Starting Tests With Easy Versus Difficult Tasks: Effects on Appraisals and Emotions

Maik Bieleke[1], Thomas Goetz[1], Maike Krannich[2], Anna-Lena Roos[3] and Takuya Yanagida[1]

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

[2] Institute of Education, University of Zurich, Zurich, Switzerland

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

This article has been accepted for publication in the *Journal of Experimental Education*, published by Taylor & Francis. The reference is:

Bieleke, M., Goetz, T., Krannich, M., Roos, A.-L., & Yanagida, T. (2021). Starting tests with easy versus difficult tasks: Effects on appraisals and emotions. *Journal of Experimental Education*. https://doi.org/10.1080/00220973.2021.1947764

Author Note

Maik Bieleke is now at the Department of Sport Science, University of Konstanz, Konstanz, Germany. This research was supported by a grant from the Committee on Research (AFF) at the University of Konstanz awarded to Thomas Goetz, Maike Krannich, and Anna-Lena Roos. Correspondence concerning this article should be addressed to Maik Bieleke, University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany. E-mail: maik.bieleke@uni-konstanz.de

Abstract

Tests in educational contexts often start with easy tasks assuming that this fosters positive experiences—a sense of control, higher valuing of the test, and more positive and less negative emotions. Although intuitive and widespread, this assumption lacks an empirical basis and a theoretical framework. We conducted a field experiment and randomly assigned 208 students to an easy-to-difficult or a difficult-to-easy condition in a mathematics test. Perceived challenge was measured along with control appraisals, value appraisals, and emotions (enjoyment, pride, anxiety, anger, boredom). While students starting with easy tasks felt less challenged than students starting with difficult tasks in Part 1, no differences emerged regarding control and value appraisals and emotions. In Part 2, students who had started with easy tasks proceeded to difficult tasks and reported higher challenge, lower value and control, and less positive and more negative emotions than students who proceeded from difficult to easy tasks. Control and value appraisals mediated these differences between conditions, especially regarding positive emotions. These results cast doubt on the preference for easy-to-difficult over difficult-to-easy task orders, revealing their potential for causing adverse experiences at the end of the test (e.g., reflecting contrast effects).

Keywords: achievement emotions; task difficulty; task order; control-value theory; perceived challenge; cognitive appraisals

Starting Tests With Easy Versus Difficult Tasks: Effects on Appraisals and Emotions

A widespread assumption among teachers and students alike is that tests should ideally start with easy tasks rather than with difficult ones (Skinner, 2009). It is assumed that presenting easy tasks at the beginning of a test motivates students and helps them build up confidence quickly, fostering positive and attenuating negative emotions throughout the test (e.g., by reducing test anxiety; Goetz & Kleine, 2006; Skinner, 2009). As plausible as this assumption seems, however, it is not grounded in empirical research. Also, the control-value theory of achievement emotions (CVT; Pekrun, 2006), a well-known theory that explains the emergence of emotions in the context of learning and achievement, suggests a more complex association between task order and emotions. According to the CVT, encountering easy tasks at the beginning of a test might indeed lead to more desirable appraisals and emotions. However, students will then inevitably encounter the difficult tasks at the end of the test, which should result in less desirable appraisals and emotions. These early desirable and later undesirable effects might cancel each other out, making the net effects on appraisals and emotions ambiguous. The resulting uncertainty regarding the link between task order and emotions is unfortunate when considering the ubiquity and prevalence of tests in academic settings (e.g., at schools; Baines & Goolsby-Smith, 2016), which makes it imperative to establish how test features like the ordering of tasks based on task difficulty can promote or compromise emotions. Examining such effects is crucial beyond the academic setting as well because tests are pervasive in various domains of life. For instance, people take tests that assess their general cognitive abilities (e.g., intelligence tests), determine their suitability for a

job (e.g., assessment centers), measure their athletic performance (e.g., in sport competition), or evaluate their artistic skills (e.g., in art or music contests). In all of these contexts, it is common practice to start with easy tasks, conforming with the general and widely shared belief that starting tests with the easier tasks allows people to “gain momentum” in terms of enhancing confidence and self-efficacy (Habbert & Schroeder, 2020).

Moreover, fostering positive and attenuating negative emotions in test situations is not merely an end in itself. Emotional and motivational aspects of tests might influence students' performance (e.g., Chan et al., 1997; Cole et al., 2008; Eklöf, 2010; Lau et al., 2009; Wise & DeMars, 2010). As such, erroneous assumptions about how the ordering of easy and difficult tasks influences these aspects can jeopardize the validity of test results as an indicator of students' ability. This in turn can have momentous consequences for students' academic and occupational prospects because tests are commonly used to grade students and to evaluate their academic performance. To address this important research gap, we designed a field experiment investigating how the ordering of easy and difficult tasks within a test affects the cognitive appraisals and emotions of students.

A Performance Measurement Perspective on the Order of Tasks

One of the fundamental decisions when designing tests is how the constituting tasks should be arranged (e.g., Lienert & Raatz, 1998; Miller et al., 2009). From the perspective of performance measurement, it is plausible to use task difficulty as a criterion for making this decision. Many tests are designed as "speeded power-tests," requiring students to work on tasks of varying difficulty during a limited time span. These tests include both easy and

difficult tasks to enable a representation of the full performance spectrum of the tested students (Lienert & Raatz, 1998; Sparfeldt, 2013). Under the assumption that students are able to solve all tasks with a difficulty level below (or identical to) their individual ability, presenting the tasks with ascending difficulty levels seems to be the optimal option. Therefore, when the goal is to measure students' maximal performance, starting with easy tasks is commonly recommended over starting with difficult tasks (Miller et al., 2009; Nagy et al., 2018; Sparfeldt, 2013).

This reasoning implies that ordering tasks according to their difficulty affects test performance; specifically, performance is expected to be worse when students start with difficult rather than with easy tasks. However, this assumption is generally not supported by empirical findings. A comprehensive review of the literature on the link between task order and performance (Hauck et al., 2017) identified a total of 12 comparisons of easy-to-difficult and difficult-to-easy orders. In two of these comparisons, an easy-to-difficult task order resulted in better performance than a difficult-to-easy task order, while in one comparison a difficult-to-easy task order resulted in better performance. However, the remaining nine comparisons yielded no significant differences between the two task orders. Based on these findings, the authors conclude that "the sequence of test items apparently has minimal effects on test performance" (p. 65), making the order of tasks seem inconsequential for performance. This is surprising if one assumes that an easy-to-difficult task order fosters positive emotions and attenuates negative emotions, as more pleasant emotions should be associated with better test performance (e.g., Pekrun et al., 2017). It should be noted,

however, that only about 50% of the non-significant results were based on samples sizes that granted sufficient power to detect medium effects of task order on performance, and none of them had sufficient power to detect small effects. The link between task order and performance is thus somewhat ambiguous. Still, the results available to date indicate that a difficult-to-easy task order does not necessarily impair performance (e.g., by preventing students from reaching the easier items later in the test).

Task Difficulty and Task Order: Associations With Emotions

Task Difficulty and Emotions

The control-value theory (CVT) of achievement emotions (Pekrun, 2006) proposes subjective control and value appraisals as important determinants of students' emotions in achievement situations. Subjective control refers to students' evaluation of their ability to have a causal impact on achievement activities and outcomes (e.g., the expectation that effort leads to success), whereas subjective value refers to the perceived valence of these activities and outcomes (e.g., intrinsic: the activity is interesting; extrinsic: the outcome is useful). Further, the CVT assumes that control and value mediate the effects of situational aspects like task difficulty on achievement emotions (Pekrun, 2000, 2006). Importantly, because control and value are subjective appraisals, they depend primarily on how students perceive aspects of the achievement situation (e.g., Frenzel et al., 2007b).

Based on the CVT, it can be hypothesized that the difficulty of the tasks that students encounter in a test situation influences their control appraisals. Here, we focus on situations that are characterized by a mismatch between perceived demands and student's abilities (in

contrast to situations in which demands and abilities match): In these situations, students perceive the demands associated with the task as either underchallenging (i.e., below their ability level) or overchallenging (i.e., above their ability level). We use the term challenge here to refer to a subjective assessment of the degree to which the demands of a task match one's ability level, which is tightly linked to, but not synonymous with, task difficulty. Specifically, easy tasks that are perceived as underchallenging should be associated with higher control than difficult tasks that are perceived as overchallenging. The difficulty of tasks might also affect value appraisals, in particular if there is a mismatch between the perceived demands and individual abilities (i.e., either under- or overchallenge; Pekrun, 2006). Empirically, the effects of difficult tasks and thus overchallenge have been associated with lower value (e.g., by undermining interest in the task; Fulmer & Tulis, 2013; Li et al., 2007). Underchallenging tasks can also be accompanied by low levels of task value, as the common experience of boredom in underchallenging situations suggests. To the extent that the degree of under- and overchallenge is similar (i.e., the discrepancies between demands and abilities are of comparable magnitude), the corresponding tasks should thus be associated with similarly low levels of value. However, in test situations at school it seems conceivable that students experience overchallenge to a larger degree than underchallenge (e.g., because even easy tasks must be solved thoroughly in a given time), which should result in lower levels of value in difficult compared to easy tasks. If this was the case, easy tasks that are perceived as underchallenging should be associated not only with higher control but also with higher value than difficult tasks that are perceived as overchallenging.

These differences in subjective control and value are assumed to be ultimately reflected in students' emotions—that is, the CVT conceives of control and value appraisals as mediators of the effects of task difficulty on students' emotions (Pekrun, 2006). Here, we focus on a set of two positive and three negative achievement emotions that are highly prevalent and important in academic settings (Pekrun et al., 2002; Pekrun et al., 2011): enjoyment, pride, anxiety, anger, and boredom. These emotions are conceptually distinct and cover three broader groups of emotions according to the circumplex model of emotions (Watson & Tellegen, 1985): positive activating (enjoyment and pride), negative activating (anxiety and anger), and negative deactivating (boredom). We did not assess positive deactivating emotions (e.g., relief and relaxation) because positive deactivating emotions tend to occur after rather than during academic situations (Pekrun et al., 2002).

Enjoyment and pride are experienced in successful situations, when students feel in control and value the task highly. Regarding pride, matters are complicated by an inverse link between difficulty and the attributability of success: On the one hand, pride is more likely in easy than in difficult tasks due to a higher probability of success. On the other hand, pride is undermined in easy versus difficult tasks by the attribution of success to task characteristics rather than to own abilities. As we have argued above, however, test situations in school are likely to be conducive to higher degrees of overchallenge than underchallenge. If this was the case, easy tasks should clearly facilitate enjoyment and pride in comparison to difficult tasks. Anxiety and anger, on the other hand, are experienced in (potentially) unsuccessful situations that are characterized by ambiguous or low levels of personal control, respectively, combined

with high value. Accordingly, experiences of failure and low control in tasks that are perceived as overchallenging should make it more likely to observe anxiety and anger in difficult than in easy tasks. However, to the extent that overchallenge might also be associated with lower value, these differences can be attenuated. Finally, boredom is likely to arise in situations of either low or high control combined with low value. Such a situation could arise during a test and we therefore included boredom although it has so far rarely been considered as a test emotion. If overchallenge is more pronounced than underchallenge, boredom could be more pronounced in difficult tasks that are perceived as overchallenging (e.g., Asseburg & Frey, 2013; Krannich et al., 2019) than in easy tasks that are perceived as underchallenging. Taken together, higher control and higher value in easy compared to difficult tasks should have a clearly facilitating effect on positive emotions. In contrast, higher control combined with higher value in easy compared to difficult tasks should attenuate negative emotions. However, these associations could be obscured by the complex association between negative emotions and control and value (e.g., all negative emotions except boredom intensify with increasing task value at a given level of control) and might accordingly be weaker than the associations with positive emotions. At any rate, control and value appraisals are expected to mediate differences between easy and difficult tasks in terms of emotions.

Task Order and Emotions

The common assumption that an easy-to-difficult task order fosters positive and attenuates negative emotions compared to a difficult-to-easy task order is mostly made with reference to differences in task difficulty at the beginning of a test. The central premise is that

students who encounter easy tasks at the beginning of a test gain higher confidence, become more motivated, and ultimately feel better than students who initially encounter difficult tasks (Goetz & Kleine, 2006; Miller et al., 2009; Skinner, 2009). For instance, the Assessment and Qualifications Alliance (AQA) is an English examination board that designs General Certificate of Secondary Education (GCSE) exams with increasing difficulty, assuming that this permits “confidence building” and makes the test an overall “rewarding experience” for students (AQA, 2016, p. 5). This is compatible with the CVT, as underchallenging tasks should lead to higher confidence as well as more positive and less negative emotions than underchallenging tasks. However, there is a second, more implicit and yet crucial premise: that the initially beneficial effects of starting with easy tasks carry over to the end of the test and thus facilitate students’ experience of the overall test. This is, however, questionable because students who start with easy tasks inevitably encounter the difficult tasks at the end of the test, while students who start with difficult tasks end up working on the easy tasks. As we have argued above, according to the CVT the perceived challenge associated with the difficulty of a task is crucial for control and value appraisals and thus for emotions. As the eventual change in task difficulty implied by the two task orders is likely accompanied by a reversal in perceived challenge, appraisals and emotions can be expected to reverse as well.

Specifically, students who benefit from easy tasks at the beginning of the test should make more adverse experience at the end of the test, compared to students who start with difficult and then proceed to easy tasks. We accordingly do not expect that one way of ordering tasks yields a generally better experience than the other. Instead, students should

perceive easier tasks as less challenging than difficult tasks irrespective of whether these are encountered at the beginning or at the end of the test. This implies a reversal of perceived challenge in the two task orders (see Figure 1), which should be accompanied by a reversal of control and value appraisals as well as of emotions. It follows that, overall, the common assumption that an easy-to-difficult task order is more beneficial than a difficult-to-easy task order in terms of control and value appraisals and emotions might be overly simplistic. Rather, with respect to predictions that can be derived from the CVT, it seems more sensible to assume that the difficulty of the tasks that students encounter in each part of the test should matter for their experience. If this was the case, tests should have similar effects on appraisals and emotions irrespective of whether they start with easy or difficult tasks.

The Present Research

We conducted a field experiment to investigate the common assumption that an easy-to-difficult task order yields more beneficial experiences in tests than a difficult-to-easy task order. This assumption rests on two empirically testable premises: First, easy tasks should foster more positive experiences than difficult tasks at the beginning of the test (e.g., by inducing feelings of control or reducing test anxiety; Goetz & Kleine, 2006; Skinner, 2009). Second, these initially beneficial effects of easy versus difficult tasks should still be present at the end of the test. The first but not the second premise is in line with the control-value theory (CVT) of achievement emotions. Based on the CVT, we instead assumed that easy tasks generally foster more positive experiences than difficult tasks, irrespective of whether these tasks are encountered at the beginning or at the end of test. And because students who start

with easy tasks inevitably proceed to difficult tasks, they should make worse experiences at the end of the test compared to students who start with difficult tasks and then proceed to easy tasks.

We tested these predictions by randomly assigning students to work on easy versus difficult tasks in Part 1 of a mathematics test. We then swapped task assignments, such that students who started with easy tasks now worked on difficult tasks in Part 2 (i.e., easy-to-difficult condition), while students who started with difficult tasks proceeded to easy tasks (i.e., difficult-to-easy condition). The common assumption that an easy-to-difficult task order yields more positive experiences (i.e., higher control and value, more positive and less negative emotions) than a difficult-to-easy task order is captured by a main effect of condition (Hypothesis 1). Our alternative hypothesis that the difficulty of the tasks students encounter in each part of the test matters for students' experience is instead captured by an interaction effect between condition and the part of the test (Hypothesis 2). As the common assumption about differences between easy-to-difficult and difficult-to-easy task order and our alternative hypothesis make the same predictions about students' experiences in Part 1 but not in Part 2 of the math test, we also directly compared the two conditions in Part 1 and in Part 2. Following up on Hypothesis 2, we assumed that students in the easy-to-difficult condition make better experiences in Part 1 (Hypothesis 3a) and worse experiences in Part 2 (Hypothesis 3b) compared to students in the difficult-to-easy condition. Finally, we tested the prediction made by the CVT that differences between the two conditions in terms of emotions are mediated by differences in control appraisals (Hypothesis 4a) and/or value appraisals

(Hypothesis 4b).

Methods

Sample

We collected data in four different upper-track schools (Gymnasium) in southern Germany. Nine 8th grade math classes with $N = 208$ students participated in the experiment and worked on a math test comprising easy and difficult tasks. Participants were randomly assigned to the easy versus the difficult tasks in Part 1 of the test and this assignment was swapped in Part 2, establishing an easy-to-difficult ($n = 103$) and a difficult-to-easy condition ($n = 105$). Testing our hypotheses involved comparisons between these conditions, and we therefore chose the sample size to be able to detect small-to-medium differences ($d \approx 0.4$) in independent t-tests with 80% power (two-tailed, $\alpha = .05$). All but two classes were taught by different teachers. Demographic data are not available for students from one class ($n = 23$) that did not participate in an initial screening session. The remaining 185 students were on average $M = 13.7$ years old ($SD = 0.4$) and comprised 97 females (54.5%), 81 males (45.5%); 7 students did not indicate their gender (3.8%). Written informed consent was obtained from students and their parents prior to the study.

Procedure

Screening Session

One to three weeks before the experiment, the students participated in a screening session during a regular math class. Demographic data were collected in this session that could be linked to the experimental data via pseudonymized codes.

Experimental Session

In the experimental session, all students worked on a math test comprising one part with easy tasks and one part with difficult tasks. Students were randomly assigned to the easy-to-difficult or the difficult-to-easy condition that the beginning of the session. They had 20 minutes to work on each of the two parts (i.e., 40 minutes for the entire test). Before each part, we assessed subjective control and value appraisals along with achievement emotions to obtain baseline measures. These baseline measures are not suitable for testing our hypotheses; rather, we used them to check that students in the two conditions did not already differ in their appraisals and emotions prior to working on the tasks. Immediately after they finished working on each part of the test, we measured students' appraisals and emotions during this part with retrospective measures, and used these assessments to test our hypotheses. We additionally assessed the perceived challenge associated with the tasks.

The math tasks used in this study served as an experimental manipulation rather than as a reliable measure of student's performance. We selected them from the database of a nationwide written mathematics test taken by students in the 8th grade as a standardized performance comparison (Institute for Educational Quality Improvement [IQB], 2019). The tasks covered four different content areas (i.e., numbers, measurement, space and form, functional relationships). The IQB classifies the tasks as easy or difficult based on their psychometric properties (e.g., solution frequencies) in representative studies with large student populations. Relying on these tasks allowed us to create an authentic, ecologically valid, and comparable test situation that—unlike an actual exam—enabled us to

experimentally vary the difficulty of the tasks in full accordance with ethical considerations. A mathematics teacher was consulted to select easy and difficult tasks in line with the regular curricula of the targeted schools. This resulted in a pool of 23 tasks for the easy part and 10 tasks for the difficult part of the test. Students had 20 minutes to work on each part; however, this limit was deliberately chosen to make sure that students would not finish all of the tasks in time to ensure a realistic speeded-test assessment. We made students aware of this at the beginning of the test and emphasized that they would not be able to solve all problems within the provided time. To give relevance to the test and encourage students to perform well, it was announced as a preparatory exam for the upcoming state-wide comparison tests. Additionally, we awarded a prize of 250 euros to the class with the best overall test performance.

Measures

The present study was conducted as part of a broader investigation, and for the sake of conciseness we focus on those tasks and measures that were collected to address our present research question. Because survey time was limited, we used single-item measures for all constructs of interest. Despite their brevity, single-item measures have evidence of reliability and validity as instruments for assessing motivational and emotional constructs (e.g., Gogol et al., 2014; Wanous et al., 1997). In all retrospective questions, we highlighted that the items related to a specific part of the test rather than to the test situation as a whole by explicitly linking them to students' "experiences while you were working on the math tasks in the first [vs. second] part of the test."

Perceived Challenge

After they had completed each part of the test, we asked students as how challenging they perceived the corresponding tasks with a single 5-point bipolar item ("The requirements of the math tasks were …") comprising five response alternatives: 1 (*much too low*), 2 (*too low*), 3 (*just right*), 4 (*too high*), and 5 (*much too high*). Accordingly, scores of 3 represent optimal challenge, whereas scores smaller than 3 represent perceived underchallenge and scores greater than 3 represent perceived overchallenge.

Subjective Control and Value Appraisals

We measured control and value with one item each. Students indicated their control and value appraisals *before* working on each part of the math test with a concurrent assessment ("I feel I have the situation under control" [see Perry et al., 2001; Weinstein et al., 2002, for a similar assessment] and "In this math task it is important to me to achieve a good result" [see Bieg et al., 2013; Frenzel et al., 2007a, for a similar assessment]). These baselines measures were used to rule out differences between conditions prior to the math tasks. To test our hypotheses, we assessed students' experiences *during* the math tasks with a retrospective assessment ("I felt I had the situation under control" and "In this math task it was important to me to achieve a good result", respectively). The item pertaining to value focused on extrinsic value (i.e., the value of achievement) because extrinsic value is likely more susceptible to experiences made during a single test than intrinsic value, which might capture stable interest in a domain. Answers were provided on 5-point Likert scales (1 = *not at all true*, 5 = *exactly true*) for all of these measures.

Achievement Emotions

We measured the five achievement emotions of enjoyment, pride, anxiety, anger, and boredom with single items adapted from the Achievement Emotions Questionnaire (AEQ; Pekrun et al., 2011). Students indicated the extent to which they experienced these emotions *before* working on the math tasks with a concurrent assessment ("How strongly do you currently experience the following emotions?"). Again, these baseline measures were meant to rule out differences between conditions prior to the math tasks. More importantly, we measured emotions *during* the math tasks with a retrospective assessment ("How strongly did you experience the following emotions while completing the math tasks?"). Answers were provided on Likert scales (1 = *not at all*, 5 = *very much*) for all of these measures.

Analytic Approach

The analyses were conducted using the statistical software *R* version 4.0.2 (R Core Team, 2020) and *Mplus* version 8.1 (Muthén & Muthén, 1998-2018). We conducted mixed ANOVA with *afex* version 0.27-20 (Singmann et al., 2020). All tests were two-sided and the Type-I error rate was set to $\alpha = .05$. A total of 1.03% of the data was missing, ranging from 0.00% to 5.29% on the level of individual variables. Little's test was not significant, $\chi^2(548) = 582.5$, $p = .149$, supporting the assumption that values are missing completely at random (MCAR). We used listwise deletion of missing data in the ANOVA analyses and full information maximum likelihood (FIML) in the mediation analyses.

We subjected perceived challenge, control and value appraisals, and achievement emotions (enjoyment, pride, anxiety, anger, and boredom) to mixed ANOVA and determined

the main effects of Condition (easy-to-difficult, difficult-to-easy), the main effects of Part (1, 2), as well as the interaction effects of Condition and Part. We then inspected simple effects of Condition using *emmeans* version 1.4.8 (Lenth, 2020). This approach allowed us to test the following set of hypotheses. First, we examined the common assertion that students assigned to an easy-to-difficult order make more positive experiences across the two parts of the math test than students assigned to a difficult-to-easy order (i.e., main effect of Condition; Hypothesis 1). Second, we examined our alternative prediction that differences between the easy-to-difficult and the difficult-to-easy task order reverse from Part 1 to Part 2 of the math test (i.e., interaction effect of Condition and Part; Hypothesis 2). Third, we examined whether students in the easy-to-difficult condition make more positive experiences in Part 1 and more negative experiences in Part 2 compared to students in the difficult-to-easy condition (i.e., simple effects of Condition; Hypothesis 3a and Hypothesis 3b). Fourth, we investigated whether control and value appraisals mediate differences between the easy-to-difficult and the difficulty-easy conditions with respect to emotions. To this end, we estimated mediation models and report direct and indirect effects. Of particular interest were the specific indirect effects of control (Hypothesis 4a) and value (Hypothesis 4b). To ascertain the significance of these effects, we relied on bias-corrected bootstrap confidence intervals based on 10,000 samples (Muthén & Muthén, 1998-2018). We considered effects as significant when their confidence intervals excluded zero.

Results

An overview of the descriptive statistics (means, standard deviations) of all variables

assessed in this study is provided in Table 1.

Perceived Challenge

The analysis of perceived challenge revealed an interaction effect of Condition and Part, $F(1, 198.28) = 163.00$, $p < .001$, in line with Hypothesis 2 (see Figure 2). Simple effects showed that participants in the easy-to-difficult condition perceived Part 1 as less challenging ($M = 3.02$, $SD = 0.65$) than participants in the difficult-to-easy condition ($M = 3.57$, $SD = 0.70$), $t(374) = 5.36$, $p < .001$, $d = 0.88$. In Part 2, participants in the easy-to-difficult condition reported to be more challenged ($M = 3.76$, $SD = 0.69$) than participants in the difficult-to-easy condition ($M = 2.74$, $SD = 0.76$), $t(372) = 10.35$, $p < .001$, $d = 1.69$. This pattern of results is in line with Hypothesis 3a and Hypothesis 3b. Besides the interaction effect, we also found a main effect of Order, $F(1, 202.66) = 9.77$, $p = .002$. However, despite its significance this effect does not support Hypothesis 1, as it reflects that participants in the easy-to-difficult condition reported an overall *higher* level of challenge ($M = 3.40$, $SD = 0.77$) across both parts of the math test than participants in the difficult-to-easy condition ($M = 3.16$, $SD = 0.84$). This unexpected finding reflects an asymmetry in the magnitudes of the difference between conditions in Part 1 ($d = 0.88$, 95% CI [0.55, 1.21]) versus Part 2 ($d = 1.69$, 95% CI [1.35, 2.04]), which are significantly different according to their non-overlapping confidence intervals. As a consequence, the lower perceived challenge that students in the easy-to-difficult condition reported in Part 1 was surpassed in magnitude by the higher perceived challenge that these students reported in Part 2, leading to an overall higher level of perceived challenge compared to students in the difficult-to-easy condition. Finally, the main effect of Part was not

significant, $F(1, 198.28) = 0.50$, $p = .479$, suggesting that perceived challenge did not merely change over time.

Did Task Order Affect Control and Value Appraisals?

Appraisals Before the Tasks

At baseline, we found main effects of Part with respect to control, $F(1, 205.34) = 96.58$, $p < .001$, but not with respect to value, $F(1, 202.59) = 2.13$, $p = .146$, reflecting that students reported more control before Part 1 ($M = 3.96$, $SD = 0.86$) than before Part 2 ($M = 2.95$, $SD = 1.35$) across conditions. Besides these general changes in baseline control appraisals over time, the main effects of Condition and the interaction effects of Condition and Part were not significant, $p \geq .146$. This indicates that students in the easy-to-difficult and the difficult-to-easy condition did not differ in their cognitive appraisals prior to Part 1 or Part 2 of the math test.

Appraisals During the Tasks

We found significant interaction effects of Condition and Part with respect to control, $F(1, 203.72) = 12.32$, $p < .001$, and value, $F(1, 203.99) = 20.45$, $p < .001$, in line with Hypothesis 2 (see Figure 3). Contrary to Hypothesis 1, the main effects of Condition were not significant for control, $F(1, 204.61) = 1.35$, $p = .247$, or value, $F(1, 205.88) = 1.46$, $p = .228$. Similarly, the main effects of Part were not significant for control, $F(1, 203.72) = 3.83$, $p = .052$, or value, $F(1, 203.99) = 2.74$, $p = .100$. The simple effects of Condition showed that students reported similar levels of control and value in Part 1 of the math test across conditions, $p \geq .249$. Compared to students in the easy-to-difficult condition, however, students in the difficult-to-easy condition

reported more control (M = 3.59, SD = 1.10 versus M = 3.11, SD = 1.16), $t(375) = 3.02$, $p = .003$, $d = 0.50$, and higher value (M = 3.07, SD = 1.28 versus M = 2.52, SD = 1.29), $t(294) = 3.04$, $p = .003$, $d = 0.70$, in Part 2 of the math test. Taken together, these results are not in line with Hypothesis 3a, whereas they are consistent with Hypothesis 3b with regard to both control and value appraisals.

Did Task Order Affect Achievement Emotions?

Emotions Before the Tasks

At baseline, we found main effects of Part with respect to enjoyment, $F(1, 205.00) = 144.19$, $p < .001$, and pride, $F(1, 203.06) = 54.08$, $p < .001$, reflecting that students reported more enjoyment before Part 1 (M = 2.80, SD = 1.07) than before Part 2 (M = 1.89, SD = 0.96) and more pride before Part 1 (M = 2.00, SD = 0.97) than before Part 2(M = 1.50, SD = 0.80). We also found main effects of Part with respect to anxiety, $F(1, 203.06) = 27.11$, $p < .001$, and boredom, $F(1, 203.51) = 7.16$, $p = .008$, but not with respect to anger, $F(1, 204.38) = 1.12$, $p = .292$. This reflects that students reported more anxiety before Part 1 (M = 1.91, SD = 1.05) than before Part 2 (M = 1.56, SD = 0.88) and more boredom before Part 1 (M = 2.02, SD = 1.06) than before Part 2 (M = 1.80, SD = 1.10). Besides these general changes in baseline emotions over time, the main effects of Condition and the interaction effects of Condition and Part were not significant, $p \geq .141$. This indicates that students in the two conditions did not differ in their emotions prior to Part 1 or Part 2 of the math test.

Emotions During the Tasks

We found significant interaction effects of Condition and Part with respect to

enjoyment, $F(1, 206.00) = 7.77$, $p = .006$, pride, $F(1, 204.72) = 4.41$, $p = .037$, and anger, $F(1, 205.18) = 7.68$, $p = .006$, in line with Hypothesis 2. Contrary to Hypothesis 2, however, no such interaction effects emerged for anxiety, $F(1, 205.46) = 1.59$, $p = .209$, and boredom, $F(1, 204.93) = 2.19$, $p = .141$. Contrary to Hypothesis 1, the main effects of Condition were not significant for enjoyment, $F(1, 206.00) = 3.03$, $p = .083$, pride, $F(1, 205.61) = 3.30$, $p = .071$, anxiety, $F(1, 205.94) = 0.56$, $p = .454$, anger, $F(1, 205.73) = 0.32$, $p = .571$, and boredom, $F(1, 205.88) = 2.27$, $p = .134$. Finally, the main effects of Part were non-significant for enjoyment, $F(1, 206.00) = 0.18$, $p = .675$, and pride, $F(1, 204.72) = 0.18$, $p = .668$, but significant for anxiety, $F(1, 205.46) = 19.88$, $p < .001$, anger, $F(1, 205.18) = 26.97$, $p < .001$, and boredom, $F(1, 204.93) = 22.15$, $p < .001$.

The pattern of results underlying these analyses is depicted in Figure 4. It shows the expected reversal between conditions in the domain of positive emotions (enjoyment, pride), whereas for negative emotions (anxiety, anger, boredom) there were similar and general developments from Part 1 to Part 2 across conditions. However, for anger and boredom the magnitude of these developments differed between conditions. The simple effects of Condition showed that students reported similar levels of enjoyment, pride, anxiety, anger, and boredom in Part 1 of the math test, $p \geq .224$. Compared to students in the easy-to-difficult condition, however, students in the difficult-to-easy condition reported more enjoyment ($M = 2.16$, $SD = 1.23$ versus $M = 1.72$, $SD = 0.92$), $t(379) = 3.06$, $p = .002$, $d = 0.51$, more pride ($M = 1.82$, $SD = 1.06$ versus $M = 1.49$, $SD = 0.71$), $t(377) = 2.71$, $p = .007$, $d = 0.45$, less anger ($M = 1.64$, $SD = 0.99$ versus $M = 1.97$, $SD = 1.16$), $t(382) = 2.12$, $p = .034$, $d = 0.35$, and less boredom ($M = 1.83$, $SD = 1.10$ versus $M = 2.15$, $SD = 1.38$), $t(336) = 2.06$, $p = .040$, $d = 0.39$, in Part 2 of the math test.

Only with regard to anxiety no differences emerged in Part 2, $t(334) = 1.29$, $p = .198$, $d = 0.25$. Taken together, this pattern of results is not in line with Hypothesis 3a but it is consistent with Hypothesis 3b for all emotions except anxiety.

Did Control and Value Mediate Effects of Task Order on Achievement Emotions?

In the analyses reported so far, we found differences between conditions with regard to control and value appraisals as well as achievement emotions in Part 2 but not in Part 1 of the math test. Accordingly, we focused our mediation hypotheses on Part 2 to examine whether the observed differences between conditions in terms of emotions were mediated by changes in control (Hypothesis 4a) and/or value (Hypothesis 4b). An overview of the mediation models and the estimated path coefficients is provided in Figure 5.

Regarding enjoyment, the total indirect effect was significant, $b = 0.22$, 95% CI [0.08, 0.36], whereas the direct effect was not significant, $b = 0.22$, 95% CI [−0.05, 0.50]. The specific indirect effects of task order on enjoyment via control, $b = 0.10$, 95% CI [0.01, 0.19], and value, $b = 0.12$, 95% CI [0.03, 0.22], were both significant. Regarding pride, the total indirect effect was significant, $b = 0.20$, 95% CI [0.08, 0.32], while the direct effect was not significant, $b = 0.14$, 95% CI [−0.08, 0.36]. The specific indirect effects via control and value were both significant, $b = 0.07$, 95% CI [0.001, 0.13], and $b = 0.13$, 95% CI [0.04, 0.23], respectively. Together, these findings indicate that higher perceptions of value and control mediated the differences between conditions regarding enjoyment and pride in Part 2 of the math test. This is in line with both Hypothesis 4a and Hypothesis 4b in the domain of positive emotions.

With regard to negative emotions, we found a significant specific indirect effect of

control on anxiety, $b = -0.10$, 95% CI [−0.18, −0.02], whereas the specific indirect effect of value was not significant, $b = 0.05$, 95% CI [−0.01, 0.12]. Also, neither the total indirect effect, $b = -0.04$, 95% CI [−0.14, 0.05], nor the direct effect were significant, $b = -0.13$, 95% CI [−0.35, 0.09]. Regarding anger, neither the direct effect, $b = -0.25$, 95% CI [−0.52, 0.01], nor the total indirect effect were significant, $b = -0.08$, 95% CI [−0.22, 0.07]. The latter, however, reflects the presence of two opposing specific indirect effects: a negative effect via control, $b = -0.16$, 95% CI [−0.29, −0.04], and a positive effect via value, $b = 0.09$, 95% CI [0.01, 0.17]. With regard to boredom, we found a significant total indirect effect, $b = -0.15$, 95% CI [−0.27, −0.03] but no direct effect, $b = -0.17$, 95% CI [−0.48, 0.15]. Also, neither the specific indirect effect via control, $b = -0.06$, 95% CI [−0.15, 0.02], nor via value were significant, $b = -0.09$, 95% CI [−0.18, 0.004]. Taken together, these findings indicate that control and value mediated the differences between conditions regarding anger in Part 2, in line with Hypothesis 4a and Hypothesis 4b. Additionally, control but not value emerged as mediators for anxiety, but this result should be taken with a grain of salt as no difference between conditions emerged in the first place. Finally, the results found with respect to boredom are not in line with Hypothesis 4a and Hypothesis 4b.

Discussion

We conducted a field experiment to investigate the common assumption that starting tests with easy tasks (i.e., easy-to-difficult condition) rather than with difficult tasks (i.e., difficult-to-easy condition) fosters positive experiences in terms of control appraisals, value appraisals, and emotions. The results of our experiment contradict this assumption in several

ways. First, we found no evidence for the predicted overall difference between conditions in terms of cognitive appraisals or emotions. The only general effect was on perceived challenge, but this reflected that students felt *more* challenged in the easy-to-difficult than in the difficult-to-easy condition. Second, while students working on easy tasks in Part 1 felt less challenged than those working on difficult tasks, this difference was not accompanied by differences in terms of control and value appraisals or emotions. Third, when students who had started with the easy tasks proceeded to the difficult tasks in Part 2, they felt more challenged than students who worked on the easy tasks after having completed the difficult tasks. This difference in perceived challenge was accompanied by lower control, lower value, more negative emotions (anger, boredom), and less positive emotions (enjoyment, pride). That is, not only had starting with easy tasks no beneficial effect on students' experience in general or at the beginning of the test, it also backfired at the end.

The pattern of results was more consistent with predictions derived from the control-value theory of achievement emotions (CVT; Pekrun, 2006). Based on the CVT, we assumed higher control and value as well as more favorable emotions when students work on easy rather than on difficult tasks, irrespective of whether these tasks are encountered at the beginning of the test or at the end. Moreover, we expected differences in emotions to be mediated by control and value appraisals. These predictions were mostly confirmed in Part 2 of the test, where students working on easy tasks indeed reported more pleasant experiences than students working on difficult tasks. Moreover, differences between conditions were mediated by both control and value for the positive emotions of enjoyment and pride and for

the negative emotion of anger. Control further mediated condition differences in anxiety, while no mediation was observed for boredom.

Differences in Perceived Challenge Varied in Magnitude Between Part 1 and 2

In contrast to Part 2, the lack of differences between conditions in Part 1 was inconsistent with common assumptions about task order as well as with predictions derived from the CVT. This is particularly surprising because students already differed in their perceived challenge in Part 1. Notably, however, this difference was significantly smaller in magnitude than in Part 2 (i.e., about half the effect size). Accordingly, the difference in perceived challenge might not have been large enough to alter control and value appraisals already in Part 1. Feelings of control are not only determined by situational characteristics but also shaped by stable appraisals, such as the academic self-concept or self-efficacy beliefs (e.g., Goetz et al., 2010). A similar argument can be made with respect to value, which captures rather stable interests in a subject domain or achievement goals (e.g., Frenzel et al., 2007a). This assumed stability of appraisals is well in line with established social-cognitive learning theories (Rotter, 1954). Accordingly, it is conceivable that rather strong differences in perceived challenge like those observed in Part 2 of the test are required to affect control and value appraisals. Unfortunately, this might thwart attempts by teachers to make test situations more pleasant by starting with easy tasks.

While this reasoning might explain the absence of differences in appraisals and emotions in Part 1, it does not explain why the magnitude of differences in perceived challenge was smaller than in Part 2 in the first place. A promising starting point for

addressing this observation is provided by the literature on contrast effects: People incorporate available information when evaluating a situation (Schwarz & Bless, 1992) and this information arguably varied as a function of task order in our study. Tasks that were encountered at the beginning of a test could not be evaluated based on information about previous tasks. Therefore, differences in how challenging students perceived these tasks have probably reflected differences in task difficulty. The tasks students encountered at the end of the test, however, could be evaluated with respect to information about previous tasks, and this might explain the greater magnitude of differences in perceived challenge. Specifically, having started with easy tasks might have made subsequently encountered difficult tasks appear more challenging than they actually were, while having started with difficult tasks might have made subsequently encountered easy tasks appear less challenging. This reasoning is well in line with our data (see Figure 2) and contrast effects like this are frequently observed in various achievement settings (e.g., as determinants of the academic self-concept; Marsh, 1987). We therefore think that contrast effects might have affected the results of our study.

Effects on Achievement Emotions

Apart from the absence of differences between conditions in Part 1, our study provides experimental evidence for key assertions regarding the antecedents of achievement emotions made by the CVT and some additional novel insights. In particular, we observed notable differences between specific emotions. For the positive emotions of enjoyment and pride, we found the predicted reversal between the two task orders from Part 1 to Part 2. In the domain of negative emotions, however, we found decreasing anger and anxiety as well as increasing

boredom across task orders. Rather than being reversed from Part 1 to Part 2, anger and boredom were less pronounced when students encountered the easy tasks last (i.e., in the difficult-to-easy condition), which nevertheless led to the predicted differences between conditions in Part 2. Moreover, while differences between task orders were mediated by control and value with regard to both positive emotions, these effects were less consistent in the domain of negative emotions. These findings converge with research showing that the amount of mediation as well as the relative importance of control and value both differ between achievement emotions (Goetz et al., 2020). They are also in line with research showing that positive emotions can be more responsive to aspects of the learning environment than negative emotions (Goetz et al., 2021). Finally, they reflect that increases in both control and value have unequivocal effects on enjoyment and pride. For negative emotions, however, a change of control and value in the same direction can have more ambiguous effects. That is, negative emotions might depend in a more complex way on the interplay of control and value than positive emotions.

To illustrate, consider our findings with respect to anxiety as the only emotion for which we did not find a difference between conditions. According to the CVT, anxiety arises when there is uncertainty about the outcome (i.e., ambiguous control) in a highly valued situation (Pekrun, 2006). These conditions were neither met during easy tasks—in which control turned out to be rather high—nor in difficult tasks—in which value turned out to be rather low. This might not only explain the lack of differences between conditions but also the generally low levels of anxiety we found especially in Part 2. The fact that the stakes of performing poorly in our test were limited for practical and ethical reasons (e.g., no grading was

involved and individual contributions to the class price were averaged) might have further contributed to this result. In summary, it seems worthwhile to continue research that investigates the interplay between control and value appraisals within distinct emotions, perhaps in other academic contexts than tests (e.g., during learning or in class).

Limitations and Future Directions

Our study has limitations that should be considered when evaluating the results and their practical implications. A first caveat is that our experimental design made sure that students always worked on the easy and difficult tasks for a given time. However, tests should not start with difficult tasks when doing so bears the risk that students do not reach the easy questions because they spend too much time on the difficult ones (Flaugher et al., 1968; Miller et al., 2009; Towle & Merrill, 1975). Although more positive and less negative emotions are always desirable, optimizing tests with regard to these experience at the expense of measurement accuracy seems undue. Thus, sufficient time should be allotted for working on all tasks or the purpose of the test should be on formative feedback rather than on summative feedback and grading. Relatedly, we examined emotions in a speeded-power test that does not represent the only way in which tests are administered in practice (e.g., when the focus is on creating test environments that allow students to demonstrate their best performance, as in the framework of the Universal Design of Learning; Ketterlin-Geller, 2005). Accordingly, it would be interesting to examine the extent to which our findings hold in tests without time limits. It is conceivable, for instance, that lifting or relaxing time limits increases the sense of control students have concerning their performance, which might counteract the adverse effects task difficulty has on perceived control.

Second, our study is based on self-reports of cognitive appraisals and emotions, which might be problematic (e.g., biased perception) and should therefore be complemented by additional data sources in future research. For instance, the behavioral and physiological components of emotions could be assessed by observational and physiological data, respectively, to obtain a more comprehensive measure. Also, we limited ourselves to investigating a subset of the various achievement emotions that are likely relevant in test situations. It would be advisable to investigate other emotions as well. For instance, we focused on pride but not on shame, which is probably another highly prevalent emotion occurring during tests. Another route for future research could involve the effects of task order on positive deactivating emotions (e.g., relief) that we omitted in our study because they tend to occur after rather than during the test situation.

Third, our sample consisted solely of 8th graders that were recruited from math classes at upper-track schools in Germany and who performed an incentivized but still low-stake test. It thus remains to be explored whether our findings generalize to different age groups (e.g., at the secondary level), school types (e.g., elementary school, university), and test types (e.g., high-stake test). And although the basic structures and functional mechanisms of emotions are assumed to be universal (Pekrun, 2006; Goetz et al., 2021), it seems also worthwhile to investigate whether our findings replicate across subject domains and contexts (e.g., graded exams).

Fourth, we did not examine the extent to which differences between conditions in terms of experiences (appraisals, emotions) also map onto differences in performance. The

relationship between task order and performance has already been investigated (Hauck et al., 2017) and our focus therefore was on emotions. Accordingly, the math test we used was designed as an experimental manipulation rather than as a reliable measure of students' performance. Nevertheless, follow-up studies could focus on emotions in test situations that are more suitable for assessing performance.

Fifth, we focused on a direct comparison of easy-to-difficult and difficult-to-easy task order, which connects our study directly to previous research (e.g., Hauck et al., 2017) and allowed us to test common conceptions about task ordering effects on emotions. However, other difficulty levels are conceivable and used in teaching; for instance, it would also be interesting to investigate whether tasks of medium difficulty could mitigate the negative effects of encountering difficult tasks at the end of the test (e.g., an easy-medium-difficult task order). Moreover, it remains unclear whether dispersing problems of varying difficulty across the test (e.g., easy-to-difficult-to-easy-to-difficult) might yield different results. Relatedly, a different distribution of tasks might also evoke context effects beyond the contrast effect we discussed above. For instance, the difficulty of the last task students encounter during a test might influence their overall evaluation (i.e., a recency effect) and thus shape their appraisals and emotions with regard to the whole test. In a similar vein, while separating the easy and difficult tasks into two distinct parts provided ideal grounds for testing our hypotheses, it might have created the impression of two separate tests rather than a single test. Despite several provisions to avoid such an impression (e.g., announcing the test upfront as a single two-part test, awarding a prize for the best overall performance), it might explain the lack of

differences in baseline ratings between Part 1 and 2 as well as the presence and size of the contrast effect discussed earlier. Future research might thus check the generalizability of our results to more integrated test situations.

Finally, it is worth noting that students in our study perceived the easy tasks as less challenging than the difficult tasks but not as underchallenging. One might argue that the tasks we used therefore do not permit the most rigorous experimental test of our hypotheses. However, the easy tasks were the easiest ones available for 8th grade math classes; creating even easier tasks would have had required to either construct artificial tasks or use tasks from lower grade levels. We decided against this approach to maintain the ecological validity of our study and ensure the practical relevance of our findings. Future research might employ a more rigorous experimental test of our hypotheses to examine the robustness of our results.

Practical Implications

The most straightforward practical implication of our research is that starting tests with easy tasks does not have the desired advantages in terms of better experience, and that instructors should consider that an easy-to-difficult task order can backfire at the end of the test. It thus seems advisable for test constructors to avoid tests with a strict distinction in easy-to-difficult tasks (which is a common practice; e.g., Miller et al., 2009). Of course, it might not be avoidable that students work through the tasks in a particular order, especially when the tasks are distributed in a single booklet or when they build upon each other in terms of their content. Yet, even then it seems likely that students initially read the tasks in the provided order and this might already be sufficient to gear their appraisals and emotions into

the directions observed in the present research. Moreover, computerized test environments make it easier to guide students through a test and determine the order in which certain tasks are encountered. It is also plausible that our findings are not limited to exams students write in school classes. We focused on the test setting to link our work to existing research on task order effects on test performance and because it is easier to establish specific task orders in tests than in unsupervised learning settings like doing homework. However, in the educational context, our results should be transferrable to situations in which students work on assigned tasks (e.g., homework assignments), in which sorting tasks from easy to difficult might have similarly negative effects. This should be investigated in subsequent research. Also, it is conceivable that our results are relevant to other academic and non-academic tests as well, informing the constructors of intelligence tests, assessment centers, sport competitions, or art and music contests about the potential consequences of ordering tasks according to their difficulty for the emotional experiences of the test takers.

Conclusion

We investigated the assumption that starting tests with easy tasks fosters a positive test experience in comparison to starting with difficult tasks. Our data provide no support for this assumption, instead suggesting that starting with easy tasks might have adverse effects on students appraisals and emotions towards the end of a test. This might be interpreted in terms of a contrast effect, which merits attention to the effects of task order in future research. Overall, our research findings align with general recommendations to “eat the frog first” when it comes to designing tests at school and beyond (e.g., Habbert & Schroeder,

2020).

References

AQA (2016). Our exams explained: GCSE science exams from summer 2018. AQA Education. Retrieved April 22, 2021, from https://filestore.aqa.org.uk/resources/science/AQA-GCSE-SCIENCE-EXAMS-EXPLAINED.PDF

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.

Baines, L., & Goolsby-Smith, R. (2016). America's obsessive-assessment disorder. *Counterpoints*, *492*, 61–74. https://doi.org/10.2307/45157503

Bieg, M., Goetz, T., & Hubbard, K. (2013). Can I master it and does it matter? An intraindividual analysis on control–value antecedents of trait and state academic emotions. *Learning and Individual Differences*, *28*, 102–108. https://doi.org/10.1016/j.lindif.2013.09.006

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *The Journal of Applied Psychology*, *82*(2), 300–310. https://doi.org/10.1037/0021-9010.82.2.300

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

Flaugher, R. L., Melton, R. S., & Myers, C. T. (1968). Item rearrangement under typical test conditions. *Educational and Psychological Measurement*, *28*(3), 813–824. https://doi.org/10.1177/001316446802800310

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007a). Girls and mathematics —A "hopeless" issue? A control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education*, *22*(4), 497–514. https://doi.org/10.1007/BF03173468

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007b). Perceived learning environment and students' emotional experiences: A multilevel analysis of mathematics classrooms. *Learning and Instruction*, *17*(5), 478–493. https://doi.org/10.1016/j.learninstruc.2007.09.001

Fulmer, S. M., & Tulis, M. (2013). Changes in interest and affect during a difficult reading task: Relationships with perceived difficulty and reading fluency. *Learning and Instruction*, *27*, 11–20. https://doi.org/10.1016/j.learninstruc.2013.02.001

Goetz, T., Bieleke, M., Gogol, K., van Tartwijk, J., Mainhard, T., Lipnevich, A. A., & Pekrun, R. (2021). Getting along and feeling good: Reciprocal associations between student-teacher relationship quality and students' emotions. *Learning and Instruction*, *71*, 101349. https://doi.org/10.1016/j.learninstruc.2020.101349

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*(1), 44–58. https://doi.org/10.1016/j.cedpsych.2009.10.001

Goetz, T., Keller, M. M., Lüdtke, O., Nett, U. E., & Lipnevich, A. A. (2020). The dynamics of real-time classroom emotions: Appraisals mediate the relation between students' perceptions of teaching and their emotions. *Journal of Educational Psychology*, *112*(6), 1243–1260. https://doi.org/10.1037/edu0000415

Goetz, T., & Kleine, M. (2006). Emotionales Erleben im Mathematikunterricht [Emotional experience in mathematics classes]. *Mathematik Lehren*, *135*, 4–9.

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., & 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*(3), 188–205. https://doi.org/10.1016/j.cedpsych.2014.04.002

Habbert, R., & Schroeder, J. (2020). To build efficacy, eat the frog first: People misunderstand how the difficulty-ordering of tasks influences efficacy. *Journal of Experimental Social Psychology*, *91*, 104032. https://doi.org/10.1016/j.jesp.2020.104032

Hauck, K. B., Mingo, M. A., & Williams, R. L. (2017). A review of relationships between item sequence and performance on multiple-choice exams. *Scholarship of Teaching and Learning in Psychology*, *3*(1), 58–75. https://doi.org/10.1037/stl0000077

Institute for Educational Quality Improvement. (2019). *Vergleichsarbeiten 3. und 8. Jahrgangsstufe (VERA-3 und VERA-8) [Comparison tests 3rd and 8th grades (VERA-3 and VERA-8]*. https://www.iqb.hu-berlin.de/vera

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment*, *4*. https://ejournals.bc.edu/index.php/jtla/article/view/1649

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 academic self-concept and boredom. *Learning and Individual Differences*, *69*, 206–218. https://doi.org/10.1016/j.lindif.2018.10.004

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education*, *58*(3), 196–217. https://doi.org/10.2307/27798138

Lenth, R. (2020). *emmeans: Estimated marginal means, aka least-squares means*. https://CRAN.R-project.org/package=emmeans

Li, W., Lee, A., & Solmon, M. (2007). The role of perceptions of task difficulty in relation to self-perceptions of ability, intrinsic value, attainment value, and performance. *European Physical Education Review*, *13*(3), 301–318. https://doi.org/10.1177/1356336X07081797

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse [Test construction and test analysis]* (6th ed.). Beltz.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

Miller, M. D., Gronlund, N. E., & Linn, R. L. (2009). *Measurement and assessment in teaching* (10th ed.). Merrill / Pearson.

Muthén, L. K., & Muthén, B. O. (1998-2018). *Mplus User's Guide* (8th ed.). Muthén & Muthén.

Nagy, G., Nagengast, B., Becker, M., Rose, N., & Frey, A. (2018). Item position effects in a reading comprehension test: An IRT study of individual differences and individual correlates. *Psychological Test and Assessment Modeling*, *60*(2), 165–187.

Pekrun, R. (2000). A Social-cognitive, control-value theory of achievement emotions. In J. Heckhausen (Ed.), *Advances in Psychology. Motivational psychology of human development: Developing motivation and motivating development* (Vol. 131, pp. 143–163). North-Holland. https://doi.org/10.1016/S0166-4115(00)80010-2

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

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

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (2017). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child Development*, *88*(5), 1653–1670. https://doi.org/10.1111/cdev.12704

Perry, R. P., Hladkyj, S., Pekrun, R., & Pelletier, S. T. (2001). Academic control and action control in the achievement of college students: A longitudinal field study. *Journal of Educational Psychology*, *93*(4), 776–789. https://doi.org/10.1037/0022-0663.93.4.776

R Core Team. (2020). *R: A language and environment for statistical computing.* Vienna, Austria. R Foundation for Statistical Computing. https://www.R-project.org/

Rotter, J. B. (1954). *Social learning and clinical psychology*. Prentice-Hall, Inc. https://doi.org/10.1037/10788-000

Schwarz, N., & Bless, H. (1992). Assimilation and contrast effects in attitude measurement: An inclusion/exclusion model. *Advances in Consumer Research*, *19*(1), 72–77.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). *afex: Analysis of Factorial Experiments.* https://CRAN.R-project.org/package=afex

Skinner, N. F. (2009). Academic folk wisdom: Fact, fiction and falderal. *Psychology Learning & Teaching*, *8*(1), 46–50. https://doi.org/10.2304/plat.2009.8.1.46

Sparfeldt, J. (2013). „Schwere Aufgaben nach hinten?“ Aufgabenreihenfolge und Mathematikleistung in schriftlichen Prüfungen [“Easy items first?“ - Item order and mathematical performance]. *Psychologie in Erziehung und Unterricht*, *60*(2), 133–142. https://doi.org/10.2378/peu2013.art11d

Towle, N. J., & Merrill, P. F. (1975). Effects of anxiety type and item-difficulty sequencing on mathematics test performance. *Journal of Educational Measurement*, *12*(4), 241–249. https://doi.org/10.1111/j.1745-3984.1975.tb01025.x

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, *82*(2), 247–252. https://doi.org/10.1037/0021-9010.82.2.247

Weinstein, F. M., Healy, C. C., & Ender, P. B. (2002). Career choice anxiety, coping, and perceived control. *The Career Development Quarterly*, *50*(4), 339–349. https://doi.org/10.1002/j.2161-0045.2002.tb00582.x

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Table 1

Descriptive Statistics of Perceived Challenge, Cognitive Appraisals, and Emotions

	Part 1						Part 2					
	Condition						Condition					
	Easy → Difficult		Difficult → Easy				Easy → Difficult		Difficult → Easy			
Variable	*M*	*SD*	*M*	*SD*	*d*	Sig.	*M*	*SD*	*M*	*SD*	*d*	Sig.
Before the tasks												
Control	3.91	0.88	4.00	0.85	0.08	ns	3.03	1.35	2.88	1.35	0.15	ns
Value	3.13	1.06	3.24	1.13	0.16	ns	3.06	1.22	3.09	1.29	0.08	ns
Enjoyment	2.88	1.01	2.71	1.12	0.21	ns	1.89	0.94	1.89	0.98	0.01	ns
Pride	2.03	0.98	1.96	0.96	0.10	ns	1.50	0.77	1.50	0.82	0.01	ns
Anxiety	1.95	1.03	1.87	1.07	0.13	ns	1.59	0.92	1.53	0.85	0.09	ns
Anger	1.70	1.00	1.80	1.10	0.11	ns	1.70	1.03	1.63	0.85	0.09	ns
Boredom	2.08	1.11	1.96	1.01	0.14	ns	1.94	1.23	1.66	0.94	0.32	ns
During the tasks												
Perceived Challenge	3.02	0.65	3.57	0.70	0.88	*	3.76	0.69	2.74	0.76	1.69	*
Control	3.26	1.10	3.07	1.18	0.19	ns	3.11	1.16	3.59	1.10	0.50	*
Value	3.00	1.36	2.85	1.22	0.20	ns	2.52	1.29	3.07	1.28	0.70	*
Enjoyment	1.92	0.96	1.89	1.04	0.04	ns	1.72	0.92	2.16	1.23	0.51	*
Pride	1.61	0.83	1.63	0.86	0.04	ns	1.49	0.71	1.82	1.06	0.45	*
Anxiety	1.76	1.04	1.76	1.11	0.00	ns	1.54	0.91	1.37	0.78	0.25	ns
Anger	2.19	1.09	2.38	1.16	0.20	ns	1.97	1.16	1.64	0.99	0.35	*
Boredom	1.66	0.98	1.57	0.92	0.10	ns	2.15	1.38	1.83	1.10	0.39	*

Note. d = effect size (Cohen's d) of simple effects comparing the two conditions. Sig. = Significance. All variables were measured on 5-point Likert scales (from 1 to 5), with higher values corresponding to a more pronounced experience.
* $p < .05$.

Figure 1

The Predicted Reversal of Perceived Challenge from Part 1 to Part 2 of the Math Test Between the Easy-to-Difficult and the Difficult-to-Easy Condition

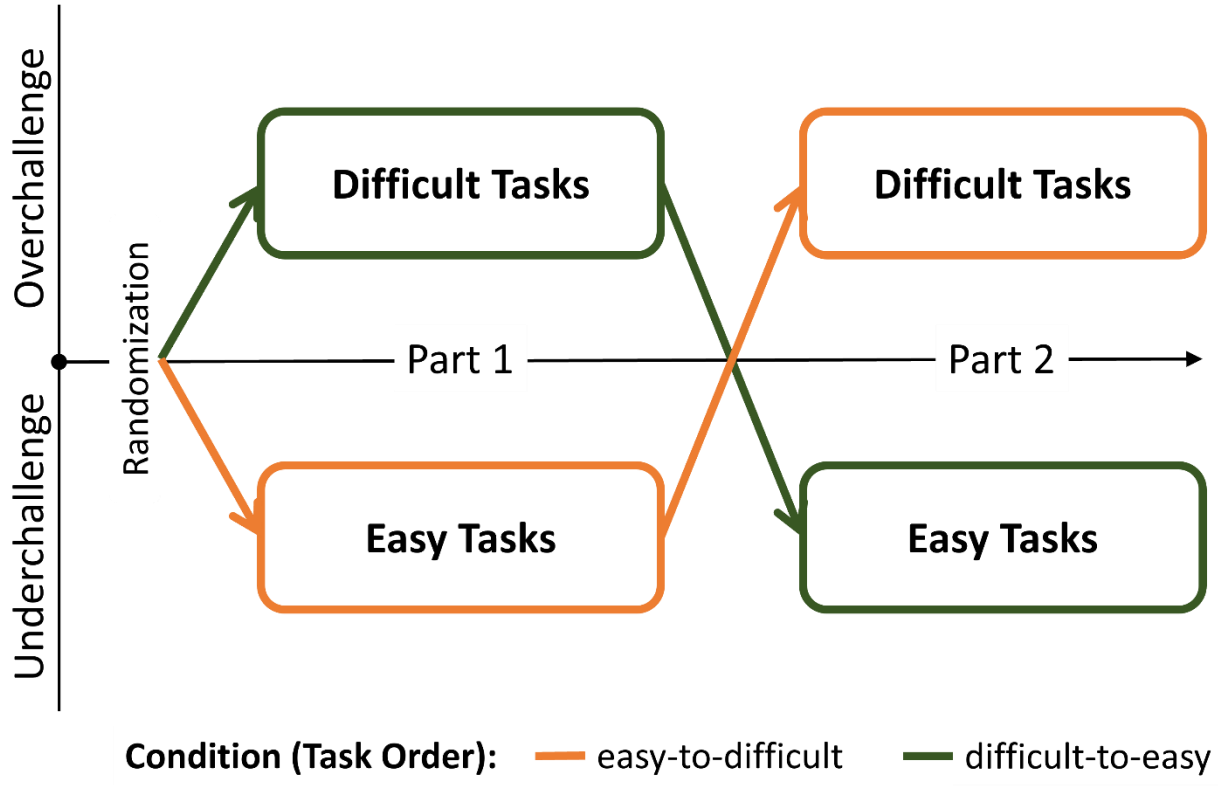
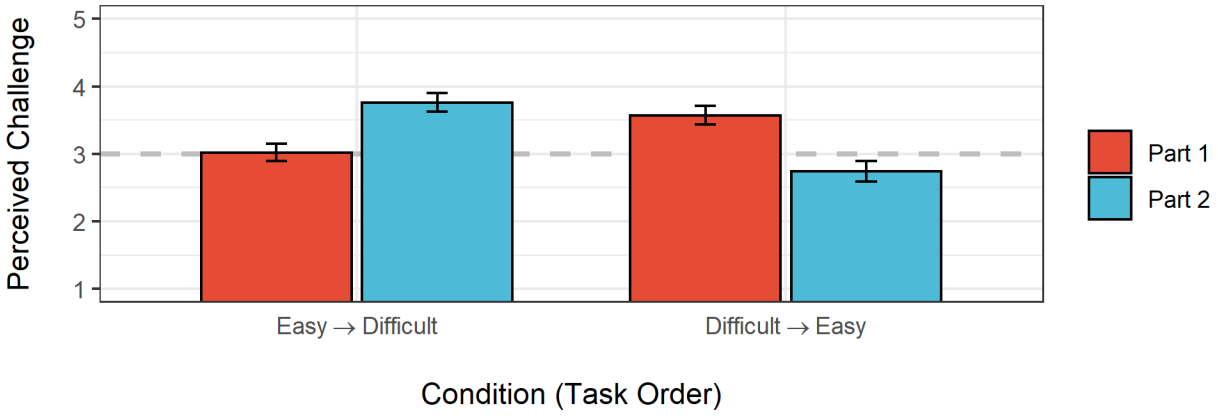


Figure 2

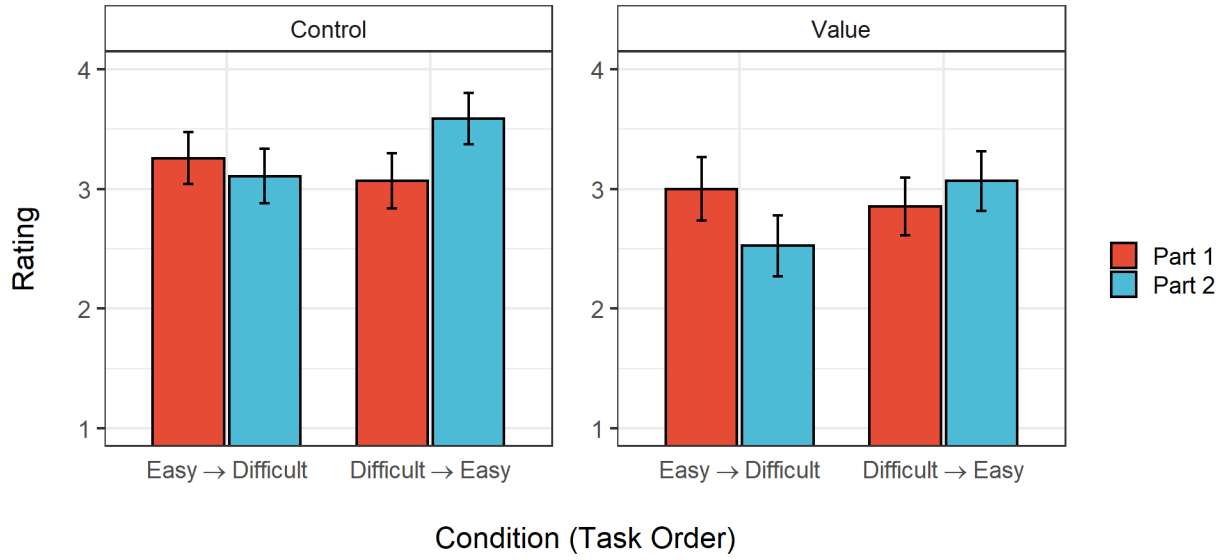
Perceived Challenge in Part 1 and 2 of the Math Test as a Function of Condition



Note: Error bars indicate 95% confidence intervals of the mean. Values below the dashed line represent underchallenge, values above the dashed line represent overchallenge.

Figure 3

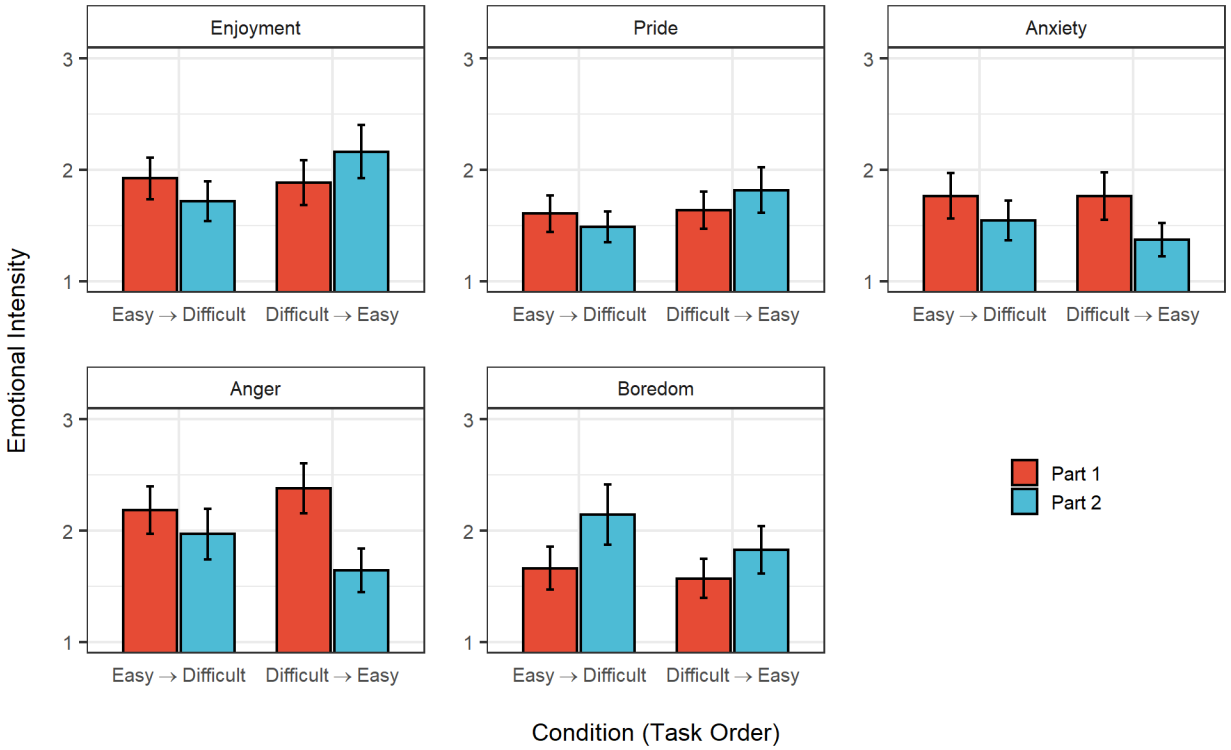
Control and Value Appraisals During Part 1 and 2 of the Math Test as a Function of Condition



Note: Error bars indicate 95% confidence intervals of the mean.

Figure 4

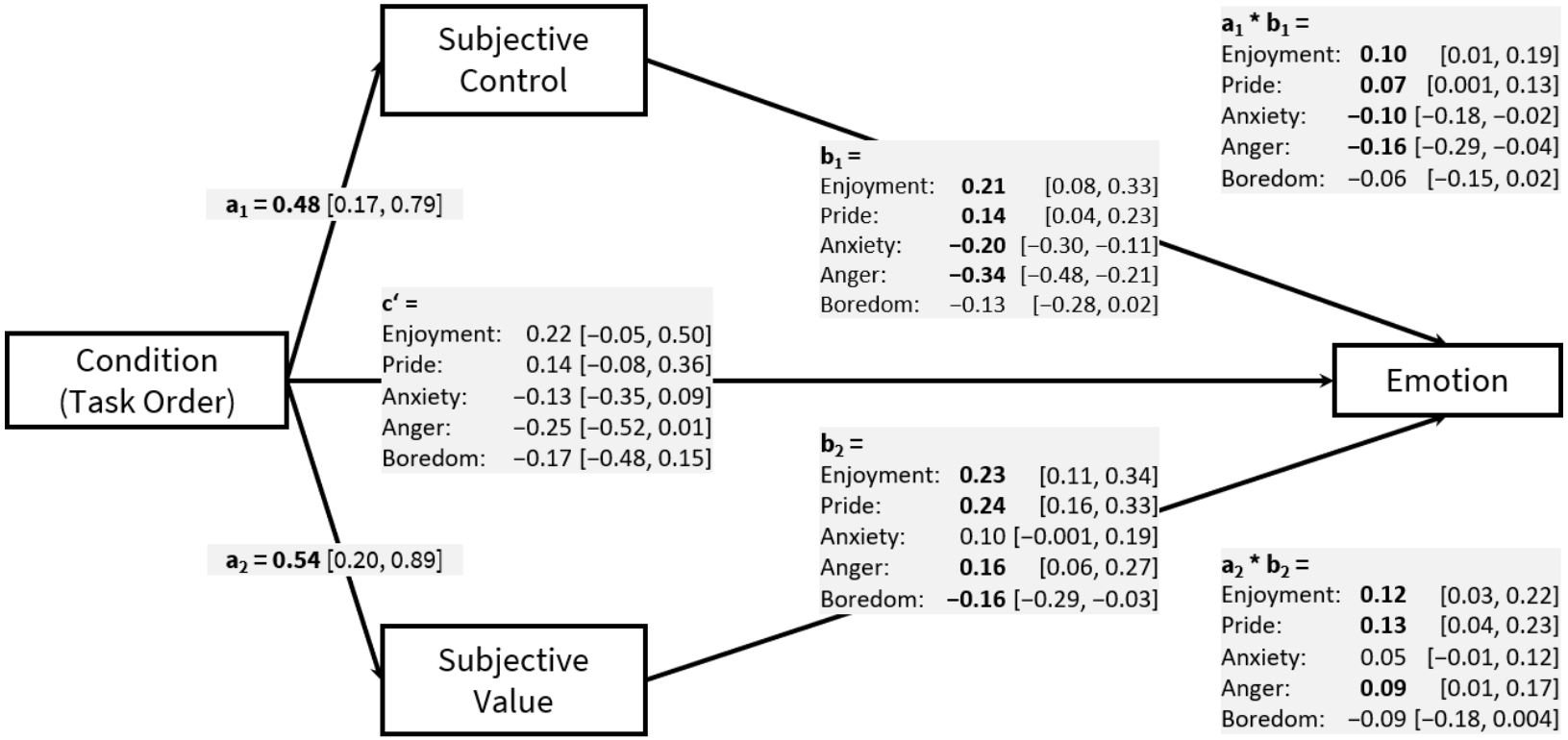
Achievement Emotions During Part 1 and 2 of the Math Test as a Function of Condition



Note: Error bars indicate 95% confidence intervals of the mean.

Figure 5

Estimated Paths of the Models Testing Whether Control and Value Mediate Differences between Conditions in Terms of Emotions During Part 2 of the Math Test



Note: Each path is labelled with the unstandardized coefficients from 5 different mediation models. The estimated paths a_1 and a_2 varied negligibly across these models. Values in square brackets represent bootstrapped 95% confidence intervals. Significant paths are highlighted in bold face. Condition represents a categorical variable with value 0 for the easy-to-difficult condition and 1 for the difficult-to-easy condition. Thus, positive values indicate higher values in the difficult-to-easy condition.