAT than in the FIT. This further supports the notion that the adaptive test being more engaging might result from this test providing both lower- and higher-ability participants with the opportunity to perform well. Since, however, our findings did not show a main effect of test type on the self-reported emotional experience, the higher levels of arousal during the adaptive test might stem from alternative engagement-related processes such as higher levels of attention (Zhang et al., 2021) or cognitive load (Nourbakhsh et al., 2017; Vanneste et al., 2021) rather than differences in emotional experience. This assumption aligns with the finding that more effort may be invested in taking an adaptive compared to a fixed-item test, with effort measured by reaction time (Akhtar & Kovacs, 2024).

Effects of Adaptive Versus Fixed-Item Testing on Subjective Emotional Experience

Regarding the subjective experience of discrete achievement emotions, we expected interactions between test type and test-takers' characteristics as predictors. Based on control-value theory (Pekrun, 2006), we hypothesized that emotional experiences in adaptive versus fixed-item testing would depend on relative perceived control related to these two tests: Whichever test the individual perceives as more controllable would be accompanied by a more positive and less negative emotional experience. From this perspective, the commonly asked question of whether emotional experiences differ between adaptive versus fixed-item testing appears too simple, given that the experience would depend on features of the two tests in relation to each other as well as features of the test-taker. We therefore investigated two possible

moderators of the effect of test type on self-reported emotions, namely, personal ability and relative perceived difficulty.

Our first interaction hypothesis was not supported, as participants' ability did not significantly interact with test type in predicting any of the discrete emotions. There were only main effects of ability, with higher-ability individuals reporting generally more positive emotions (joy and pride) and less negative emotions (frustration and anger) while working on both tests. Notably, though, we observed no main effects of ability on anxiety or boredom.

For interpreting the lack of an interaction, it needs to be noted that our reasoning that ability would moderate the effect of test type was built on the assumption that the adaptive test would be more difficult for higher-ability individuals. As such, we expected that higher-ability individuals would solve fewer items on the adaptive test than on the fixed-item test. In contrast, for lower-ability individuals, we expected the adaptive test to be easier, with more items solved correctly than on the fixed-item test. This assumption was partly supported, as the CAT indeed provided individuals with higher ability with more difficult items.

However, there was also an unexpected, strong correlation between ability and success rate (i.e., test scores) on the CAT. This means that despite receiving relatively easy items in the CAT, participants with lower ability only solved a few of them correctly, whereas participants with high ability solved many items correctly despite facing relatively difficult items in the CAT. It is unclear whether positive relations between ability and success rates are a common phenomenon in adaptive testing. Some studies reported non-significant correlations (e.g., Ortner et al., 2013; Tonidandel et al., 2002), some did not report success rates (e.g., Ortner et al., 2014), and others found high correlations, similar to the present findings (e.g., Ling et al., 2017). For the present study, it appears that the CAT has been adaptive, but not sufficiently so to achieve a similar level of success for all individuals, regardless of their ability. A possible reason might have been a lack of very easy or very hard items in the item bank. Another possible reason is the relatively low number of items in the test, so that the CAT may have been terminated before settling in at a difficulty level that would correspond to the pre-defined success rate of 50%. Therefore, while ability significantly predicted several emotions as expected, it may not have done so in interaction with test type because the experience of control has not differed as much as expected across the two test types at different levels of ability.

To investigate the second interaction hypothesis, we obtained a measure of relative perceived difficulty, indicating which test was perceived as easier and to what extent. We observed significant interactions between test type and relative perceived difficulty for all emotions except boredom. The more one test was perceived as easier than the other, the more joy and pride, and the less frustration, anxiety, and anger were experienced on this test compared to the other test. This finding supports the idea that, regardless of adaptivity, an individual's emotional experience is more positive and less negative on whichever test they find easier, likely due to higher levels of perceived control.

Early work on the emotional advantages of adaptive testing (e.g., Betz & Weiss, 1976) has claimed that adaptive testing would be beneficial for the emotional experience due to the reduction of boredom in high-ability individuals and the reduction of anxiety in low-ability individuals. Our findings do not support this assumption. They show that the level of anxiety was related to differences in the perceived difficulty of the two tests rather than their adaptability. For boredom, although the interaction was not statistically significant, the results descriptively show a similar pattern: The more one test was perceived as easier than the other, the less boredom was experienced on this test compared with the other test (for similar findings,

see Asseburg & Frey, 2013). However, it is worth noting that boredom was generally very low in the lab setting of the present study (see also Goetz et al., 2023). The low levels of boredom may have been due to the challenging nature of the task and the induction of ego-threat by introducing the test as a measure of intelligence, a highly desirable trait for any individual.

Taken together, our findings indicate that it is not individuals' objectively assessed ability that determines differences in their emotional experience in an adaptive versus a fixed-item test. Rather, their subjective perceptions of difficulty and how they compare between the tests may be driving these differences. This finding corroborates previous studies showing that only perceived, not actual performance is associated with the preference for a certain test (Betz & Weiss, 1976), and that perceived success mediates the effect of actual success on metacognitive experiences like satisfaction (Ortner et al., 2013; Tonidandel et al., 2002).

There was some correspondence between ability and relative perceived difficulty, as participants who found the CAT more difficult than the FIT had higher average ability. However, this difference in mean ability was small and not statistically significant. As such, there were participants with high ability who found the FIT more difficult, as well as some with low ability who found the CAT more difficult.

In sum, our findings on self-reported emotions underline the importance of considering individual perceptions of the test-taker to understand the effects of adaptive testing on emotional experience. Especially the subjective experience of control, indicated by which test is perceived as easier and to what extent, seems to determine how adaptive and fixed-item tests compare in terms of the emotions they trigger. This finding contradicts the widespread notion that adaptive testing is emotionally beneficial for all test-takers due to a better fit between ability and test difficulty (e.g., Betz & Weiss, 1976).

Limitations and Directions for Future Research

Several limitations need to be considered in interpreting the findings and can be used to inform directions for future research. Some features of the present study may limit its generalizability to real-life applications. First, the present work was conducted in a lab setting with a sample mainly consisting of university students. While the lab setting allowed us to control variables that impact data quality, it may limit the generalizability of the present findings, in particular to high-stakes testing situations. By informing participants that the task would measure a component of intelligence and that they would receive feedback, we aimed to increase the perceived value of the task and thereby intensify the emotional experience on both tests. Although this instruction may have made the procedure more similar to a real-life testing situation, the setting likely did not fully resemble a high-stakes situation with strongly adverse consequences in the case of failure (e.g., not getting access to a desired study program) or highly desirable consequences in case of success (e.g., getting a desired job offer). In such situations, perceptions of control, and especially a loss or gain of control, might have more profound effects on test-takers' emotional experiences. In particular, when stakes are very high and a test is way too difficult, anxiety might be the dominant emotion, and "switching off," as it was possible in the present context, might be rare.

Second, in the present study, participants were not informed about the adaptivity of the test. In real-life settings, test-takers might be provided with such information, which might lead to perceiving difficult items on an adaptive test not as a loss of control but rather as a sign of having performed well. Third, the study employed numerical reasoning tasks in both tests. Research is needed to replicate the current findings in other domains and establish whether the effects are generalizable across different areas. Fourth, the procedure included both an adaptive

and a fixed-item test. This design does not fully mirror typical testing situations, where individuals would rarely encounter both an adaptive and a non-adaptive test. The two tests were administered consecutively, with participants determining the length of the break in between. The break was included to minimize carry-over effects and reinforce the notion in participants that a new test would commence after the break. However, especially participants who experienced negative emotions may have used this break to regulate them, which may have altered their self-report of emotions.

The finding of higher physiological arousal in the adaptive test provides new evidence suggesting that this form of testing is more emotionally arousing for test-takers. The increased arousal was likely triggered by a better balance of task demands and ability that might generate a stimulating level of challenge, increased task engagement, and reduced "switching off." However, since we were the first to consider psychophysiological reactions to adaptive testing, replication of this effect and further exploration are needed. Specifically, future research could explore how the detected increases in SCR relate to test-takers' performance and well-being during and after testing.

By considering additional variables, future research could also enhance our understanding of how adaptive testing influences the emotional experience. The present research focused on six common achievement emotions (joy, pride, anger, anxiety, frustration, boredom). While the assessment of these emotions likely covered a significant portion of participants' emotional experiences, future research could explore how adaptive testing influences other achievement emotions. In addition to assessing a pre-defined set of emotions, this could also be done by including a free-text option to describe the emotional experience. Future research could also include an assessment of perceived control to explore whether control perceptions indeed mediate the effect of test features on physiological and emotional states.

Based on control-value theory, we argued that it would be too simple to assume main effects of adaptive testing on the subjective experience of achievement emotions, as emotional differences between adaptive and fixed-item testing were explained by interactions between individual characteristics and features of the two tests. The findings corroborated this claim by showing that individuals experienced more positive and less negative emotions the more a test appeared easier to them. Surprisingly, neither ability nor objective mean difficulty or test scores could fully explain whether participants would find a certain test easier or not. Research is needed to determine factors that generate differential perceptions of difficulty and could be used as leverage points to positively influence test-takers' emotional experience. Until these factors are identified, we concur with previous recommendations to increase the success rate in adaptive testing (Asseburg & Frey, 2013). Especially since precision in adaptive testing is still considerably high at success rates of 60% or 70% (Eggen & Verschoor, 2006), it seems sensible to create an adaptive test that allows for relatively low difficulty, thereby providing a more positive emotional experience.

Finally, the present findings represent estimates of within-person effects, which may not necessarily be transferable to a between-person level (Hunter et al., 2024). They also represent effects with a fixed slope, that is, aggregates of within-person effects. As such, the findings might not hold true for each individual. Future research could use random slopes modeling to investigate generalizability across individuals.

Conclusion

The rising popularity of (computerized) adaptive testing in high-stakes and large-scale assessments raises the question of how adaptive testing affects test-takers' performance and emotional well-being. To the best of our knowledge, the present investigation is the first that used psychophysiological measures to answer this question. The findings show that adaptive testing leads to higher levels of sympathetic arousal compared to fixed-item testing. At the same time, the results imply that neither test format bears a systematic risk of emotionally harming participants. Instead, our results indicate that test-takers' perceptions of difficulty determine how their emotional experience compares between the two types of tests: Whichever test was perceived as easier by the test-taker was accompanied by more positive and less negative emotions. These findings alleviate concerns regarding potential negative effects of adaptivity on test-takers' affective states, thereby encouraging the use of adaptive testing given their psychometric benefits.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author(s) used GPT-4 and Grammarly to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Akhtar, H., & Kovacs, K. (2024). Measurement precision and user experience with adaptive versus non-adaptive psychometric tests. *Personality and Individual Differences*, *225*, 112675. https://doi.org/10.1016/j.paid.2024.112675

Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2022). The effect of computerized adaptive testing on motivation and anxiety: A systematic review and meta-analysis. *Assessment 30*(5), 1379–1390. https://doi.org/10.1177/10731911221100995

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.

Betz, N. E., & Weiss, D. J. (1976). *Psychological effects of immediate knowledge of results and adaptive ability testing. Research Report 76-4*.

Boucsein, W. (2012). *Electrodermal Activity* (2nd ed.). Springer US. https://doi.org/10.1007/978-1-4614-1126-0

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E., & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2015). *A guide for analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for psychological experiments*. https://www.birmingham.ac.uk/documents/college-les/psych/saal/guide-electrodermal-activity.pdf

Christopoulos, G. I., Uy, M. A., & Yap, W. J. (2019). The body and the brain: Measuring skin conductance responses to understand the emotional experience. *Organizational Research Methods*, *22*(1), 394–420. https://doi.org/10.1177/1094428116681073

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, *30*(5), 379–393. https://doi.org/10.1177/0146621606288890

Frey, A., & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. In M. Prenzel, I. Gogolin, & H.-H. Krüger (Eds.), *Kompetenzdiagnostik* (pp. 169–184). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90865-6_10

Frey, A., Liu, T., Fink, A., & König, C. (2024). Meta-analysis of the effects of computerized adaptive testing on the motivation and emotion of examinees. *European Journal of Psychological Assessment, 40*(5), 1–17. https://doi.org/10.1027/1015-5759/a000821

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

Goetz, T., Preckel, F., Pekrun, R., & Hall, N. C. (2007). Emotional experiences during test taking: Does cognitive ability make a difference? *Learning and Individual Differences*,

17(1), 3–16. https://doi.org/10.1016/j.lindif.2006.12.002

Hunter, M. D., Fisher, Z. F., & Geier, C. F. (2024). What ergodicity means for you. *Developmental Cognitive Neuroscience*, *68*, 101406. https://doi.org/10.1016/j.dcn.2024.101406

Ikezawa, S., Corbera, S., Liu, J., & Wexler, B. E. (2012). Empathy in electrodermal responsive and nonresponsive patients with schizophrenia. *Schizophrenia Research*, *142*(1–3), 71–76. https://doi.org/10.1016/j.schres.2012.09.011

iMotions (Version 9.3). (2022). [Computer software].

Ishikawa, M. (2023). Measuring the autonomic nervous system as a window into the mind and brain: A selective review. *European Psychologist*, *28*(2), 67–82. https://doi.org/10.1027/1016-9040/a000500

Kiuru, N., Malmberg, L.-E., Eklund, K., Penttonen, M., Ahonen, T., & Hirvonen, R. (2022). How are learning experiences and task properties associated with adolescents' emotions and psychophysiological states? *Contemporary Educational Psychology*, *71*, 1–16. https://doi.org/10.1016/j.cedpsych.2022.102095

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*(3), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

Kreibig, S. D., & Gendolla, G. H. E. (2014). Autonomic nervous system measurement of emotion in education and achievement settings. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International Handbook of Emotions in Education* (pp. 625–642). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9780203148211.ch31

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Loe, B., Sun, L., Simonfy, F., & Doebler, P. (2018). Evaluating an automated number series item generator using linear logistic test models. *Journal of Intelligence*, *6*(2), 20. https://doi.org/10.3390/jintelligence6020020

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1–31. https://doi.org/10.18637/jss.v048.i08

Nourbakhsh, N., Chen, F., Wang, Y., & Calvo, R. A. (2017). Detecting users' cognitive load by galvanic skin response with affective interference. *ACM Transactions on Interactive Intelligent Systems*, *7*(3), 1–20. https://doi.org/10.1145/2960413

Obermeier, R., Schlesier, J., Meyer, S., & Gläser-Zikuda, M. (2022). Trajectories of scholastic well-being: The effect of achievement emotions and instructional quality in the first year of secondary school (fifth grade). *Social Psychology of Education*, *25*(5), 1051–1070. https://doi.org/10.1007/s11218-022-09726-2

OECD. (2022). *PISA 2022 Technical Report*. OECD Publishing. https://www.oecd.org/pisa/data/pisa2022technicalreport/PISA-2022-Technical-Report-

Ch-01-PISA-Overview.pdf

Ortner, T. M., Weißkopf, E., & Gerstenberg, F. X. R. (2013). Skilled but unaware of it: CAT undermines a test taker's metacognitive competence. *European Journal of Psychology of Education*, *28*(1), 37–51. https://doi.org/10.1007/s10212-011-0100-7

Ortner, T. M., Weißkopf, E., & Koch, T. (2014). I will probably fail: Higher ability students' motivational experiences during adaptive achievement testing. *European Journal of Psychological Assessment*, *30*(1), 48–56. https://doi.org/10.1027/1015-5759/a000168

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

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*, *93*(S1), 227–238. https://doi.org/10.1111/bjep.12575

Pekrun, R. (2024). Control-Value Theory: From Achievement Emotion to a General Theory of Human Emotions. *Educational Psychology Review*, *36*(3), 83. https://doi.org/10.1007/s10648-024-09909-7

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

Powell, Z.-H. E. (1994). The psychological impacts of computerized adaptive testing methods. *Educational Technology*, *34*(8), 41–47.

Roos, A.-L., Goetz, T., Krannich, M., Donker, M., Bieleke, M., Caltabiano, A., & Mainhard, T. (2023). Control, anxiety and test performance: Self-reported and physiological indicators of anxiety as mediators. *British Journal of Educational Psychology*, *93*(S1), 72–89. https://doi.org/10.1111/bjep.12536

Rosebrock, L. E., Hoxha, D., Norris, C., Cacioppo, J. T., & Gollan, J. K. (2017). Skin Conductance and Subjective Arousal in Anxiety, Depression, and Comorbidity: Implications for Affective Reactivity. *Journal of Psychophysiology*, *31*(4), 145–157. https://doi.org/10.1027/0269-8803/a000176

The Psychometrics Centre. (n.d.). *Concerto Adaptive Testing Platform*. University of Cambridge. Retrieved 2 April 2024, from https://www.psychometrics.cam.ac.uk/newconcerto

Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, *16*(1), 1–9. https://doi.org/10.7275/WQZT-9427

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*(2), 320–332. https://doi.org/10.1037/0021-9010.87.2.320

Vanneste, P., Raes, A., Morton, J., Bombeke, K., Van Acker, B. B., Larmuseau, C., Depaepe, F., & Van Den Noortgate, W. (2021). Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work*, *23*(3), 567–585. https://doi.org/10.1007/s10111-020-00641-0

Venables, P. H., & Mitchell, D. A. (1996). The effects of age, sex and time of testing on skin conductance activity. *Biological Psychology*, *43*(2), 87–101. https://doi.org/10.1016/0301-0511(96)05183-6

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, *37*(2), 70–84. https://doi.org/10.1080/07481756.2004.11909751

Weissman, D. G., & Mendes, W. B. (2021). Correlation of sympathetic and parasympathetic nervous system activity during rest and acute stress tasks. *International Journal of Psychophysiology*, *162*, 60–68. https://doi.org/10.1016/j.ijpsycho.2021.01.015

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, *2*(1), 1–17. https://doi.org/10.7333/1401-0201001

Zhang, J., Wang, K., & Zhang, Y. (2021). Physiological characterization of student engagement in the naturalistic classroom: A mixed-methods approach. *Mind, Brain, and Education*, *15*(4), 322–343. https://doi.org/10.1111/mbe.12300

Highlights

- Test-takers show stronger physiological arousal, indicated by skin conductance response, during an adaptive compared to a fixed-item test.
- Subjective emotional experiences do not differ systematically between an adaptive and a fixed-item test.
- Test-takers report more positive and less negative emotions on whichever test they perceived to be easier, independent of adaptivity.

Abstract

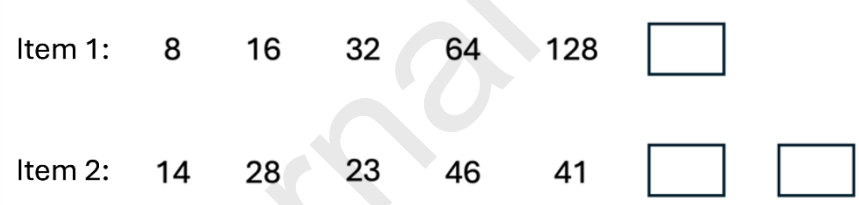
In light of the increasing use of computerized adaptive testing, we investigated how adaptive testing impacts test-takers' subjective emotional experiences and their psychophysiological arousal. Applying a within-person design ($N = 89$), we compared participants' affective states while working on an adaptive and a fixed-item test of numerical reasoning ability. During both

tests, we continuously recorded participants' skin conductance response. In addition, they filled in a self-report questionnaire after each of the three item blocks per test, assessing discrete achievement emotions (joy, pride, anger, boredom, frustration, and anxiety) and perceived level of task difficulty. As expected, participants showed higher levels of psychophysiological arousal in the adaptive compared to the fixed-item test, indicating that the adaptive test was more stimulating, independent of emotional valence. For subjective achievement emotions, we expected disordinal interaction effects between test type and ability (objective control experience) and between test type and relative perceived difficulty of the two tests (subjective control experience). This was supported for relative perceived difficulty, as participants indeed reported more joy and pride, and less frustration, anxiety, and anger on whichever test they subjectively perceived as easier. Meanwhile, no main effects of test type and no interaction between test type and ability were found. This is in line with the control-value theory and shows that it is not the adaptivity of a test that influences subjective emotional experience, but rather how difficult the adaptive test is perceived by test-takers compared to a fixed-item test. Directions for future research and implications for practice are discussed.

Keywords: emotions, achievement emotions, adaptive testing, computerized assessment, psychophysiological measures, galvanic skin response

Figure 1

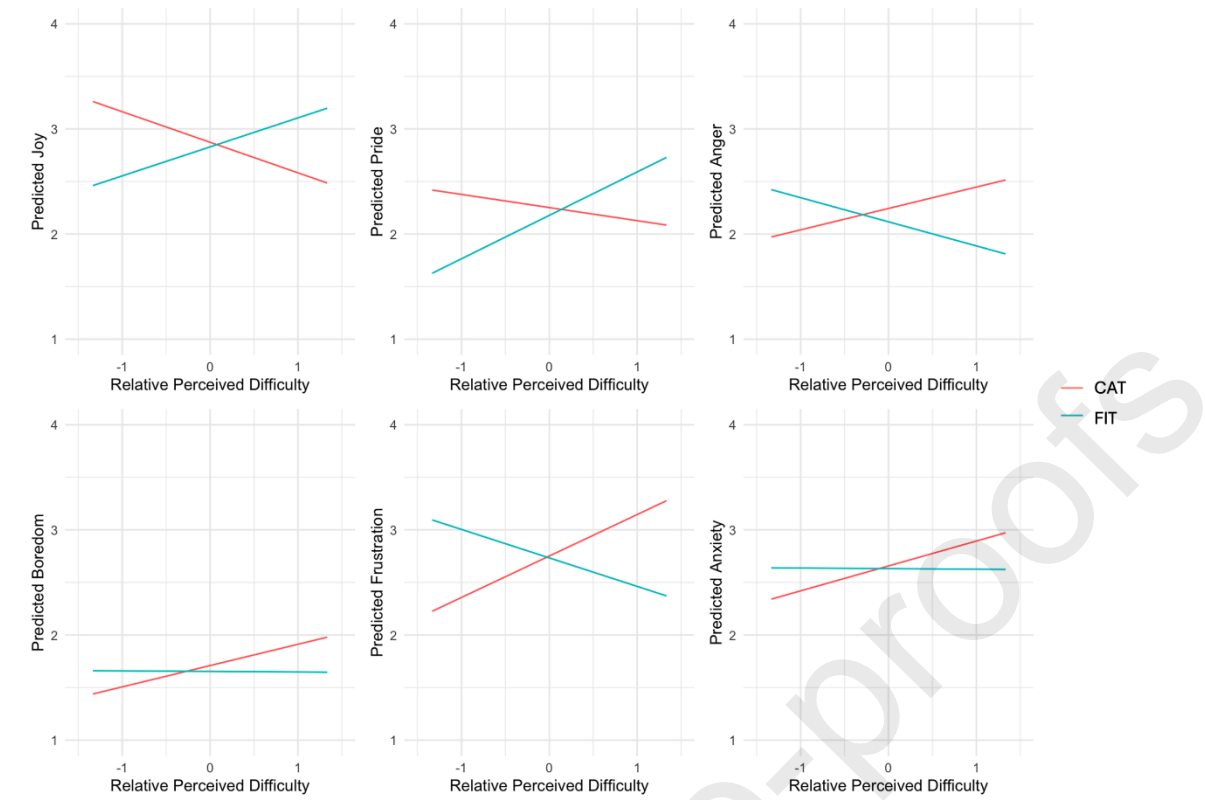
Exemplary Numerical Reasoning Items



Note. These items are examples for the type of items, not actual items used in the present study. Solutions are: "256" for Item 1, and "82; 77" for Item 2.

Figure 2

Interactions between Relative Perceived Difficulty and Test Type in Predicting Self-Reported Achievement Emotions



Note. Relative Perceived Difficulty < 0 indicates that the CAT was perceived as easier and > 0 that the FIT was perceived as easier. For boredom, the interaction effect was not significant.