Stress or failure? An experimental protocol to distinguish the environmental determinants of decision-making

Martina Vecchi (mmv5343@psu.edu)
The Pennsylvania State University https://orcid.org/0000-0002-8239-3180
Nicolai Vitt
University of Bristol

Research Article

Keywords: Acute stress, failure, online experiment, experimental protocol

Posted Date: November 7th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3570228/v1

Stress or failure? An experimental protocol to distinguish the environmental determinants of decision-making*

Martina Vecchi† Nicolai Vitt‡

November 4, 2023

Abstract

Are economic decisions affected by short-term stress, failure, or both? Such effects have not been clearly distinguished in previous experimental research, and have the potential to worsen economic outcomes, especially in disadvantaged socioeconomic groups. We validate a novel experimental protocol to examine the individual and combined influences of stress, failure, and success. The protocol employs a 2x3 experimental design in two sessions and can be used online or in laboratory studies to analyse the impact of these factors on decision-making and behaviour. The stress protocol was perceived as significantly more stressful than a control task, and it induced a sizeable and significant rise in state anxiety. The provision of negative feedback ("failure") significantly lowered participants' assessment of their performance, induced feelings of failure, and raised state anxiety.

Keywords: Acute stress; failure; online experiment; experimental protocol.

JEL Classification code: C9, D91.

*__Competing Interests and Funding__: 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. We gratefully acknowledge funding of this project from the University of Edinburgh, School of Economics and from Pennsylvania State University, Department of Agricultural Economics, Sociology and Education. Nicolai Vitt furthermore received funding from the Economic and Social Research Council, award 1651922. This work was furthermore supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under Project number PEN04709 and Accession number 1019915. **Acknowledgments**: We thank the LEMA lab at Penn State University, and in particular Anthony Kwasnica and Rashmi Sharma, for their help with the study. We thank Michèle Belot, Pablo Brañas-Garza, Steven Dieterle, Lorenz Götte, Edward Jaenicke, Jonathan James, Tatiana Kornienko, Trent Smith, Andreas Steinhauer, Till Stowasser, Roberto A. Weber, Conny Wollbrant, Katherine Y. Zipp, as well as seminar participants at the SGPE conference, the University of Stirling, the University of Edinburgh, the Pennsylvania State University and the University of Bath for their helpful comments.

†Department of Agricultural Economics, Sociology, and Education, The Pennsylvania State University, University Park, PA, United States. Email: martinavecchi@psu.edu, Corresponding author

‡School of Economics, University of Bristol, Bristol, United Kingdom. Email: nicolai.vitt@bristol.ac.uk

1 Introduction

Individuals making economic decisions often operate in stressful environments and with previous experiences of both success and failure on their minds. The experience of situational acute stress can create anxiety, divert attention, and impact the body and mind[1]. Additionally, individuals who recall past poor decisions may experience reduced confidence and diminished willingness to take further actions. Prior successes, failures, and acute stress can have complex simultaneous effects on individuals' decisions. Previous experimental studies have been unable to clearly distinguish between the influences of these factors. We conducted a behavioural online experiment to validate a novel experiment protocol that separates the effects of short-term stress and failure, allowing researchers to assess their effects individually and in combination.

Acute stress can arise from the decision-making process itself and from external factors such as financial and relationship concerns, and high-pressure work environments. Stress is the body's response to a short-term demand or pressure and manifests as physiological changes such as a rapid heart rate and as psychological effects such as feelings of anxiety (Daviu *et al.*, 2019; Giannakakis *et al.*, 2017). Experimental studies in behavioural economics and the social and biological sciences have shown that acute stress affects decision-making in settings from financial to health care (Rutters *et al.*, 2009; Von Dawans *et al.*, 2012; Delaney *et al.*, 2014; Cahlíková and Cingl, 2017; Bendahan *et al.*, 2017; Haushofer *et al.*, 2018). Several found that stress caused a shift in cognitive effort towards the stressor and away from other tasks (Cohen, 1980; Allen and Armstrong, 2006; Starcke *et al.*, 2016), and that stress impaired working memory and cognitive flexibility (see Shields *et al.*, 2016, for a review). Self-preservation instincts can also translate to a greater effort to preserve financial and other resources (Dickerson and Kemeny, 2004).

Failures are past experiences that we view as unsuccessful because we did not meet our own or others' expectations. Feelings of failure may arise from past decisions that did not yield favorable outcomes or from a broader perception of inadequacy, possibly due to unemployment, stalled career progress, or financial challenges. Experimental research has shown failure to affect how we make decisions (Gill and Prowse, 2014; Buser and Dreber, 2016; Buser, 2016; Cassar and Klein, 2019). Failures have been found to provoke negative emotions such as guilt and shame (Carver and Scheier, 1990; Bohns and Flynn, 2013), alter individuals' sense of self-worth (Heatherton and Polivy, 1991; Crocker and Wolfe, 2001; Crocker and Park, 2004), and drive them to abandon the task (Crocker and Wolfe, 2001), or attempt to compensate for prior losses (Thaler and Johnson, 1990). A history of failure may generate a self-reinforcing cycle of "poor" choices and setbacks, contributing to the persistence of poverty (Stevens, 1999) and creating a barrier to socioeconomic mobility (Corak, 2013).

Previous experimental research could not separate the effects of acute stress from the effects of failure because the protocols incorporated elements of both. The Trier Social Stress Test ((Kirschbaum *et al.*, 1993) asks participants to give a presentation and complete a mental arithmetic task and informs the participants that their performance will be evaluated, likely provoking fear of failure. Other stress protocols have asked participants to complete unsolvable riddles or extremely difficult cognitive tasks (Habhab *et al.*, 2009; Rutters *et al.*, 2009) in

[1]In the following, acute stress, short-term stress, and stress will be used interchangeably when describing stress or stressors.

which failure is used as the stressor. Similarly, experimental studies of failure have relied on competitive settings which inherently also involve elements of stress (Buckert *et al.*, 2017).

Stress and failure are not intrinsically linked. Within educational and occupational environments, elevated workloads can elicit stress independently of any failure. Similarly, perceptions of failure, such as those stemming from insufficient stimulation or negative performance feedback in academic or professional settings, do not invariably coincide with stress. Given that stress and failure can exert distinct influences on decision-making processes, a more effective protocol to discern their respective roles will help pinpoint affected groups and contexts, informing targeted policy interventions.

Our experiment employs a 2x3 two-session design to manipulate participants' decision-making environment. First, participants complete a stress-inducing task in either the first or second session, creating exogenous variation in acute stress levels. In the second treatment, we varied whether participants received no performance feedback, feedback relative to a low threshold (success), or relative to a high threshold (failure) before decision-making tasks in the second session. This introduced variation in perceived success or failure.

The two session design is crucial to induce failure/success separately from stress: All participants received feedback related to the stress task, ensuring consistency. Those who completed the stress task in the first session received success or failure feedback in the second session when no longer exposed to stress, while those who completed the stress task in the second session received feedback while still experiencing stress. This design generates six different decision environments and allows between-subject and within-subject comparisons across the two sessions.

We find strong evidence that the protocol induced short-term psychological stress among the participants, who reported significantly greater levels of stress from the stress-inducing task than from the control task and reported statistically significant increases in their anxiety following the stress task. We also find strong evidence that negative feedback affected participants' assessments of their performance, triggered emotional responses, and induced stronger feelings of failure. The response to positive feedback, on the other hand, did not differ strongly from the response to no feedback.

Our study makes several contributions to the literature. First, we develop an experimental design that allows researchers to distinguish the individual effects of acute stress, failure, and success. We also establish a protocol for experiments that induces realistic cognitive stress in student populations and propose a novel feedback protocol that allows researchers to vary participants' perceptions of success and failure and induce emotional responses. These protocols can be used in online and laboratory settings to investigate the impacts of stress and failure separately and in combination.

This paper is structured as follows. Section 2 describes the experimental design. Section 3 presents the results. Section 4 concludes.

2 Experimental design

2.1 Sample and recruitment

Our online experiment validated a protocol to test the effects of stress and feedback on decisions participants made in two sessions that were scheduled at least seven days apart. The first session lasted approximately 50 minutes and the second session lasted approximately 60 minutes. We conducted the experiment between 04 June and 31 July, 2020, and recruited students at Pennsylvania State University after ethical approval of the study from the university's Institutional Review Board. We pre-registered the experiment and hypotheses tested in the AEA RCT registry under the following trial ID: AEARCTR-0005946. Details can be found at http://www.socialscienceregistry.org/trials/5946. The experiment design was pre-tested in May 2020 with a small sample of students.

In the experiment, participants performed four different decision-making outcome tasks. Difficulties associated with the COVID-19 pandemic restricted our ability to recruit participants in 2020, leading to a smaller sample size (final sample of 269 students) than specified and reducing the power of the study to detect treatment effects on the decision-making outcome tasks. Future research will focus on the outcomes of decisions under stress and failure/success after collecting data from additional participants.

2.2 Procedure and randomization

As shown in Table 1, our 2x3 two-session experiment design produced six conditions that varied in the timing of stress and experience of failure/success. The experimental conditions were randomized at the participant level and introduced random variation in the decision environment of participants.

Treatment 1 varied the timing of the stress task. At the beginning of the first experiment session, half of the participants were assigned the stress-neutral control task (S2 in Table 1): answer demographics questions, read a short text and answer questions about it. The other half were assigned the stress task and completed an incentivized cognitive task (S1). In the second session about a week later, participants were assigned to the task they did not complete in the first session, so each participant completed one control session and one stress session.

The incentive structure for the stress task involved a potential deduction from participants' pay-off if they performed below a threshold. Treatment 2 varied whether participants received feedback about their performance in the stress task relative to the threshold and the difficulty of the threshold. Feedback was presented in the second session. At that point, half the participants had completed the stress task that day and the other half had completed the stress task approximately a week earlier in the first session. This temporal break between stress and feedback allowed us to analyse the impacts of stress and failure/success separately. The three feedback conditions were no feedback (S1-N, S2-N in Table 1), a success condition that provided feedback relative to a low threshold for success (S1-S, S2-S), and a failure condition that gave feedback relative to a high threshold for success (S1-F, S2-F).

Table 1: Experimental treatments and conditions

(a) **During the first session:**

Treatment 1: Timing of stress task	
Stress in the 1st session (S1)	Stress in the 2nd session (S2)
S1: Stress only	S2: Control

(b) **During the second session:**

		Treatment 1: Timing of stress task	
		Stress in the 1st session (S1)	Stress in the 2nd session (S2)
Treatment 2: Feedback	No feedback (N)	S1-N: Control	S2-N: Stress only
	Low threshold (S)	S1-S: Success only	S2-S: Stress & Success
	High threshold (F)	S1-F: Failure only	S2-F: Stress & Failure

The experiment presented two experimental conditions during the first session (stress and no-stress per Treatment 1), and six in the second session (stress and no-stress combined with the three types of feedback from Treatment 2). In the second stage of each session, participants completed the decision-making outcome tasks. At three points in each session (at the beginning, after completion of the task, and at the end), participants completed the six-item short-form state anxiety inventory developed by Marteau and Bekker (1992). They also completed brief questionnaires. First-session questions addressed their perceptions of the assigned task. Second-session questions again addressed their perceptions of the task assigned that day and collected information about their demographic characteristics, actual stress experiences, failures, and decision-making processes. The full timeline of the experimental sessions is displayed in Online Appendix A. The experimental instructions shown to participants can be found in Online Appendix B.

Participants received a fixed participation fee of \$5 at the end of their second session. 50% of the participants received payment based on their performance in the stress task. The other 50% was paid based on one of the decision-making outcome tasks. Minimum compensation was \$5, maximum compensation was \$45, and the mean for the study was \$20.96. All payments were made via PayPal within 48 hours of completion of the second session. This arrangement minimized potential attrition effects between the first and second session, avoided influence from wealth and income effects, and provided no information about which task would be for payment until the end of the second session.

2.3 Stress protocol

The stress protocol was designed to mimic stressors commonly experienced by students and was framed as a "block of several tasks that are similar to test questions you may face in assessment tests when applying for jobs or for graduate school."

To induce stress, we incorporated a financial incentive with potential losses of pay-off, cognitive pressure, time pressure, and distractions in the stress task. All have been shown to induce stress in study participants. The stress task required them to complete up to 18 short cognitive tasks in 10 minutes, and they were penalized for incorrect answers and leaving tasks undone. As a distraction, the program intermittently displayed brief incentivized knowledge and arithmetic questions on the screen during the ten-minute task block to induce additional stress.

The task was loosely based on the stress task used in Vitt *et al.* (2021), which was perceived by the study sample of low-income mothers as significantly more stressful than a control task and induced a significant increase in participants' heart rate. We adapted this protocol to the different population of interest, university students. Differently from other studies (e.g. the Tier Social Stress Test by Kirschbaum *et al.*, 1993, or the cold pressor test used in Delaney *et al.*, 2014), and similarly to Vitt *et al.* (2021), our aim is not to trigger strong physiological responses (such as the stress hormone cortisol), but to mimic stressors university students experience in real life.

Participants were initially allocated a maximum potential incentive of 5000 tokens ($40). Participants' performance in the cognitive tasks and pop-up questions determined how much of the initial endowment they "lost." The incentives are framed as a loss to avoid inducing positive feelings from "winning." Participants could lose up to 4050 tokens ($32.40) from the 18 cognitive tasks and 200 tokens ($1.60) from the pop-up questions. In addition, 750 tokens ($6.00) would be deducted if their performance in the stress task was below an undisclosed threshold. The level of this threshold was experimentally varied by Treatment 2 (described below) to induce failure or success. Participants were informed that they would be randomly assigned to a group of participants who all faced the same threshold.

A detailed description of the stress task and the incentive structure, as well as several sample screenshots can be found in Online Appendix C.

2.4 Control tasks

In the control sessions, participants were given four minutes to complete a short demographics questionnaire. They then had ten minutes to read a text about the possibility of life on Mars (518 words) and a text about the evolution of languages (629 words) and to answer three simple questions about each text.

This task required participants to pay attention but was not meant to induce stress. The time required was equal to the time required of participants to review the instructions for the stress task plus time allotted to complete it. No financial or time pressure was induced by the control task.

2.5 Feedback on failure or success

Treatment 2 assigned each participant in the second session to a feedback condition. Following the stress or control task, the program informed participants assigned to the no-feedback control (S1-N and S2-N) that there would be a 60-second wait before proceeding. For participants assigned to the feedback treatments (S1-S, S2-S, S1-F, or S2-F), the screen displayed failure/success feedback as a statement that noted only whether they had met the threshold to avoid the 750-token penalty. It was displayed for 60 seconds. Screenshots are presented in Figures C.5 - C.7 of Online Appendix C.

The threshold in the success condition (S1-S, S1-S) was 1,250 tokens such that approximately 95% of participants were expected to succeed. The threshold in the failure condition (S1-F, S2-F) was 4,300 tokens such that approximately 95% of participants were expected to fail.

This random variation in perception of success and failure (conditional on participants' performance) was independent from the timing of the stress task, allowing us to overcome any confounding factors like their ability, and to identify the influence of success/failure.

We assess our protocol's effectiveness in manipulating feelings of failure by collecting participants' self-reports of their perceived successfulness, performance, and post-feedback emotions. We expected failure feedback to increase feelings of failure, decrease self-assessments of performance, and induce negative emotions. Details on these measures of failure / success are in Online Appendix D.2.

2.6 Empirical approach

To be able to identify the impacts of stress and failure individually and when combined, participants completed the outcome tasks of interest at the end of the first and the second sessions. This design allows a between-subject and within-subject comparison across the two sessions. The fixed-effects model used for this analysis and the corresponding power calculations are presented in Online Appendices E.1 and F. Furthermore, to deal with non-random attrition between sessions, we used a Heckman model of selection (Heckman, 1979; Wooldridge, 2010), which is outlined in Online Appendix E.2.

3 Results

3.1 Descriptive statistics

The demographic characteristics of the sample in our validation study are provided in Table G.1 in Online Appendix G.1. The average age of participants was 23.5 years with 66% of participants identifying as female and 32% as male. All participants were students; 72% were undergraduates and 28% were post-graduates. Our comparison of responses from participants assigned to the stress condition in the first session (S1) with those assigned to the stress condition in the second session (S2) shows a significant difference in average participant age. We report our demographic comparisons by feedback condition in Table G.2 in Online Appendix G.1.

Of the 317 participants who completed the first session, 269 (84.86%) returned and successfully completed the second session, 12 (3.79%) provided invalid or incomplete responses in the second session and were dropped, and 36 (11.36%) did not return for the second session. In Online Appendices G.2 and G.3 we further examine these attrition issues and show that our results are robust to accounting for non-random attrition.

3.2 Effectiveness of stress protocol

Our protocol was designed to induce acute stress, and we used psychological stress measures to analyse its effectiveness. As shown in Table 2, the stress task was rated as substantially more stressful than the control task on a Likert scale from 1 (not at all) to 5 (very much). The mean stressfulness score was 3.5 for the stress task and 2.1 for the control task. The sizeable difference of 1.4 is equivalent to 1.33 standard deviations and is highly statistically significant. We found only small and not statistically significant differences in perceived stress by feedback condition. The results in Table 3 show that the stress task was perceived, on average, as significantly less relaxing, less easy, more difficult, less enjoyable, less successful, and more tiring than the control task.

Table 2: Perceived stressfulness of the stress / control task

	(1)	(2)	(3)	Differences:	
	No feedback	Success	Failure	(2) - (1)	(3) - (1)
Stress	3.483	3.583	3.763	0.100	0.280
	(0.082)	(0.178)	(0.169)	(0.196)	(0.187)
Control	2.181	1.902	2.186	−0.278	0.005
	(0.082)	(0.151)	(0.153)	(0.171)	(0.173)
Difference:	1.302***	1.681***	1.577***	0.379	0.275
Stress - Control	(0.113)	(0.234)	(0.228)	(0.272)	(0.257)
Observations	342	77	81		
Participants	250	77	81		

Note: Means and mean differences were obtained using the sample of participants who responded to the task stressfulness question in both sessions. Participants with a performance level above the high or below the low threshold were excluded from the sample for simplicity. Standard errors were clustered at the participant level and are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

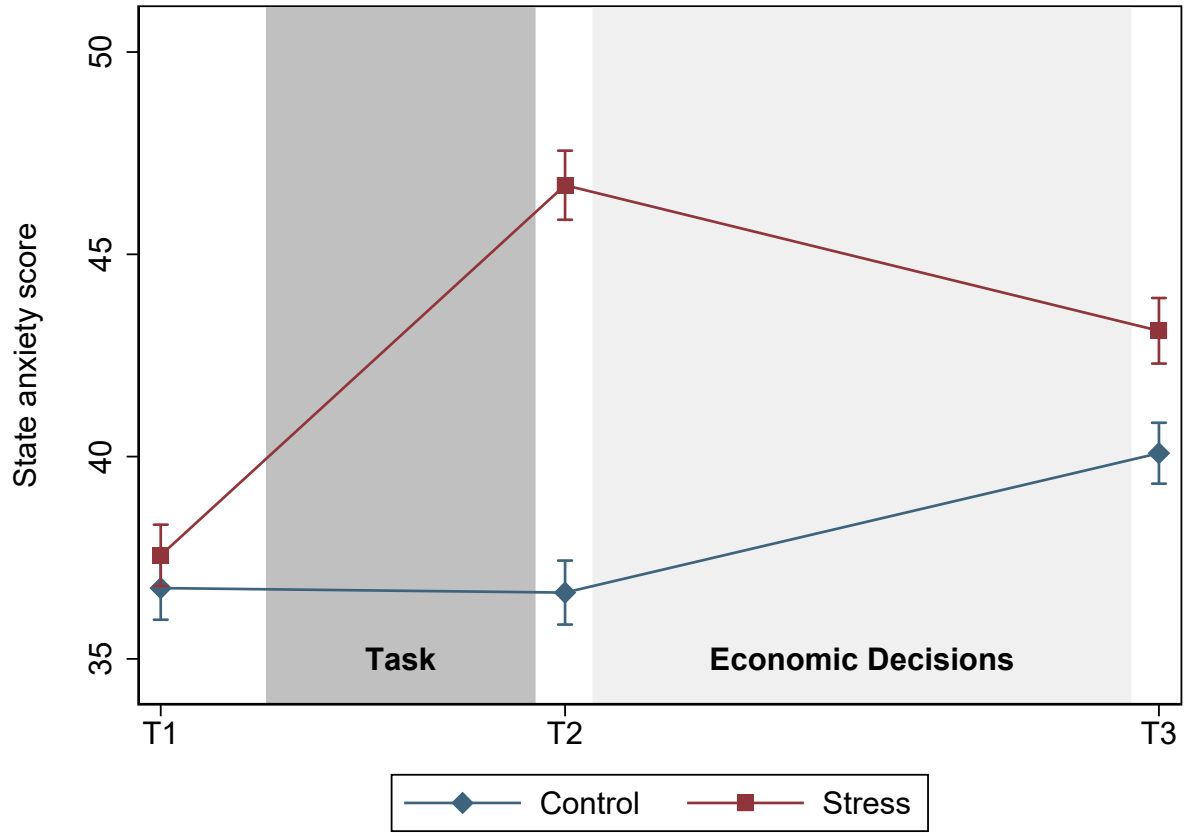
Figure 1 presents mean state anxiety scores in the stress and control conditions for the three measurements taken in each session. A difference-in-difference comparison of state anxiety levels is provided in Table G.3 of Online Appendix G. We find baseline mean anxiety levels (T1) of 37.6 and 36.8 in the stress and control conditions respectively, which are not statistically different and are in line with values reported in Marteau and Bekker (1992). The second anxiety measurement (T2) was conducted shortly after completion of the stress or control task. At that point, the mean anxiety level of the control participants was 36.7, nearly unchanged from the baseline level. The mean anxiety level of the stress-task participants rose from 37.6 to 46.7, a substantial increase of 9.1 (0.77 baseline standard deviations). We find that the increase in anxiety among stress-task participants and the difference in anxiety levels between the control and stress-task participants are both statistically significant. They are larger than ones reported for the Trier Social Stress Test for Groups (increase by approximately 7 points, Von Dawans *et al.*, 2011), and

Table 3: Perceptions of the stress / control task

	Stress	Control	Difference	Observations	Participants
Stressful	3.540	2.136	1.404***	500	250
	(0.069)	(0.066)	(0.089)		
Relaxing	2.235	3.195	−0.960***	502	251
	(0.067)	(0.064)	(0.082)		
Easy	2.779	4.237	−1.458***	498	249
	(0.070)	(0.055)	(0.081)		
Difficult	3.248	1.776	1.472***	500	250
	(0.070)	(0.052)	(0.085)		
Enjoyable	2.520	3.064	−0.544***	500	250
	(0.070)	(0.066)	(0.083)		
Successful	2.867	3.791	−0.924***	498	249
	(0.061)	(0.064)	(0.080)		
Tiring	2.908	2.156	0.752***	500	250
	(0.074)	(0.068)	(0.084)		

Note: Perceptions of the stress / control task were scored from 1 for 'not at all' to 5 for 'very much'. Means and mean differences were obtained using the sample of participants who responded to the task perception question in both sessions. Participants with a performance level above the high or below the low threshold were excluded from the sample for simplicity. Standard errors were clustered at the participant level and are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

Figure 1: State anxiety response to the stress / control task



Note: Means were calculated for the state anxiety scores at the three measurement points. Bands indicate +/- 1 standard error. Participants with a performance level above the high or below the low threshold were excluded here for simplicity.

for stress protocols based on unsolvable arithmetic tasks (increase by approximately 3 points, Rutters *et al.*, 2009). By the end of the session (T3) the gap in anxiety levels between the stress

and control conditions closed somewhat, indicating that the acute stress dissipated slightly when their attention was shifted from the task.

Overall, we find strong evidence that the stress protocol was effective in inducing mild short-term psychological stress.

3.3 Effectiveness of feedback

By randomizing the feedback and threshold for success we aim to induce feelings of failure or success among participants. We analyse whether the feedback provided novel information to participants about their performance and whether it modified perceived success in the stress task.

For the feedback to provide novel information, the protocol must prevent participants from keeping track of their performance. We find that participants' two self-assessment ratings are only weakly correlated with actual performance: a correlation coefficient of 0.068 for the Likert-scale measure and -0.120 for expectations about losing tokens. Figures G.1 and G.2 in Online Appendix G show the distribution of participants' performance prior to deductions. Given the mean performance score of 2386 tokens, participants lost, on average, 61.5% of the 4250 tokens associated directly with performance. With a standard deviation of 621 tokens (14.6%), the distribution is sufficiently narrow for feedback to provide meaningful information to participants. Nearly all of the participants (97.8%) performed at levels between the low and high thresholds.

Table G.4 in Online Appendix G summarizes participants' self-assessments of likely exceeding the success threshold. In the no-feedback condition, 66% (70%) of those who exceeded (did not exceed) their assigned thresholds expected to exceed the threshold. There is no significant difference, indicating that success was difficult to predict without feedback.

Table 4 compares perceived success in the stress and control tasks by feedback condition rated on a Likert scale of 1 (not at all successful) to 5 (very much). Participants in the stress condition who received failure feedback reported the lowest success scores: a mean of 2.4 relative to 2.9 for reports by participants who received no feedback, a statistically significant difference (0.56 standard deviations). Though the feedback did not relate to the control task, S1 participants who received failure feedback rated their success lower than S1 participants who received no feedback. The difference of 0.3 is smaller and is not statistically significant. Success feedback instead did not lead to significantly greater self-perceptions of success.

At the end of the second session, participants were asked to judge their performance in the stress task using a Likert scale of 0 (very bad) to 4 (very good). Table 5 presents a comparison of their responses by feedback condition. S2 participants who received failure feedback rated their performance significantly lower than S2 participants who received no feedback (difference of 0.6) and positive feedback (difference of 0.8 points / 0.76 SDs). S1 participants rated their performance significantly higher than the S2 participants, suggesting that participants' self-assessments improves with time. S1 participants who received failure feedback rated their performance lower on average than S1 participants who received no feedback (difference of 0.3) and positive feedback (difference of 0.5 points / 0.45 SDs), and only the former difference was

Table 4: Perceived successfulness of the stress / control task

	(1)	(2)	(3)	Differences:	
	No feedback	Success	Failure	(2) - (1)	(3) - (1)
Stress	2.932	3.028	2.405	0.096	−0.526**
	(0.071)	(0.145)	(0.161)	(0.161)	(0.175)
Control	3.830	3.927	3.512	0.097	−0.319
	(0.079)	(0.145)	(0.149)	(0.164)	(0.168)
Difference:	−0.898***	−0.899***	−1.106***	−0.001	−0.208
Stress - Control	(0.096)	(0.205)	(0.219)	(0.231)	(0.241)
Observations	341	77	80		
Participants	249	77	80		

Note: Means and mean differences were obtained using the sample of participants who responded to the task successfulness question in both sessions. Participants with a performance level above the high or below the low threshold were excluded from the sample for simplicity. Standard errors were clustered at the participant level and are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

statistically significant. We also once again find that responses following positive feedback were not significantly different from responses following no feedback.

Table 5: Perceived performance in the stress task

	(1)	(2)	(3)	Differences:		
	No feedback	Success	Failure	(2) - (1)	(3) - (1)	(3) - (2)
S2: Stress - Session 2	1.979	2.132	1.366	0.153	−0.613**	−0.766***
	(0.157)	(0.161)	(0.120)	(0.224)	(0.197)	(0.200)
S1: Stress - Session 1	2.574	2.767	2.295	0.193	−0.279	−0.472*
	(0.145)	(0.166)	(0.161)	(0.220)	(0.217)	(0.231)
Difference:	−0.596**	−0.636**	−0.930***	−0.040	−0.334	−0.294
Session 2 - Session 1	(0.213)	(0.231)	(0.201)	(0.314)	(0.293)	(0.306)
N	94	81	85			

Note: Performance perceptions were scored from 0 for 'very bad' to 4 for 'very good'. Participants with a performance level above the high or below the low threshold were excluded from the sample for simplicity. Standard errors are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

Table G.5 in Online Appendix G compares participant responses regarding their expected token losses by feedback condition. Among S2 participants, those who received failure feedback rated their expected loss of tokens significantly higher than those who received success feedback.

Finally, as shown in Table 6, we find that participants who received failure feedback felt significantly less pleased, calm, confident, encouraged, and successful and felt more angry, anxious, disappointed, sad, and embarrassed than participants who received success feedback. We further observed a significant increase from baseline in the state anxiety of participants who received failure feedback (see Figure 2). The participants in the control condition who received failure feedback had a greater average increase in state anxiety than participants in the stress condition who received success feedback or no feedback. These results confirm the ability of our feedback protocol to induce negative emotions and that success feedback does not mitigate increases in anxiety triggered by stress.

Overall, these findings provide strong evidence that the provision of feedback affected participants' assessment of their performance, their perceptions of failure/success, and their emotions.

Table 6: Recalled emotions after receiving feedback

	(1) Success	(2) Failure	Difference: Failure - Success
Pleased	3.580	1.894	−1.686***
	(0.114)	(0.096)	(0.149)
Angry	1.753	3.106	1.353***
	(0.105)	(0.125)	(0.163)
Calm	3.321	2.447	−0.874***
	(0.115)	(0.110)	(0.159)
Anxious	2.395	2.906	0.511**
	(0.143)	(0.121)	(0.188)
Confident	3.185	2.000	−1.185***
	(0.118)	(0.093)	(0.150)
Disappointed	2.123	3.929	1.806***
	(0.125)	(0.106)	(0.163)
Encouraged	3.284	1.929	−1.355***
	(0.122)	(0.091)	(0.152)
Sad	1.630	2.918	1.288***
	(0.102)	(0.122)	(0.159)
Embarrassed	1.704	3.047	1.343***
	(0.113)	(0.133)	(0.174)
Successful	3.457	1.812	−1.645***
	(0.105)	(0.088)	(0.137)
N	81	85	

Note: Emotions after receiving the feedback were scored from 1 for 'not at all' to 5 for 'very much'. Participants with a performance level above the high or below the low threshold were excluded from the sample for simplicity. Standard errors are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

Failure feedback, in particular, induced a strong response, while the response to success feedback did not differ substantially from receiving no feedback.

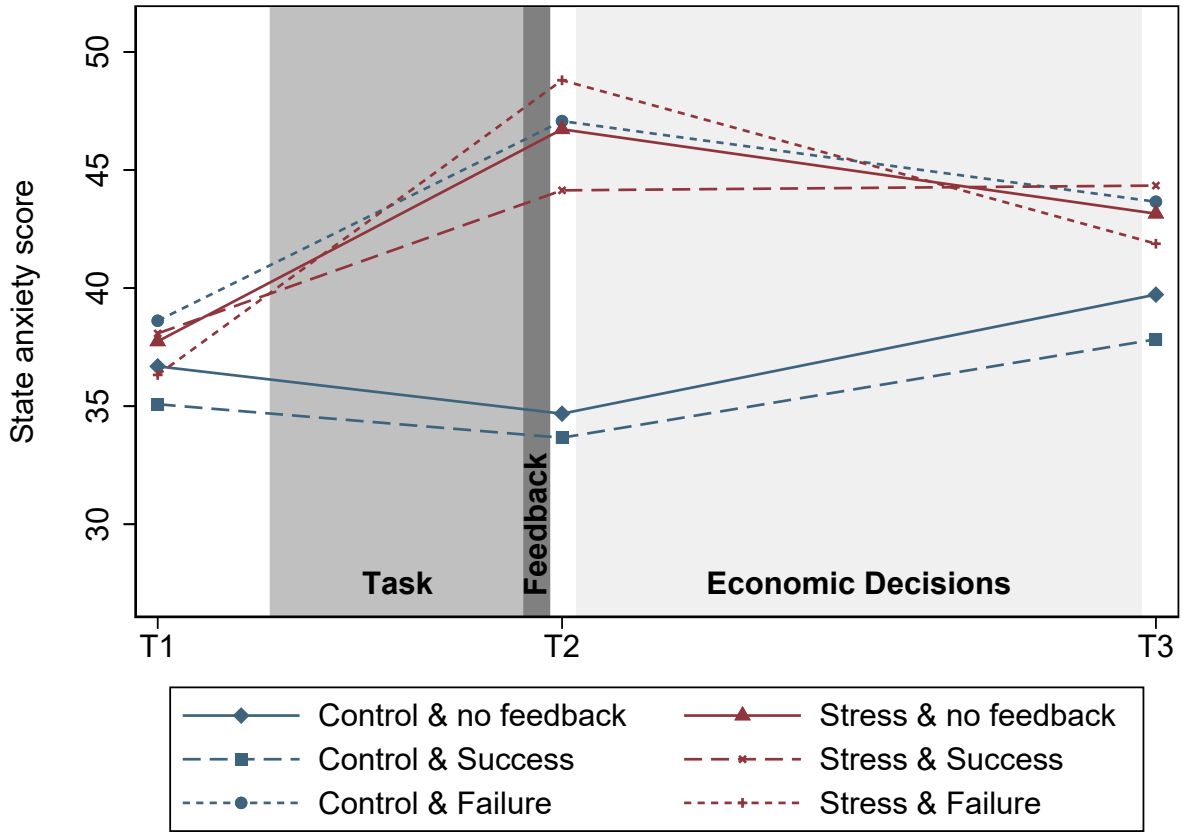
4 Conclusion

In this study, we develop an experimental protocol to identify the individual and combined effects of acute stress and failure on decision-making. The protocol uses a two-session design to vary participants' exposure to stress, failure, and success. Exogenous variation in acute stress levels is introduced by assigning an incentivized cognitive task designed to induce mild stress. Half of the participants complete the stress task in the first session and half complete it in the second session. Those in a session not assigned to the stress task complete a stress-neutral control task during the same time period.

The experiment introduces variation in participants' perceptions of success or failure via a feedback protocol in the second session. The incentive for the stress task deducts a portion of the performance pay-off for participants whose performance falls below an undisclosed threshold. The threshold level and the provision of feedback are randomly assigned. Participants receive either no feedback, feedback with a low threshold level (success condition), or feedback with a high threshold level (failure condition).

We validate the protocol using an online experiment with a sample of 269 university students and find that participants perceived the incentivized stress task as significantly more stressful than the control task. State anxiety also increased substantially and significantly after com-

Figure 2: State anxiety response to the stress / control task and the feedback



Note: Means were calculated for the state anxiety scores at the three measurement points. Participants with a performance level above the high or below the low threshold were excluded from the sample for simplicity.

pleting the stress task. These results provide strong evidence that our stress protocol induced short-term psychological stress.

Participants who received failure feedback reported significantly lower self-assessments of their performance and success in the cognitive stress task, significantly greater anxiety, and strong emotional responses. Responses from participants who received success feedback did not differ substantially from responses of participants who received no feedback. Thus, we find strong evidence that our provision of negative feedback negatively affected participants' perceptions of themselves and evoked negative emotions.

The ability of our protocol to induce short-term stress and feelings of failure among participants in online and laboratory experiments will allow future studies of the impacts of those two dimensions on decision-making. Our findings indicate that receiving success feedback does not alter individuals' assessments of their performance or how successful they feel. Therefore, it would be more efficient to concentrate future studies solely on a comparison between failure feedback and no feedback.

Stress and failure likely affect decision-making differentially in a variety of settings. At school and at work, stress could improve individuals' focus on a task, increase their cognitive load, and decrease their attention on other tasks. Such stress could be addressed by adapting the workloads and providing coping strategies. Failure, on the other hand, likely affects self-worth and self-esteem, decreases confidence, and induces compensatory behaviours. Redesigning goal

settings and feedback structures could improve individuals' reactions to a poor performance. Future studies could use the protocol in this study to examine the relative importance of stress and failure for various types of decisions, thereby uncovering effective interventions.

References

Allen, T.D. and Armstrong, J., 2006. Further Examination of the Link Between Work-Family Conflict and Physical Health, *American Behavioral Scientist*, 49 (9), 1204–1221.

Bendahan, S., Goette, L., Thoresen, J., Loued-Khenissi, L., Hollis, F., and Sandi, C., 2017. Acute stress alters individual risk taking in a time-dependent manner and leads to anti-social risk, *European Journal of Neuroscience*, 45 (7), 877–885.

Bohns, V.K. and Flynn, F.J., 2013. Guilt by Design: Structuring Organizations to Elicit Guilt as an Affective Reaction to Failure, *Organization Science*, 24 (4), 1157–1173.

Buckert, M., Schwieren, C., Kudielka, B.M., and Fiebach, C.J., 2017. How stressful are economic competitions in the lab? an investigation with physiological measures, *Journal of Economic Psychology*, 62, 231–245.

Buser, T., 2016. The impact of losing in a competition on the willingness to seek further challenges, *Management Science*, 62 (12), 3439–3449.

Buser, T. and Dreber, A., 2016. The flipside of comparative payment schemes, *Management Science*, 62 (9), 2626–2638.

Cahlíková, J. and Cingl, L., 2017. Risk preferences under acute stress, *Experimental Economics*, 20 (1), 209–236.

Carver, C.S. and Scheier, M.F., 1990. Origins and Functions of Positive and Negative Affect: A Control-Process View, *Psychological Review*, 97 (1), 19–35.

Cassar, L. and Klein, A.H., 2019. A Matter of Perspective: How Failure Shapes Distributive Preferences, *Management Science*, 65 (11), 5050–5064.

Certo, S.T., Busenbark, J.R., Woo, H.s., and Semadeni, M., 2016. Sample selection bias and Heckman models in strategic management research, *Strategic Management Journal*, 37 (13), 2639–2657.

Cohen, S., 1980. Aftereffects of stress on human performance and social behavior: A review of research and theory, *Psychological Bulletin*, 88 (1), 82–108.

Corak, M., 2013. Income Inequality, Equality of Opportunity, and Intergenerational Mobility, *Journals of Economic Perspectives*, 27 (3), 79–102.

Crocker, J. and Park, L.E., 2004. The costly pursuit of self-esteem, *Psychological Bulletin*, 130 (3), 392–414.

Crocker, J. and Wolfe, C.T., 2001. Contingencies of self-worth, *Psychological Review*, 108 (3), 593–623.

Daviu, N., Bruchas, M.R., Moghaddam, B., Sandi, C., and Beyeler, A., 2019. Neurobiological links between stress and anxiety, *Neurobiology of Stress*, 11.

Delaney, L., Fink, G., and Harmon, C., 2014. Effects of stress on economic decision-making: Evidence from laboratory experiments, *IZA Discussion Paper*, 8060.

Dickerson, S.S. and Kemeny, M.E., 2004. Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research, *Psychological Bulletin*, 130 (3), 355–391.

Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P.G., Marias, K., and Tsiknakis, M., 2017. Stress and anxiety detection using facial cues from videos, *Biomedical Signal Processing and Control*, 31, 89–101.

Gill, D. and Prowse, V., 2014. Gender differences and dynamics in competition: The role of luck, *Quantitative Economics*, 5 (2), 351–376.

Habhab, S., Sheldon, J.P., and Loeb, R.C., 2009. The relationship between stress, dietary restraint, and food preferences in women, *Appetite*, 52 (2), 437–444.

Haushofer, J., Jang, C., Lynham, J., and Abraham, J., 2018. Stress and temporal discounting: Do domains matter?, *Mimeo*.

Heatherton, T.F. and Polivy, J., 1991. Development and Validation of a Scale for Measuring State Self-Esteem, *Journal of Personality and Social Psychology*, 60 (6), 895–910.

Heckman, J.J., 1979. Sample Selection Bias as a Specification Error, *Econometrica*, 47 (1), 153–161.

International Cognitive Ability Resource Team, 2014. The International Cognitive Ability Resource.

Kirschbaum, C., Pirke, K.M., and Hellhammer, D.H., 1993. The 'Trier Social Stress Test' - A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting, *Neuropsychobiology*, 28, 76–81.

Kudielka, B.M., Buske-Kirschbaum, A., Hellhammer, D.H., and Kirschbaum, C., 2004. HPA axis responses to laboratory psychosocial stress in healthy elderly adults, younger adults, and children: Impact of age and gender, *Psychoneuroendocrinology*, 29 (1), 83–98.

List, J.A., Shaikh, A.M., and Xu, Y., 2019. Multiple hypothesis testing in experimental economics, *Experimental Economics*, 22 (4), 773–793.

Marteau, T.M. and Bekker, H., 1992. The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI), *British Journal of Clinical Psychology*, 31 (3), 301–306.

Rutters, F., Nieuwenhuizen, A.G., Lemmens, S.G., Born, J.M., and Westerterp-Plantenga, M.S., 2009. Acute stress-related changes in eating in the absence of hunger, *Obesity*, 17 (1), 72–77.

Shields, G.S., Sazma, M.A., and Yonelinas, A.P., 2016. The effects of acute stress on core executive functions: A meta-analysis and comparison with cortisol, *Neuroscience & Biobehavioral Reviews*, 68, 651–668.

Šidák, Z., 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions, *Journal of the American Statistical Association*, 62 (318), 626–633.

Starcke, K., Wiesen, C., Trotzke, P., and Brand, M., 2016. Effects of acute laboratory stress on executive functions, *Frontiers in Psychology*, 7, 461.

Stevens, A.H., 1999. Climbing out of Poverty, Falling Back in: Measuring the Persistence of Poverty Over Multiple Spells, *The Journal of Human Resources*, 34 (3), 557–588.

Sunde, U. and Dohmen, T., 2016. Aging and preferences, *The Journal of the Economics of Ageing*, 7, 64–68.

Thaler, R.H. and Johnson, E.J., 1990. Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice, *Management Science*, 36 (6), 643–660.

Vitt, N., James, J., Belot, M., and Vecchi, M., 2021. Daily stressors and food choices: A lab experiment with low-SES mothers, *European Economic Review*, 136, 103754.

Von Dawans, B., Fischbacher, U., Kirschbaum, C., Fehr, E., and Heinrichs, M., 2012. The Social Dimension of Stress Reactivity: Acute Stress Increases Prosocial Behavior in Humans, *Psychological Science*, 23 (6), 651–660.

Von Dawans, B., Kirschbaum, C., and Heinrichs, M., 2011. The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format, *Psychoneuroendocrinology*, 36 (4), 514–522.

Wölbert, E. and Riedl, A., 2013. Measuring time and risk preferences: Reliability, stability, domain specificity, *CESifo Working Paper*, 4339.

Wooldridge, J.M., 2010. *Econometric analysis of cross section and panel data.*, Cambridge, MA: MIT Press, 2nd ed.

Electronic Supplementary Material

Appendix A Full timeline of the experimental sessions

Table A.1: Timeline of the experimental sessions

(a) First session

Experimental conditions:	
S1: Stress only	**S2: Control**
General instructions & preview of experiment stages	
Example of incentive calculation	
1st State anxiety measure	
Stress task	Control task
Short wait	
2nd State anxiety measure	
Outcome tasks	
3rd State anxiety measure	
Task perceptions	

(b) Second session

Experimental conditions:			
S1-N: Control	**S1-S: Success only** **S1-F: Failure only**	**S2-N: Stress only**	**S2-S: Stress & Success** **S2-F: Stress & Failure**
General instructions & preview of experiment stages			
Example of incentive calculation			
1st State anxiety measure			
Control task		Stress task	
Short wait	Feedback	Short wait	Feedback
2nd State anxiety measure			
Outcome tasks			
3rd State anxiety measure			
Task perceptions			
Performance expectations			
Questionnaire on control variables, real-life stress, failure, and decision-making			

Appendix B Experimental instructions

The experimental instructions below were shown to participants on screen. Instructions differed between the first and the second session, and between the different experimental conditions. Any parts of the instructions that are specific to a session or experimental condition are noted as such below.

B.1 Introduction - First session

a) General instructions

Economic Decision Making – An Online Experiment

Laboratory for Economics, Management and Auctions (LEMA), Smeal College of Business, The Pennsylvania State University

Decision Study – First Survey

Experimental Instructions

Welcome and thank you for participating in this experiment.

The study you are participating in consists of two surveys, so we ask you to complete a second survey in approximately one week. For participating in the two surveys, you will receive a guaranteed $5 after the second survey. Please note that you will receive your payment through Paypal, and you will not be paid if you do not complete the second survey. You can earn additional money based on the tasks you complete and the decisions you make during this survey and the second survey. Specifically, you will be asked to complete several tasks and decisions. The computer will randomly choose one task or decision during the two surveys to be paid to you.

Unless otherwise specified, this additional payment (of up to $40) will be made **after the second survey.** Again, please note that you will not be paid if you do not complete the second survey.

All the decisions you make and the information we collect will remain anonymous and will not be shared with other participants in the study.

This is an online experiment and all decisions, tasks and surveys will be completed using the web browser on your computer. **Please do not close the browser window at any point throughout the survey, as you will not be able to return to the survey and you therefore would not receive any payment.**

[*New page*]
For some of the tasks, you may want to use a calculator or pen & paper. Please make sure you have these available before proceeding.

This survey will last approximately 60 minutes. Please only start this survey if you have sufficient time to complete it in whole. It is not possible to pause, stop, resume or retake the survey. There are time limits for each stage of the experiment. A timer will indicate how much time remains for each stage.

All further instructions throughout the experiment will be displayed on your computer screen. Please make sure you read them carefully.

Note that you are free to drop out of the study at any point. However, we are only able to compensate participants who complete both surveys.

Your participation in this study is voluntary, and you are free to skip any tasks or questions if you wish so. Skipping or not completing a task can however affect your payment at the end of the experiment.

b) Preview of the session

What will happen during this session:

1. [*Stress session:* You will be asked to complete a **set of computer-based tasks**. You will have 10 minutes to complete as many tasks as possible. You might be able to earn an additional fee based on your performance in this task. There is a probability of 50% that this task will be chosen for your payment.]
   [*Control session:* You will have 5 minutes to **answer a few questions about yourself**. Afterwards we will ask you to **read some short texts and answer questions** about the texts.]

2. We will ask you to **make a range of decisions and complete some tasks**. You may be paid additional money based on of your decisions and performance in these tasks. There is a probability of 25% that one of these tasks or decision will be chosen for your payment.

3. We will ask you a few questions about your experience of today's study.

From time to time throughout the survey we will furthermore ask you short questions about how you feel.

Please do not close the browser window on your computer at any point during this survey.

c) Incentive explanations and example

For participating in the two surveys, you will receive $5 after the second survey. Please note that you will receive your payment via Paypal. We will send you the payment at the email address you provided us at sign-up, and you will not be paid if you do not complete the second survey.

You can earn additional money based on the tasks you complete and decisions you make during this survey and the second survey. Specifically, you will be asked to complete several tasks and decisions in which you can earn tokens. The computer will randomly choose one task or decision during the two surveys to be paid to you. The tokens you earned in the randomly selected task or decision will be converted to dollars for your payment: **125 tokens are worth**

1$. Unless otherwise specified, this additional payment (of up to 5000 tokens = $40) will be made **after the second survey**.

[*Stress session:* With a 50% probability, the first task you will face today will be selected for payment. With a 25% probability, one of the other decisions or tasks you will face today will be selected for payment. Finally, with a 25% probability, one of the decisions or tasks you will face in the second survey will be selected for payment.]

[*Control session:* With a 25% probability, one of the decisions or tasks you will face today will be selected for payment. With a 75% probability, one of the decisions or tasks you will face in the second survey will be selected for payment.]

Please check you have understood these instructions by answering the questions below about the following example.

Example:

A study participant has earned the following potential payments throughout the two surveys:

- Survey 1 – Task 1: 3750 tokens = $30
- Survey 1 – Task 2: 0 tokens = $0
- Survey 2 – Task 1: 2500 tokens = $20
- Survey 2 – Task 2: 1250 tokens = $10

The computer randomly chooses ONE of these tasks to be paid after the second survey.

If the computer chooses Task 1 in Survey 1 for payment, how much is paid to the participant?

- $5 for participation and $0 additionally
- $5 for participation and $10 additionally
- $5 for participation and $20 additionally
- $5 for participation and $30 additionally
- $5 for participation and $40 additionally

If the computer chooses Task 2 in Survey 2 for payment, how much is paid to the participant?

- $5 for participation and $0 additionally
- $5 for participation and $10 additionally
- $5 for participation and $20 additionally
- $5 for participation and $30 additionally
- $5 for participation and $40 additionally

[*A "check your answers" button is displayed for participants to check whether their answers are correct. Participants can only proceed once they have given the correct answers.*]

d) Start of timed part

The timed part of the survey will begin on the next page. It will last approximately 45 minutes. It is not possible to pause the survey once you proceed.

e) First state anxiety measure

Before we start with the first stage of the survey, we will ask you a few questions about how you feel.

B.2 Introduction - Second session

Economic Decision Making – An Online Experiment

Laboratory for Economics, Management and Auctions (LEMA), Smeal College of Business, The Pennsylvania State University

Decision Study – Second Survey

Experimental Instructions

Welcome back, and thank you for participating in the second part of this experiment.

For participating in the study, you will receive $5 after completing today's survey. Please note that you will receive your payment via Paypal. We will send the payment to the email address you provided to us during sign-up. You can earn additional money based on the tasks you complete and decisions you make during the two surveys. Specifically, the computer will randomly choose one task or decision during the two surveys to be paid to you.

Unless otherwise specified, this additional payment (of up to $40) will be made after today's survey.

All the decisions you make and the information we collect will remain anonymous and will not be shared with other participants in the study.

This is an online experiment and all decisions, tasks and surveys will be completed using the web browser on your computer. **Please do not close the browser window at any point throughout the survey, as you will not be able to return to the survey and you therefore would not receive any payment.**

[*New page*]

For some of the tasks, you may want to use a calculator or pen & paper. Please make sure you have these available before proceeding.

This survey will last approximately 75 minutes. Please only start this survey if you have sufficient time to complete it in whole. It is not possible to pause, stop, resume or retake the survey. There are time limits for each stage of the experiment. A timer will indicate how much time remains for each stage.

All further instructions throughout the experiment will be displayed on your computer screen. Please make sure you read them carefully.

Note that you are free to drop out of the study at any point. However, we are only able to compensate participants who complete the survey.

Your participation in this study is voluntary, and you are free to skip any tasks or questions if you wish so. Skipping or not completing a task can however affect your payment at the end of the experiment.

a) Preview of the session
What will happen during this session:

1. [*Stress session:* You will be asked to complete a **set of computer-based tasks**. You will have 10 minutes to complete as many tasks as possible. You might be able to earn an additional fee based on your performance in this task. There is a probability of 50% that this task will be chosen for your payment.]
   [*Control session:* You will have 5 minutes to **answer a few questions about yourself**. Afterwards we will ask you to **read some short texts and answer questions** about the texts.]

2. We will ask you to **make a range of decisions and complete some tasks**. You may be paid additional money based on of your decisions and performance in these tasks. There is a probability of 25% that one of these tasks or decision will be chosen for your payment.

3. We will then ask you to complete a short survey about your experience of today's study, about yourself and your lifestyle.

4. **You will be paid after you have completed the survey.** The payment will be made via Paypal and will be sent to the email address you provided to us during sign-up.

From time to time throughout the survey we will furthermore ask you short questions about how you feel.

Please do not close the browser window on your computer at any point during this survey.

b) Incentive explanations and example
For participating in the study, you will receive $5 after completing today's survey. Please note that you will receive your payment via Paypal. We will send you the payment at the email address you provided to us at sign-up.

You can earn additional money based on the tasks you complete and decisions you make during this survey and the first survey you completed. Specifically, you will be asked to complete several tasks and decisions in which you can earn tokens. The computer will randomly choose one task or decision during the two surveys to be paid to you. The tokens you earned in the randomly selected task or decision will be converted to dollars for your payment: **125 tokens are worth 1$**. Unless otherwise specified, this additional payment (of up to 5000 tokens = $40) will be made **after today's survey**.

[*Stress session:* With a 50% probability, the first task you will face today will be selected for payment. With a 25% probability, one of the other decisions or tasks you will face today will be selected for payment. Finally, with a 25% probability, one of the decisions or tasks you faced in the first survey will be selected for payment.]

[*Control session:* With a 25% probability, one of the decisions or tasks you will face today will be selected for payment. With a 75% probability, one of the decisions or tasks you faced in the first survey will be selected for payment.]

Please check you have understood these instructions by answering the questions below about the following example.

Example:

A study participant has earned the following potential payments throughout the two surveys:

- Survey 1 – Task 1: 3750 tokens = $30
- Survey 1 – Task 2: 0 tokens = $0
- Survey 2 – Task 1: 2500 tokens = $20
- Survey 2 – Task 2: 1250 tokens = $10

The computer randomly chooses ONE of these tasks to be paid after the second survey.

If the computer chooses Task 1 in Survey 1 for payment, how much is paid to the participant?

- $5 for participation and $0 additionally
- $5 for participation and $10 additionally
- $5 for participation and $20 additionally
- $5 for participation and $30 additionally
- $5 for participation and $40 additionally

If the computer chooses Task 2 in Survey 2 for payment, how much is paid to the participant?

- $5 for participation and $0 additionally
- $5 for participation and $10 additionally
- $5 for participation and $20 additionally
- $5 for participation and $30 additionally
- $5 for participation and $40 additionally

[*A "check your answers" button is displayed for participants to check whether their answers are correct. Participants can only proceed once they have given the correct answers.*]

c) Start of timed part

The timed part of the survey will begin on the next page. It will last approximately 55 minutes. It is not possible to pause the survey once you proceed.

d) First state anxiety measure

Before we start with the first stage of the survey, we will ask you a few questions about how you feel.

B.3 Stress task

Assessment-style tasks

In the following, we ask you to complete a block of several tasks that are similar to test questions you may face in assessment tests when applying for jobs or for graduate school. There are 18 tasks in total and you have 10 minutes to complete as many tasks as you can. You can take a maximum of 90 seconds to complete a single task. A countdown timer at the top of the screen will indicate how much time you have left for the current task.

After the 90 seconds have passed, your current answer will be submitted and the next task will appear automatically. If you complete a task before the 90 seconds have passed, you can submit your answer and continue with the next task by clicking the red button at the bottom of the page.

You have been allocated 5000 tokens ($40) for this task. Your performance in these tasks determines how many tokens you lose. Providing a wrong answer or taking too long to answer will cause a deduction from the 5000 tokens, according to the rules below. There is a 50% probability that this task will be chosen for your payment, in which case the remainder of the 5000 tokens will be converted to dollars and paid to you after the second survey.

Each of the 18 assessment-style questions has only one correct answer. Your performance in each of the 18 tasks can cost you up to 225 tokens:

- If you submit a correct answer to a task within 60 seconds of starting the task, there is no deduction.
- If you submit a correct answer in the last 30 seconds of a task, 60 tokens are deducted.
- If you submit a wrong answer, 225 tokens are deducted.
- If no answer is submitted within 90 seconds, 225 tokens are deducted.

In addition to these tasks, simple knowledge and math questions will appear on screen over the course of the 10 minutes. Please answer these questions by clicking on the correct answer, the pop-up will then close automatically. Each wrong answer in these pop-up questions will result in a deduction of 20 tokens from your potential pay-off. Answer as fast as you can, as the timer for the current task will continue to run down while you answer the pop-up questions.

In addition to the above deductions for your performance in each task and each pop-up question, 750 tokens will be deducted from your potential pay-off if your overall performance in this task is below a threshold. You will be randomly assigned by the computer to a group of participants, which all face the same threshold.

Please do not close the browser window at any point as you will not be able to return to the survey and participate in the study.

At this point, please wait for the task to start. The task will start automatically once the current timer has elapsed.

Reminder: You can earn up to $45 if you complete both surveys of this study. If you start one of the surveys more than once, you will not be paid.

B.4 Control task

a) Instructions for first part of control task

Survey

We now ask you to complete a short questionnaire on you and your lifestyle. You have 4 minutes to complete the questions.

b) Instructions for second part of control task

Reading and Questions

We now ask you to **read two short texts and answer questions** about the texts. You have **10 minutes in total** to read the texts and answer the questions. You should have ample time to read each text and answer each question, you can take your time. After the 10 minutes have passed, your answers will be submitted automatically. You can go back and forth between different texts by using the arrows at the bottom of the page. The answers to all questions can be found in the texts.

B.5 Feedback

a) Message shown in the first session and in the "no feedback" condition of the second session

You have now completed the first stage of this survey. The survey will continue shortly. The next stage will start automatically.

Reminder: You can earn up to $45 if you complete both surveys of this study. If you start one of the surveys more than once, you will not be paid.

b) Success feedback

You succeeded to achieve the threshold

[*Stress session:* In the previous stage you were asked to complete a **task similar to an assessment test.**]

[*Control session:* At the beginning of the first survey, you were asked to complete a **task similar to an assessment test.**]

In this task **you succeeded to achieve the threshold for your group.** If this task is selected for payment, the additional 750 tokens will not be deducted from your pay-off. If any of the other tasks is selected for payment, no deduction from your pay-off will be made.

The next stage of the survey will start automatically.

Reminder: You can earn up to $45 if you complete both surveys of this study. If you start one of the surveys more than once, you will not be paid.

c) Failure feedback

You failed to achieve the threshold

[*Stress session:* At the beginning of today's survey, you were asked to complete a **task similar to an assessment test.**]

[*Control session:* At the beginning of the first survey, you were asked to complete a **task similar to an assessment test.**]

In this task **you failed to achieve the threshold for your group.** If this task is selected for payment, the additional 750 tokens will be deducted from your pay-off. If any of the other tasks is selected for payment, no deduction from your pay-off will be made.

The next stage of the survey will start automatically.

Reminder: You can earn up to $45 if you complete both surveys of this study. If you start one of the surveys more than once, you will not be paid.

B.6 Economic decision-making tasks

a) Second state anxiety measure

Before we continue with the second stage of the survey, we will ask you a few questions about how you feel.

b) General instructions for economic decision-making tasks

Decisions and Tasks

We now ask you to make a range of decisions and complete some tasks. There is a 25% probability that one of the decisions or tasks will be chosen for your payment. Tasks 1 and 3 have a 5% probability of being chosen. Tasks 2 and 4 have a probability of 7.5% each of being chosen.

[*First session & stress session:* If this stage will not be chosen for payment, either the first task you faced today will be paid (50% probability), or one of the decisions or tasks you will face in the second survey will be paid (25% probability).]

[*First session & control session:* If this stage will not be chosen for payment, one of the decisions or tasks you will face in the second survey will be paid (75% probability).]

[*Second session & stress session:* If this stage will not be chosen for payment, either the first task you faced today will be paid (50% probability), or one of the decisions or tasks you faced in the first survey will be paid (25% probability).]

[*Second session & control session:* If this stage will not be chosen for payment, one of the decisions or tasks you faced in the first survey will be paid (75% probability).]

Prior to each task or block of decisions, instructions will be shown on screen. Please make sure you read these instructions thoroughly.

Please wait for the first task to start. It will start automatically once the current timer has elapsed.

Reminder: You can earn up to $45 if you complete both surveys of this study. If you start one of the surveys more than once, you will not be paid.

c) Altruism task

Task 1 - Instructions:

You have 5 minutes in total to read these instructions and to complete the task (4 questions).

For the following part, we ask you 4 questions about how you would like to allocate between you and a charitable organization. You could give all or none of the tokens to the charitable organization, or any amount in between.

One of these 4 questions will be randomly chosen by the computer, and if this task is selected for payment the will be paid to you and the charitable organization according to your choice. Any of the 4 questions could be chosen by the computer, so you should think carefully about each question.

Your decisions in this task will remain anonymous and will not be shared with other participants in the study.

Before beginning with the task, please check you have understood these instructions by answering the questions below about the following example. Once you have checked your answer using the blue button below, you will be able to proceed to the task.

Example:

Suppose a player selected to give 0 tokens to charity A, 1000 tokens to charity B, 1500 tokens to charity C and 2500 tokens to charity D. If this task is selected for the player's payment and charity C is randomly selected,

- How many tokens will be paid to the player? 0 / 500 / 1500 / 2000 / 2500
- How many tokens will be paid to charity A? 0 / 500 / 1500 / 2000 / 2500
- How many tokens will be paid to charity B? 0 / 500 / 1500 / 2000 / 2500
- How many tokens will be paid to charity C? 0 / 500 / 1500 / 2000 / 2500
- How many tokens will be paid to charity D? 0 / 500 / 1500 / 2000 / 2500

[*A "check your answers" button is displayed for participants to check whether their answer to the question is correct. The correct answer is displayed once the button is clicked and participants can proceed after clicking the button irrespective of whether they have given a correct or incorrect answer.*]

d) Risk taking task

Task 2 - Instructions:

You have 5 minutes in total to read these instructions and to complete the task (15 decisions).

For the following task you will be asked to make 15 decisions between two alternatives:

- Option A: getting a payment for sure.

  OR

- Option B: a lottery, in which with a 50% probability you can receive 3200 tokens, and with a 50% probability you can receive 0 tokens.

The decisions only differ in the amount of tokens paid for sure when option A is chosen.

One of these 15 decisions has been chosen in advance by the computer, and you will be paid based on this decision if this task is selected for payment: If you have chosen the sure payment (option A), you get it. If you have chosen the lottery (option B), the lottery will be played by the computer and you get paid based on the outcome of the lottery.

Any of the 15 decisions could be the one chosen by the computer, so you should think carefully about each decision you make.

Before beginning with the task, please check you have understood these instructions by answering the questions below about the following example choice. Once you have checked your answer using the blue button below, you will be able to proceed to the task.

Example choice:

Which of the following would you prefer?

A) receive 1000 for sure

B) play this lottery: 50% chance of 3200 tokens – 50% chance of 0 tokens

Suppose this decision was randomly chosen by the computer to be paid.

- If you choose A, what do you receive?
  0 tokens / 1000 tokens / 1600 tokens / 3200 tokens
- If you choose B and you WIN the lottery, what do you receive?
  0 tokens / 1000 tokens / 1600 tokens / 3200 tokens
- If you choose B and you LOSE the lottery, what do you receive?
  0 tokens / 1000 tokens / 1600 tokens / 3200 tokens

[*A "check your answers" button is displayed for participants to check whether their answer to the question is correct. The correct answer is displayed once the button is clicked and participants can proceed after clicking the button irrespective of whether they have given a correct or incorrect answer.*]

e) Intertemporal substitution task

Task 3 - Instructions:

You have 6 minutes in total to read these instructions and to complete the task (30 decisions).

For the following task you will be asked to make 30 decisions between two alternatives:

- Option A: receiving a payment at a certain point in time

  OR

- Option B: receiving a larger payment at a later point in time

The payment amounts and timing of the payments will vary between the decisions.

One of the 30 decisions has been chosen in advance by the computer and you will be paid based on this decision, if this task is selected for payment. The payment will be made using

Paypal and we will email it to you at the time specified in the choice. Any of the 30 decisions could be the one chosen by the computer, so you should think carefully about each decision you make.

Before beginning with the task, please check you have understood these instructions by answering the questions below about the following example choice. Once you have checked your answer using the blue button below, you will be able to proceed to the task.

Example choice:

Which of the following would you prefer?

A) 1000 tokens TODAY

B) 2000 tokens in 3 MONTHS

Suppose this decision was randomly chosen by the computer to be paid.

- If you choose A, what do you receive?
  1000 tokens today / 2000 tokens today / 1000 tokens in 3 months / 2000 tokens in 3 months / No payment
- If you choose B, what do you receive?
  1000 tokens today / 2000 tokens today / 1000 tokens in 3 months / 2000 tokens in 3 months / No payment

[*A "check your answers" button is displayed for participants to check whether their answer to the question is correct. The correct answer is displayed once the button is clicked and participants can proceed after clicking the button irrespective of whether they have given a correct or incorrect answer.*]

f) Effort task

Task 4 - Instructions:

You have 2 minutes to read these instructions.

The following task will consist of a screen with 100 sliders. Below these instructions you can see an example of what these sliders look like. Each slider is initially positioned at 0 and can be moved as far as 100. Each slider has a number to its right showing its current position. You can use the mouse in any way you like to move each slider. You can readjust the position of each slider as many times as you wish.

You will have 120 seconds to move as many of the sliders to exactly 50 as you can. For each slider positioned at exactly 50 at the end of the 120 seconds, you will be paid 50 tokens if this task is chosen for payment.

You now have some time to make yourself familiar with the example sliders shown below. The task will start automatically once the current timer has elapsed.

B.7 End of session - First session

a) Third state anxiety measure and task perceptions

Before concluding this survey, we will now ask you a few questions about how you feel and about your experience during today's survey.

b) End

End of first survey

This is the final stage of the first survey. We will send you an email with the link for the second survey in 7 days. Please make sure you complete the second survey as well.

[*No payment at end of first session:* You will receive payments for the tasks and decisions you made or will make during the study after completing the second survey. Therefore, please make sure you complete the second survey once you receive the link by email.]

[*Payment at end of first session:* You will receive a payment of $ XX.XX in the next 48 hours via Paypal. We will email the payment to the address you provided to us at sign-up. You will receive any other payments at the end of the second survey. Therefore, please make sure you complete the second survey once you receive the link by email.]

You may close this browser window now.

Thank you for participating in our study

B.8 End of session - Second session

a) Third state anxiety measure, task perceptions, and questionnaire on control variables, real-life stress, failure, and decision-making

In this stage, we will ask you a few questions about how you feel and about your experience during the study. We will also ask you questions about yourself and your lifestyle. Please ensure that you have answered all questions before proceeding to the next stage. Once you have completed the questionnaire, please DO NOT CLOSE the browser window.

b) End

End of study

This is the final stage of the study.

You will receive a payment of $ XX.XX in the next 48 hours via Paypal. We will email the payment to the address you provided to us at sign-up. If you do not have a Paypal account, you have 30 days to open an account and accept the payment.

[*Delayed payment:* Additionally, you will receive a payment of $ XX.XX within 48 hours of DELAYED PAYMENT DATE, again via Paypal to the email address provided.]

[*Altruism task chosen for payment:* Additionally, a donation of $ XX.XX will be made to CHARITY NAME based on your choice.]

You may close this browser window now.

Thank you for participating in our study

Appendix C Experimental tasks

C.1 Stress task

After initial instructions, participants had 10 minutes to complete a block of 18 short incentivized cognitive tasks. 77.32 % of participants did not complete all 18 tasks, suggesting that the time limit was tight and induced time pressure. There were also individual time limits of 90 seconds for each of the 18 cognitive tasks, indicated by a countdown timer at the top of the screen. After 55 seconds, the timer changed colour from grey to red to indicate that time was running out and that a first pay-off deduction (after 60 seconds – details below) was imminent. If no answer had been submitted after 90 seconds, the selected answer was submitted automatically and the following task started.

The cognitive tasks were chosen to mimic questions encountered in graduate assessment tests or intelligence tests and comprised three types of tasks: matrix reasoning tasks, mental rotation tasks, and quantitative reasoning tasks. Tasks increased in difficulty throughout the 10 minutes, to counteract any learning effects. The matrix reasoning tasks, designed by the International Cognitive Ability Resource Team (2014), are similar to Raven's Progressive Matrices. Participants were shown a 3x3 matrix of geometric forms with one of the nine forms missing. They were asked to identify which form among six possible choices was the missing one.

In the mental rotation tasks, also designed by the International Cognitive Ability Resource Team (2014), participants were shown images of three-dimensional cubes and asked to identify which cube among six possible choices could be a rotation of a given target cube.

The quantitative reasoning tasks were similar to those used in graduate admission tests and were designed by the authors. Participants were asked to identify the correct solution to these quantitative questions from six possible choices. They were asked at the beginning of the experiment sessions to ensure that they had a calculator as well as pen and paper available.

To create additional stress through distractions, 10 simple knowledge and arithmetic questions (e.g., "What is the capital of the US?") appeared as pop-ups throughout the block of cognitive tasks. The pop-up questions were programmed to be displayed at predetermined times within the 10 minutes. When a pop-up question was shown, participants were unable to continue their work on the main task until they answered the pop-up question. Meanwhile, the countdown timer for the current cognitive task remained visible and continued to run down.

To check that our protocol successfully induced stress, we collected measures of perceived stressfulness at the end of each session (see e.g., Kudielka *et al.*, 2004; Von Dawans *et al.*, 2011; Vitt *et al.*, 2021). We also collected anxiety levels using the six-item short-form state anxiety inventory developed by Marteau and Bekker (1992) at three points during each session (see e.g., Rutters *et al.*, 2009; Von Dawans *et al.*, 2011; Starcke *et al.*, 2016): prior to starting the stress or control task (T1), after completing the stress or control task (T2) and after completing the economic decision-making tasks (T3). and state anxiety (see e.g., Rutters *et al.*, 2009; Von Dawans *et al.*, 2011; Starcke *et al.*, 2016). Details on the measures of stress are in the Online Appendix D.1.

Participants were initially allocated a maximum potential incentive of 5000 tokens ($40) for

the stress task, with a conversion rate of 125 tokens for $1.Participants' performance in the cognitive tasks and the pop-up knowledge questions determined how much of the initial 5000 tokens they "lost". The incentives are framed as "losses" rather than "gains", to avoid inducing positive sentiments.

In each of the 18 cognitive tasks, participants could lose up to 225 tokens ($1.80). There was no deduction if the correct answer was given within 60 seconds of the start of a task. If a correct answer was given after more than 60 seconds, 60 tokens ($0.48) were deducted. If an incorrect answer or no answer was given, a total of 225 tokens ($1.80) was deducted. This also applied to any tasks that were not yet completed when the overall time limit elapsed. Each of the 10 pop-up knowledge and arithmetic questions was worth 20 tokens ($0.16). No deduction was made when a correct answer is given, 20 tokens ($0.16) are deducted otherwise.

In addition to these deductions, participants faced a potential deduction of 750 tokens ($6.00) if their overall performance in this task was below an unknown threshold. The level of this threshold was experimentally varied by Treatment 2 to induce failure or success. Participants were informed that they would be randomly assigned to a group of participants who all faced the same threshold.

In total, participants could lose up to 4050 tokens ($32.40) based on their performance in the 18 cognitive tasks, 200 tokens ($1.60) based on their answers to the pop-up questions, and 750 tokens ($6.00) from the additional deduction. This incentive structure was designed to ensure that participants' performance in each task and question affected their potential pay-off, hence reducing the risk of participants giving up if they struggle to solve some of the tasks.

Figure C.1: Stress task - Screenshot of a matrix reasoning task (sample)

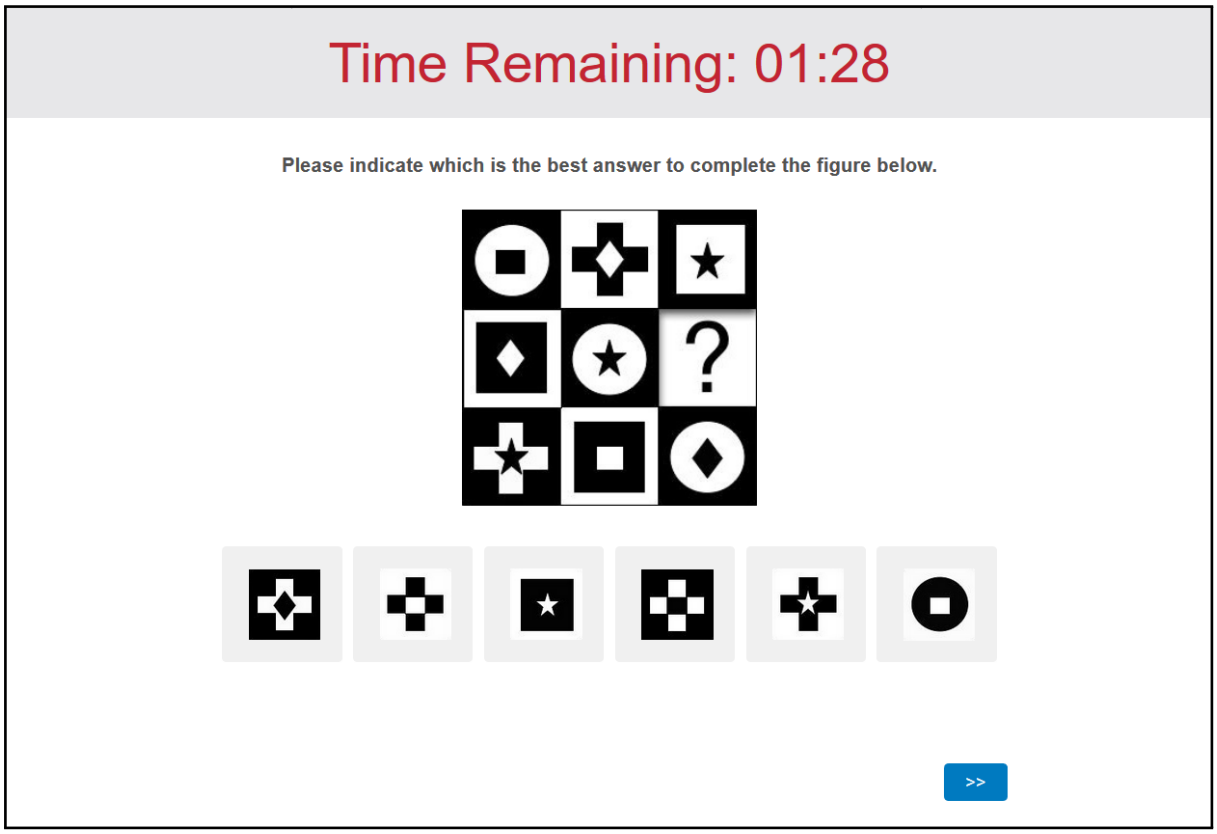


Figure C.2: Stress task - Screenshot of a mental rotation task (sample)

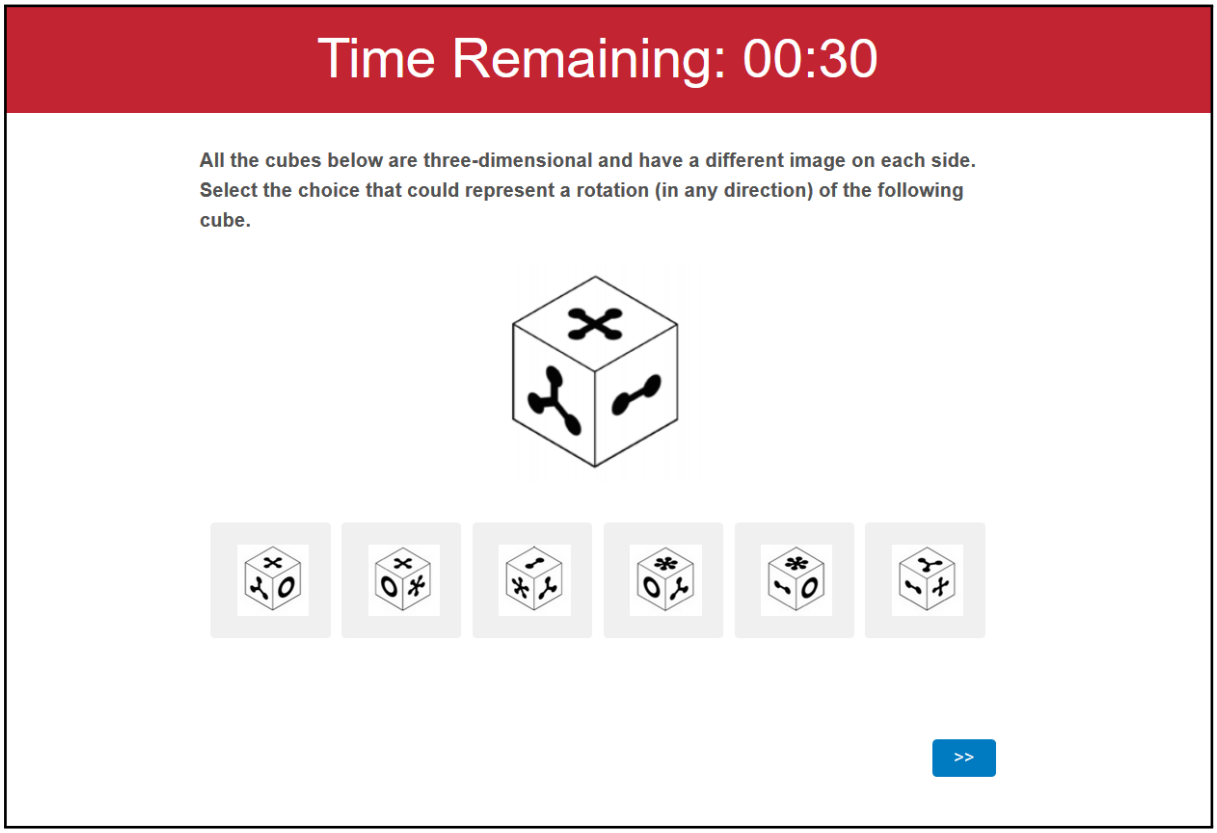


Figure C.3: Stress task - Screenshot of a quantitative reasoning task

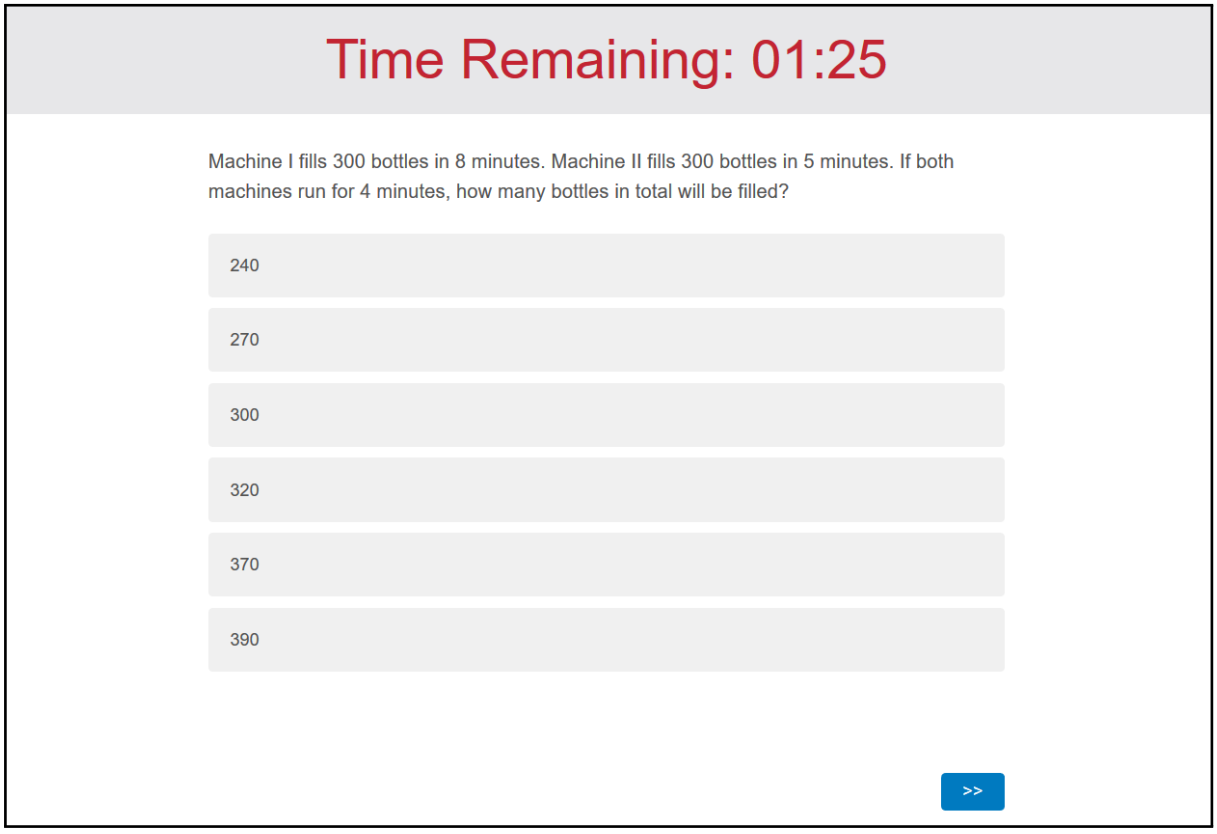
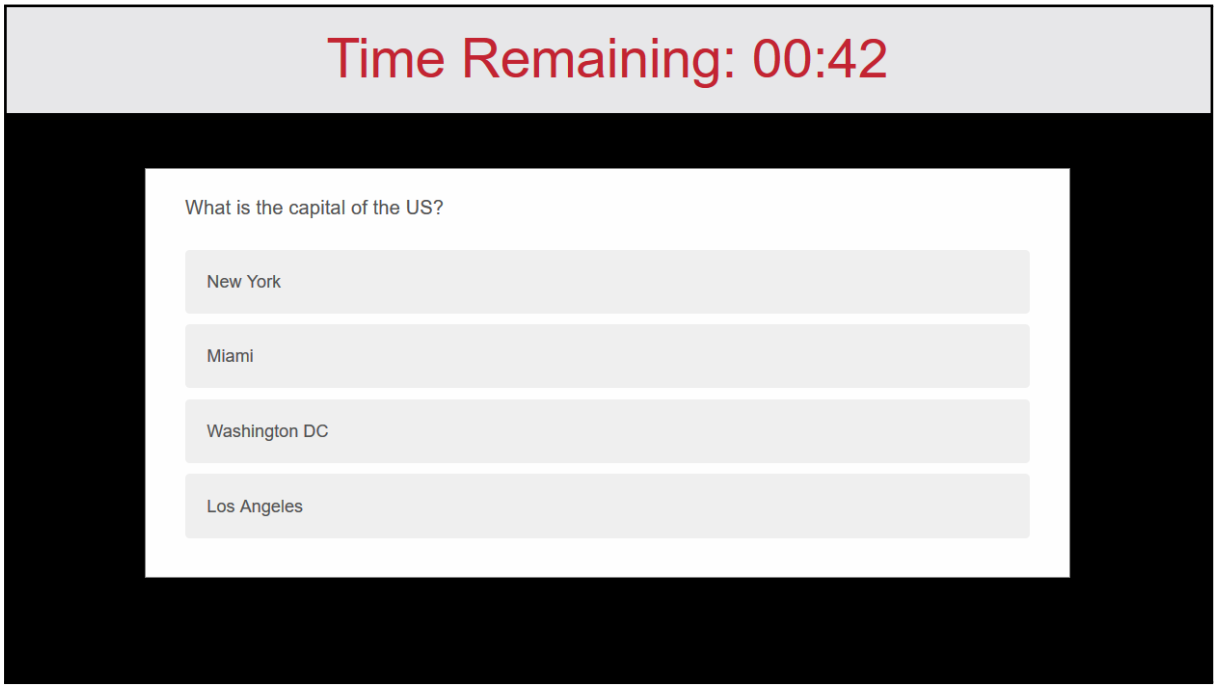


Figure C.4: Stress task - Screenshot of a pop-up knowledge question



C.2 Feedback on failure or success

Figure C.5: Screenshot of the success feedback during a stress session

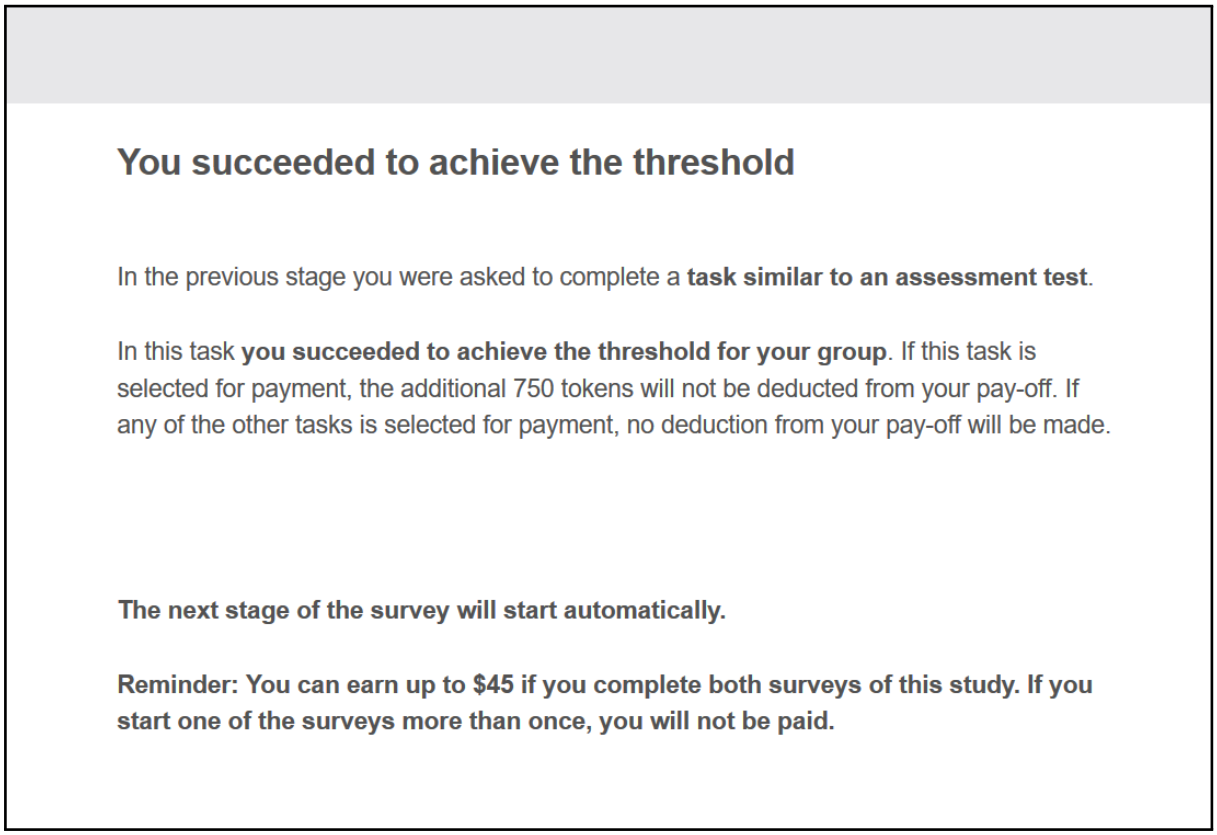


Figure C.6: Screenshot of the failure feedback during a control session

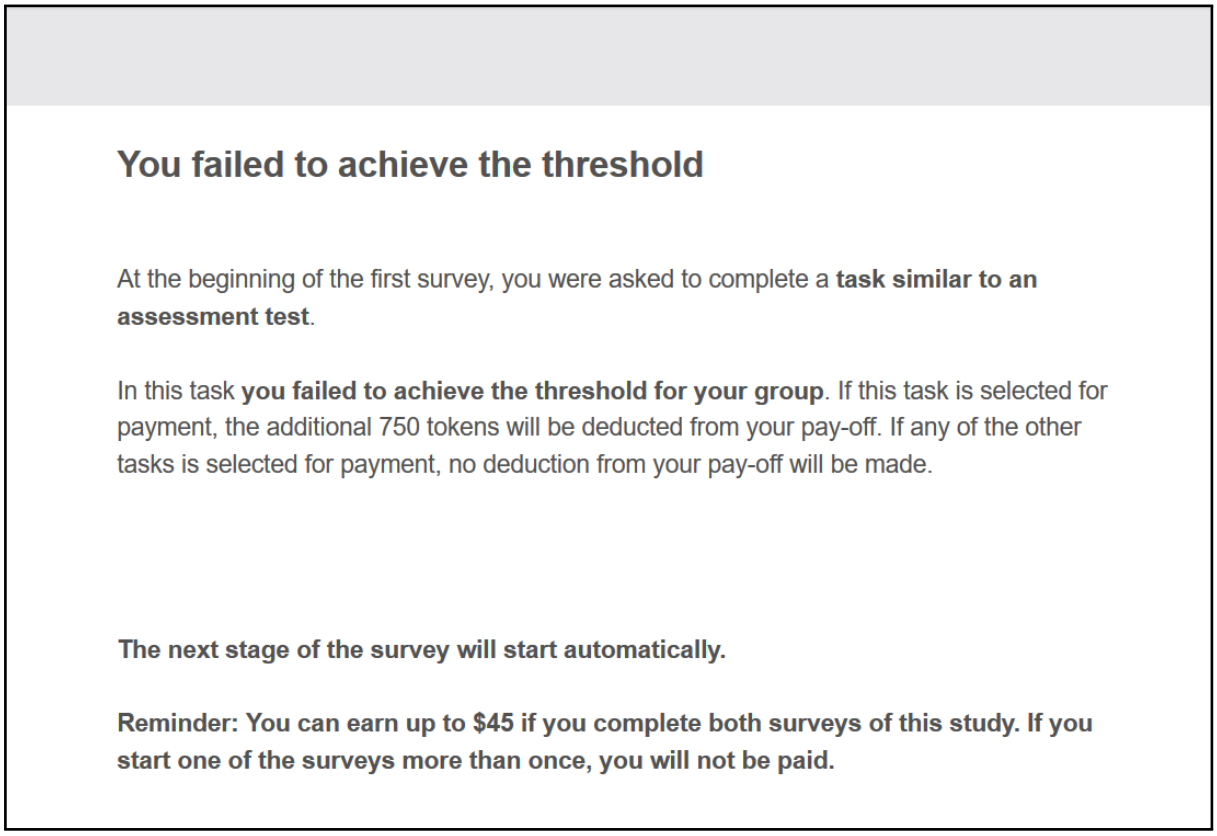
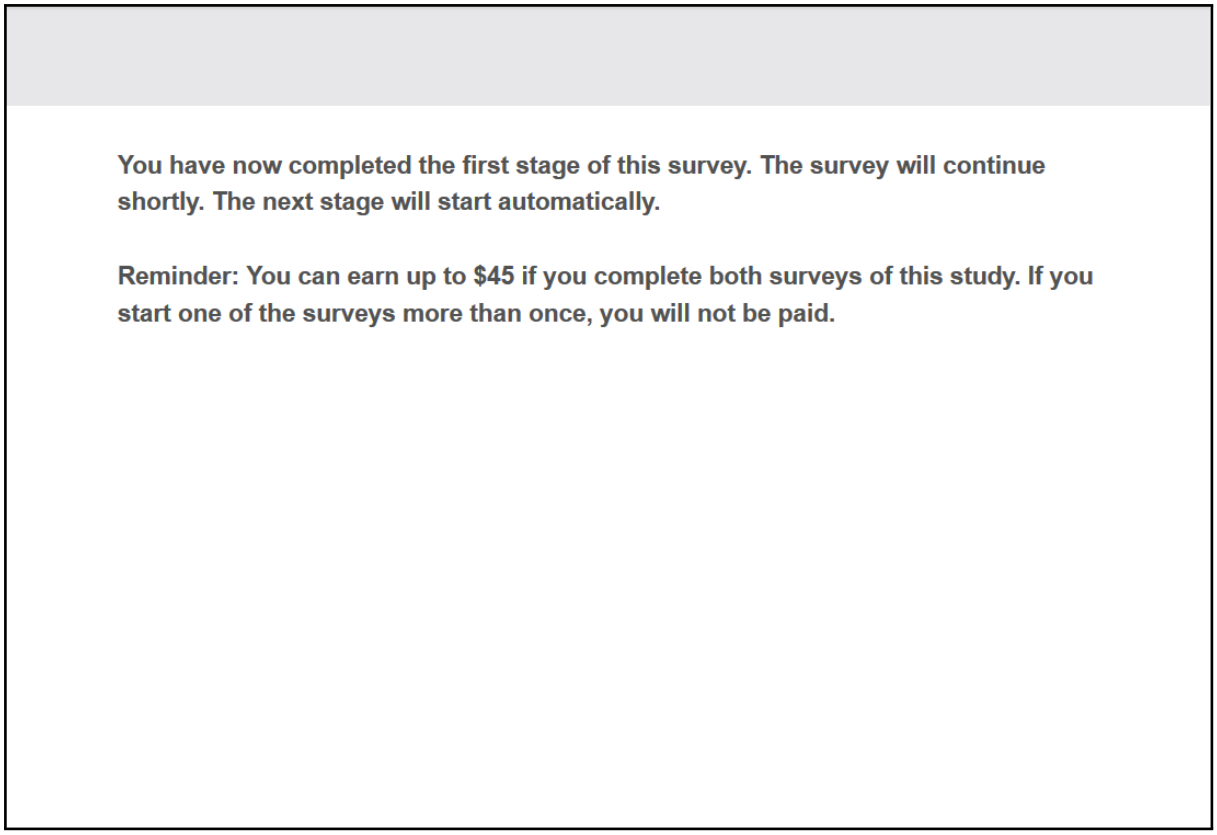


Figure C.7: Screenshot of the message shown in the "no feedback" condition



Appendix D Data

D.1 Measures of stress

To check that our protocol successfully induced stress, we collected measures of perceived stressfulness (see e.g., Kudielka *et al.*, 2004; Von Dawans *et al.*, 2011; Vitt *et al.*, 2021) and state anxiety (see e.g., Rutters *et al.*, 2009; Von Dawans *et al.*, 2011; Starcke *et al.*, 2016). If the protocol induces stress, we expect the stress task to be perceived as more stressful than the control task, and to temporarily increase state anxiety.

At the end of each session, participants were asked to indicate their perceptions of the stress or control task. They were asked to rate their agreement with each of the following perceptions on a 5-point Likert scale from 1 ("not at all") to 5 ("very much"): relaxing, easy, stressful, difficult, enjoyable, tiring, and successful. Perceived stressfulness is used to assess whether the stress task successfully induces psychological stress.

To measure changes in participants' anxiety levels throughout the experimental sessions, participants were asked to complete the six-item short-form state anxiety inventory developed by Marteau and Bekker (1992) at three points during each session: prior to starting the stress or control task (T1), after completing the stress or control task (T2) and after completing the economic decision-making tasks (T3). For the state anxiety inventory, participants were asked to rate their current feelings (calm, tense, upset, relaxed, content, worried) on a 4-point scale from "not at all" to "very much". These individual ratings are combined to give a state anxiety score, ranging from 20 to 80.

D.2 Measures of failure

To assess whether our protocol successfully manipulated participants' feelings of failure, we collected measures of perceived successfulness, self-assessed performance, and post-feedback emotions. We expect the feedback to alter participants' perception of their performance in the stress task: we expect the failure feedback to increase feelings of failure, decrease self-assessed performance, and induce negative emotions.

As part of the questions regarding participants' perceptions of the stress or control task, we asked whether the task was perceived as successful. This is used to assess whether the feedback provision was able to influence perceived success in the stress task.

At the end of the second session, participants were asked a variety of questions to capture the self-assessment of their performance in the stress task. Firstly, they were asked to rate their performance in the stress task on a 5-point scale from "very good" to "very bad". They were then asked to provide an estimate of how many of the initial 5000 tokens they lost in the stress task including the potential additional deduction. Finally, we asked participants about their expectations of having performed above or below the threshold. Participants in the "feedback" conditions were asked to recall their expectations prior to receiving feedback, while participants in the "no feedback" condition were asked about their current expectations. These measures

are used to examine whether the feedback provision was able to affect participants' assessment of their performance in the stress task.

To explore participants' emotional responses to receiving positive or negative feedback, we asked participants in the "feedback" conditions at the end of the second session to recall the emotions they felt when they were given the feedback. Specifically, they were asked to what extent they felt pleased, angry, calm, anxious, confident, disappointed, encouraged, sad, embarrassed, and successful. Each emotion was measured on a 5-point Likert scale from "not at all" to "very much".

D.3 Stress task performance

In the stress task, participants were initially allocated 5000 tokens. They could lose up to 4250 tokens based on their performance in the cognitive tasks and pop-up questions, and 750 tokens from the additional deduction. The token amount earned prior to the additional deduction captures participants' performance in the task and ranges from 750 to 5000 tokens.

The feedback provision (Treatment 2) in the second session is randomized conditionally on participants' performance in the stress task. Specifically, the feedback is randomized within the following three performance brackets: (1) Participants with a performance between the low and the high threshold (1250-4299 tokens) received either failure, success, or no feedback depending on the experimental conditions they were randomly assigned to. (2) Participants with a performance below the low threshold (750-1249 tokens) received no feedback if randomly assigned to the "no feedback" condition, or a failure feedback if randomly assigned to one of the feedback conditions. (3) Participants with a performance above the high threshold (4300-5000 tokens) received no feedback if randomly assigned to the "no feedback" treatment, or success feedback if randomly assigned to one of the feedback conditions. Since the feedback provided in the second session is randomized conditionally on the performance bracket, including indicator variables for the low and high bracket (interacted with a second session indicator variable) in all estimations is necessary.

Appendix E Empirical strategy

E.1 Fixed effects estimation

We analyse the impact of acute stress and failure or success on decision-making or behaviour by estimating models of the following form for the outcomes of interest:[2]

$$\begin{aligned} Y_{it} = {} & \beta_1 Stress_{it} + \beta_2 Failure_{it} + \beta_3 Success_{it} + \beta_4 Stress_{it} \times Failure_{it} \\ & + \beta_5 Stress_{it} \times Success_{it} + \delta_1 LowBracket_i \times Session2_t \\ & + \delta_2 HighBracket_i \times Session2_t + \gamma_t + \alpha_i + \varepsilon_{it} \end{aligned} \tag{E.1}$$

where Y_{it} denotes the outcome variable for participant i during session $t = 1, 2$. β_1, β_2, β_3, β_4 and β_5 are the coefficients of interest. $Stress_{it}$ indicates whether participant i was assigned to the "stress" condition during session t. $Failure_{it}$ and $Success_{it}$ indicate whether participant i received the "failure" or "success" feedback respectively.

The feedback shown to participants is randomized within the three performance brackets, hence controls for the performance bracket are included for observations in the second session: $LowBracket_i$ and $HighBracket_i$ are indicator variables for a performance in the stress task below the low and above the high threshold respectively. $Session2_t$ is an indicator variable for second session observations.

γ_t is a session fixed effects capturing systematic differences between the first and second session. α_i is a participant fixed effect. ε_{it} is an idiosyncratic error term.

The above equation is estimated using the fixed-effects estimator, which is identical to the first-difference estimator in the two period case. This estimation is based on a combination of within-subject and between-subject comparisons: Each participant is observed in two of the six experimental conditions across time, hence allowing a comparison of the participant's choices in these two decision environments. Additionally, different participants are observed in different experimental conditions, allowing a comparison of choices made by different participants in different decision environments.

The fixed-effects estimations are more efficient and powerful than between-subject comparisons if the variance of the individual fixed effects α_i is larger than the variance of the idiosyncratic error ε_{it}.

To account for potential heteroscedasticity and within-participant serial correlation, cluster-robust standard errors at the participant level are used in the estimations. Depending on the number of hypotheses tested, we suggest reporting p-values adjusted for multiple hypothesis testing using the bootstrapping method developed by List *et al.* (2019).

[2] As discussed in our pre-analysis plan, we intended to impose a linear structure on the effects of feedback provision, if the data approximately supported this. Imposing the assumption that failure and success feedback have effects of the same size, but in opposite directions, would have allowed for increased efficiency. Since the data does not support this assumption, we estimate the full model described here.

E.2 Attrition model

If attrition differs by experimental conditions and outcome variables in the first session, estimation focusing on participants who returned for the second session may be subject to attrition bias. To correct for this potential bias, we suggest estimating a Heckman selection model (Heckman, 1979; Wooldridge, 2010) exploring random variation in the delay (or number) of reminder emails for the second session.

In our validation study, participants automatically received an email seven days after they had completed the first session of the experiment, inviting them to complete the second session. Additionally, participants received periodical email reminders if they had not yet completed the second session. We suggest randomizing the delay (or number) of these reminder emails for the second session.[3] The random variation in the timing (or number) of reminder emails provides an exclusion restriction for the selection model.

The first difference of equation (E.1) forms the main equation of the selection model:

$$\begin{aligned}\Delta Y_{i2} = {} & \beta_1 \Delta Stress_{i2} + \beta_2 \Delta Failure_{i2} + \beta_3 \Delta Success_{i2} \\ & + \beta_4 \Delta Stress_{i2} \times \Delta Failure_{i2} + \beta_5 \Delta Stress_{i2} \times \Delta Success_{i2} \\ & + \delta_1 LowBracket_i + \delta_2 HighBracket_i + \Delta\gamma_2 + \Delta\varepsilon_{i2}\end{aligned} \tag{E.2}$$

Attrition between the two sessions is then captured by the following selection equation:

$$P(S_{i2}) = \Phi(\theta_0 + \theta_1 ReminderDelay_i + \theta_2 Stress_{i1}) \tag{E.3}$$

where S_{i2} indicates whether participant i completed the second session and $\Phi(.)$ is the cumulative distribution function of the normal distribution. $ReminderDelay_i$ is the potential delay between the automatic invite email for the second session and the first reminder email.[4] We impose the exclusion restriction that this delay affects selection, but does not directly affect the outcome variables of our experiment. In our validation study, gaps between these email reminders varied between 1 and 9 days.

In addition to the delay variable, we include $Stress_{i1}$ in the selection equation, which captures whether a participant was assigned to the "stress" condition during the first session. All other regressors from the main equation are not included in the selection equation as they are only determined during the second session for some or all participants (feedback condition and performance bracket), or are not appropriate in the attrition equation (session fixed effects).

[3] In our validation study, the reminder timing was not explicitly randomized. Instead, the experimental team manually sent periodical email reminders to all participants eligible for the second session who had not yet completed it. The combination of automatic initial email invites and manually sent email reminders creates variation in the time gap between the initial invite and the first reminder. Based on the date of participants' first session and the dates on which manual reminders were sent, we calculated the potential delay of the first reminder in days, even for participants who completed the second session prior to receiving the reminder.

[4] Since some participants will complete the second session prior to receiving the first reminder email, it is crucial to randomize the potential delay of the reminder email for all participants and use this potential delay as the exclusion restriction.

Appendix F Power calculations

In the following, we calculate the minimum detectable effects at a significance level of $\alpha = 5\%$ and with a power of 80% for the fixed effects model described in Online Appendix E.1:

$$\begin{aligned} Y_{it} = {} & \beta_1 Stress_{it} + \beta_2 Failure_{it} + \beta_3 Success_{it} + \beta_4 Stress_{it} \times Failure_{it} \\ & + \beta_5 Stress_{it} \times Success_{it} + \delta_1 LowBracket_i \times Session2_t \\ & + \delta_2 HighBracket_i \times Session2_t + \gamma_t + \alpha_i + \varepsilon_{it} \end{aligned} \tag{F.1}$$

where $\alpha_i \sim N(0, {\sigma_\alpha}^2)$ and $\varepsilon_{it} \sim N(0, {\sigma_\varepsilon}^2)$. We calculate the minimum detectable effects for sample sizes $N = \{150, 240, 300, 450, 600\}$ and assume that the sample size for each experimental condition is $N/6$.

While we recommend the use of the more efficient bootstrap approach proposed in List *et al.* (2019) to correct for multiple hypothesis testing, for simplicity we use the Šidák correction (Šidák, 1967) in the context of these power calculations.[5] We present minimum detectable effects based on Šidák corrected significance thresholds $\alpha_m = 1 - (1 - \alpha)^{1/m}$ for $m = \{1, 2, 3, 4, 5, 10, 15, 20\}$ hypotheses.

Based on previously reported test-retest correlations of 0.6 and higher for risk and time preferences over time periods between 2 and 4 weeks (Sunde and Dohmen, 2016; Wölbert and Riedl, 2013), we make the following assumption:

$$\frac{\sigma_\alpha^2}{\sigma^2} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2} = 0.6 \tag{F.2}$$

Data from our validation study show test-retest correlations between 0.65 and 0.68 for the different economic decision outcomes, confirming that a value of 0.6 is a cautious assumption.

The classical covariance estimator for the fixed effects estimator is given by:

$$\widehat{Var(\beta_{FE})} = \sigma_\varepsilon^2 (\widetilde{X}'\widetilde{X})^{-1} = 0.4 \times \sigma^2 (\widetilde{X}'\widetilde{X})^{-1} \tag{F.3}$$

where $\widetilde{X}$ denotes the demeaned regressors. The second equality follows from the assumption on test-retest correlation given in equation (F.2). The minimum detectable effect (MDE) of a hypothesis test at significance level α_m with 80% power ($\kappa = 0.8$) is given by:

$$\begin{aligned} MDE &= \left(t_{N-k,0.5\alpha_m} + t_{N-k,(1-\kappa)}\right) \times \sqrt{\widehat{Var(\beta_{FE})}} \\ &= \left(t_{N-8,\alpha_m/2} + t_{N-8,(1-0.8)}\right) \times \sqrt{0.4 \times \sigma^2 (\widetilde{X}'\widetilde{X})^{-1}} \\ &= \left(t_{N-8,\alpha_m/2} + t_{N-8,(0.2)}\right) \times \sqrt{0.4 \times (\widetilde{X}'\widetilde{X})^{-1}} \times \sigma \end{aligned} \tag{F.4}$$

Due to the feedback randomization in the experiment being conditional on participants' performance, the number of participants receiving success and failure feedback in each condition is sample-dependent and therefore $(\widetilde{X}'\widetilde{X})^{-1}$ is a random variable since. Hence the variance and

[5] Due to the underlying assumption of independent hypotheses, the use of the Šidák correction will result in a lower bound for power / upper bound for minimum detectable effects.

the minimum detectable effects are sample-dependent random variables as well. We conducted Monte Carlo simulation (with 1000 replications) for each sample size N and report the corresponding mean values for the minimum detectable effects (expressed in standard deviations of the outcome variable) in Table F.1.

Table F.1: Minimum detectable effects

(a) $N = 150$

Coefficient (Variable)	Minimum detectable effect (in standard deviations) when testing m hypotheses							
	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 10$	$m = 15$	$m = 20$
β_1 $(Stress)$	0.359	0.396	0.415	0.429	0.439	0.470	0.487	0.499
β_2 $(Failure)$	0.723	0.795	0.835	0.863	0.883	0.945	0.979	1.003
β_3 $(Success)$	0.721	0.794	0.834	0.861	0.882	0.943	0.977	1.001
β_4 $(Stress \times Failure)$	1.019	1.121	1.177	1.216	1.245	1.332	1.380	1.414
β_5 $(Stress \times Success)$	1.015	1.117	1.173	1.212	1.241	1.327	1.376	1.409

(b) $N = 240$

Coefficient (Variable)	Minimum detectable effect (in standard deviations) when testing m hypotheses							
	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 10$	$m = 15$	$m = 20$
β_1 $(Stress)$	0.282	0.311	0.326	0.337	0.345	0.368	0.382	0.391
β_2 $(Failure)$	0.568	0.625	0.656	0.677	0.693	0.741	0.768	0.786
β_3 $(Success)$	0.567	0.624	0.655	0.676	0.692	0.740	0.766	0.785
β_4 $(Stress \times Failure)$	0.800	0.879	0.923	0.953	0.976	1.043	1.080	1.106
β_5 $(Stress \times Success)$	0.799	0.879	0.923	0.953	0.975	1.042	1.080	1.106

(c) $N = 300$

Coefficient (Variable)	Minimum detectable effect (in standard deviations) when testing m hypotheses							
	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 10$	$m = 15$	$m = 20$
β_1 $(Stress)$	0.252	0.277	0.291	0.301	0.308	0.329	0.341	0.349
β_2 $(Failure)$	0.507	0.557	0.585	0.604	0.618	0.661	0.684	0.700
β_3 $(Success)$	0.507	0.557	0.585	0.604	0.618	0.661	0.684	0.700
β_4 $(Stress \times Failure)$	0.714	0.785	0.824	0.850	0.871	0.930	0.964	0.987
β_5 $(Stress \times Success)$	0.714	0.785	0.824	0.850	0.871	0.930	0.964	0.987

(d) $N = 450$

Coefficient (Variable)	Minimum detectable effect (in standard deviations) when testing m hypotheses							
	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 10$	$m = 15$	$m = 20$
β_1 $(Stress)$	0.206	0.226	0.237	0.245	0.251	0.268	0.277	0.284
β_2 $(Failure)$	0.413	0.454	0.477	0.492	0.504	0.538	0.557	0.570
β_3 $(Success)$	0.412	0.453	0.476	0.491	0.503	0.537	0.556	0.569
β_4 $(Stress \times Failure)$	0.582	0.639	0.671	0.693	0.709	0.757	0.784	0.803
β_5 $(Stress \times Success)$	0.581	0.639	0.670	0.692	0.708	0.757	0.784	0.802

(e) $N = 600$

Coefficient (Variable)	Minimum detectable effect (in standard deviations) when testing m hypotheses							
	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 10$	$m = 15$	$m = 20$
β_1 $(Stress)$	0.178	0.195	0.205	0.212	0.217	0.231	0.240	0.245
β_2 $(Failure)$	0.357	0.393	0.412	0.425	0.435	0.465	0.481	0.493
β_3 $(Success)$	0.357	0.393	0.412	0.425	0.435	0.465	0.481	0.493
β_4 $(Stress \times Failure)$	0.503	0.553	0.580	0.599	0.613	0.654	0.678	0.694
β_5 $(Stress \times Success)$	0.503	0.553	0.580	0.599	0.613	0.655	0.678	0.694

Note: Minimum detectable effects reported here are expressed in terms of standard deviations of the outcome variable. We report the mean values from Monte Carlo simulations with 1000 replications, for each sample size N and number of hypotheses m.

Appendix G Additional results

G.1 Other results

Table G.1: Demographic characteristics across experimental conditions

	S1: Stress in Session 1	S2: Stress in Session 2	S1 vs S2 P-Value
Age	24.33	22.58	0.02
	(7.29)	(5.03)	
Gender identity:			
female	0.68	0.65	0.60
male	0.30	0.35	0.44
other/unknown	0.02	0.01	0.34
Undergraduate student	0.70	0.74	0.53
	(0.46)	(0.44)	
Subject area:			
business	0.33	0.36	0.60
science/engineering	0.38	0.37	0.91
arts/humanities/social sciences	0.17	0.16	0.68
health	0.10	0.08	0.55
other/unknown	0.02	0.04	0.46
Out of state	0.36	0.42	0.34
	(0.48)	(0.49)	
International student	0.15	0.20	0.31
	(0.36)	(0.40)	
Living on campus	0.12	0.19	0.09
	(0.32)	(0.40)	
Monthly income:			
$0 - $599	0.23	0.25	0.65
$600 - $999	0.16	0.18	0.68
$1000 - $1999	0.34	0.34	0.89
$2000 and higher	0.28	0.23	0.34
Employment during term time:			
none	0.33	0.38	0.38
part-time	0.49	0.50	0.81
full-time	0.18	0.12	0.13
Employed during summer break	0.45	0.50	0.40
	(0.50)	(0.50)	
Exam accommodations	0.04	0.08	0.15
	(0.19)	(0.28)	
N	139	130	269

Table G.2: Demographic characteristics across the feedback conditions

	(1) N: no feedback	(2) S: low threshold	(3) F: high threshold	F-test P-value
Age	23.20	23.49	23.77	0.83
	(6.87)	(5.40)	(6.66)	
Gender identity:				
female	0.65	0.65	0.69	0.80
male	0.33	0.34	0.30	0.83
other/unknown	0.02	0.01	0.01	0.82
Undergraduate student	0.78	0.70	0.67	0.21
	(0.41)	(0.46)	(0.47)	
Subject area:				
business	0.33	0.33	0.36	0.90
science/engineering	0.40	0.37	0.36	0.86
arts/humanities/social sciences	0.15	0.20	0.15	0.67
health	0.09	0.07	0.10	0.77
other/unknown	0.03	0.04	0.02	0.88
Out of state	0.47	0.30	0.38	0.06
	(0.50)	(0.46)	(0.49)	
international student	0.21	0.14	0.17	0.45
	(0.41)	(0.35)	(0.38)	
living on campus	0.17	0.09	0.19	0.18
	(0.38)	(0.29)	(0.40)	
Monthly income:				
$0 - $599	0.17	0.25	0.30	0.13
$600 - $999	0.20	0.14	0.16	0.57
$1000 - $1999	0.40	0.35	0.26	0.12
$2000 and higher	0.22	0.25	0.28	0.64
Employment during term time:				
none	0.36	0.33	0.37	0.87
part-time	0.49	0.54	0.45	0.48
full-time	0.15	0.12	0.18	0.59
Employed during summer break	0.46	0.51	0.46	0.78
	(0.50)	(0.50)	(0.50)	
Exam accommodations	0.08	0.04	0.06	0.51
	(0.27)	(0.19)	(0.24)	
N	94	85	90	269

Table G.3: State anxiety response to the stress / control task

	T1 Baseline	T2 Post-Task	T3 End	Difference: T2 - T1
Stress	37.559	46.709	43.110	9.149***
	(0.757)	(0.852)	(0.809)	(0.667)
Control	36.750	36.639	40.084	−0.112
	(0.781)	(0.789)	(0.751)	(0.525)
Difference:	0.809	10.070***	3.026***	9.261***
Stress - Control	(0.753)	(0.854)	(0.666)	(0.811)

Note: Means and mean differences were obtained using the sample of participants (N=239) who responded to all state anxiety questions at times T1, T2 and T3 in both sessions. Participants with a performance level above the high or below the low threshold were excluded from the sample for simplicity. Standard errors were clustered at the participant level and are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

Figure G.1: Distribution of the stress task performances

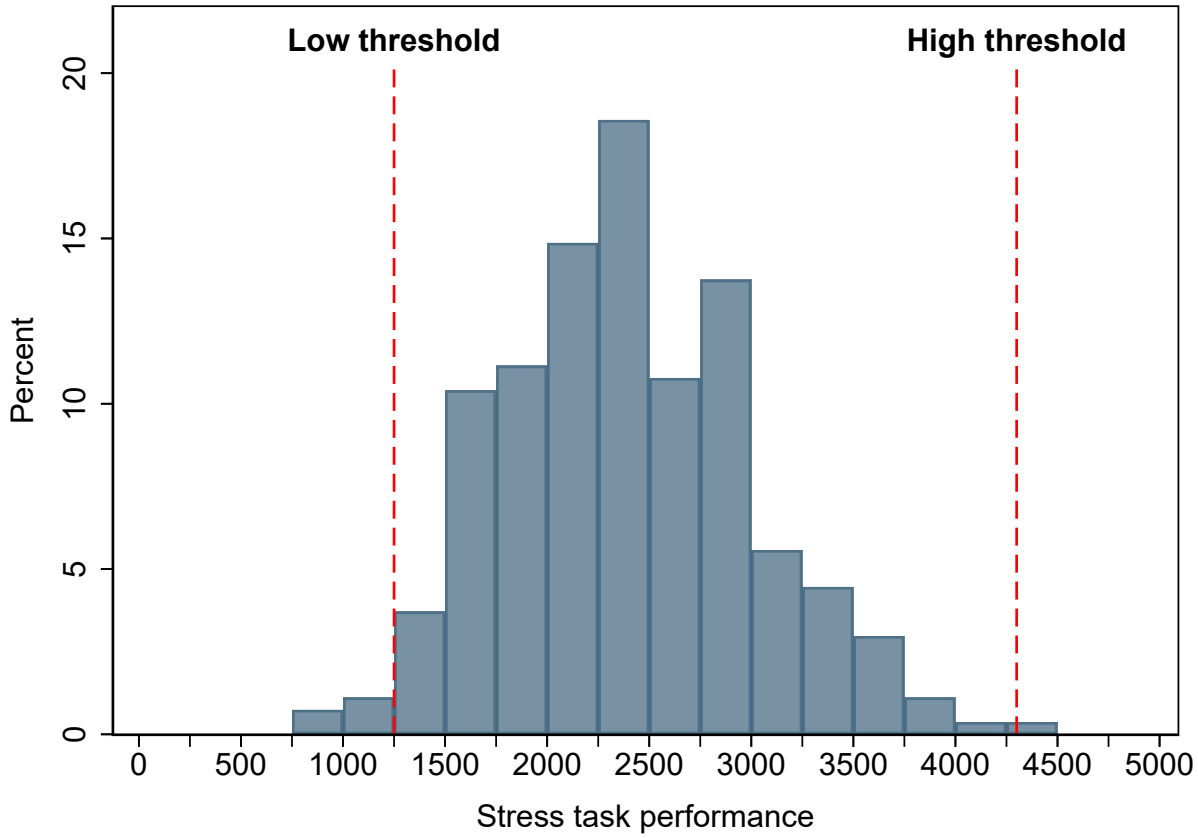


Figure G.2: Comparison of the stress task performances between the first and second session

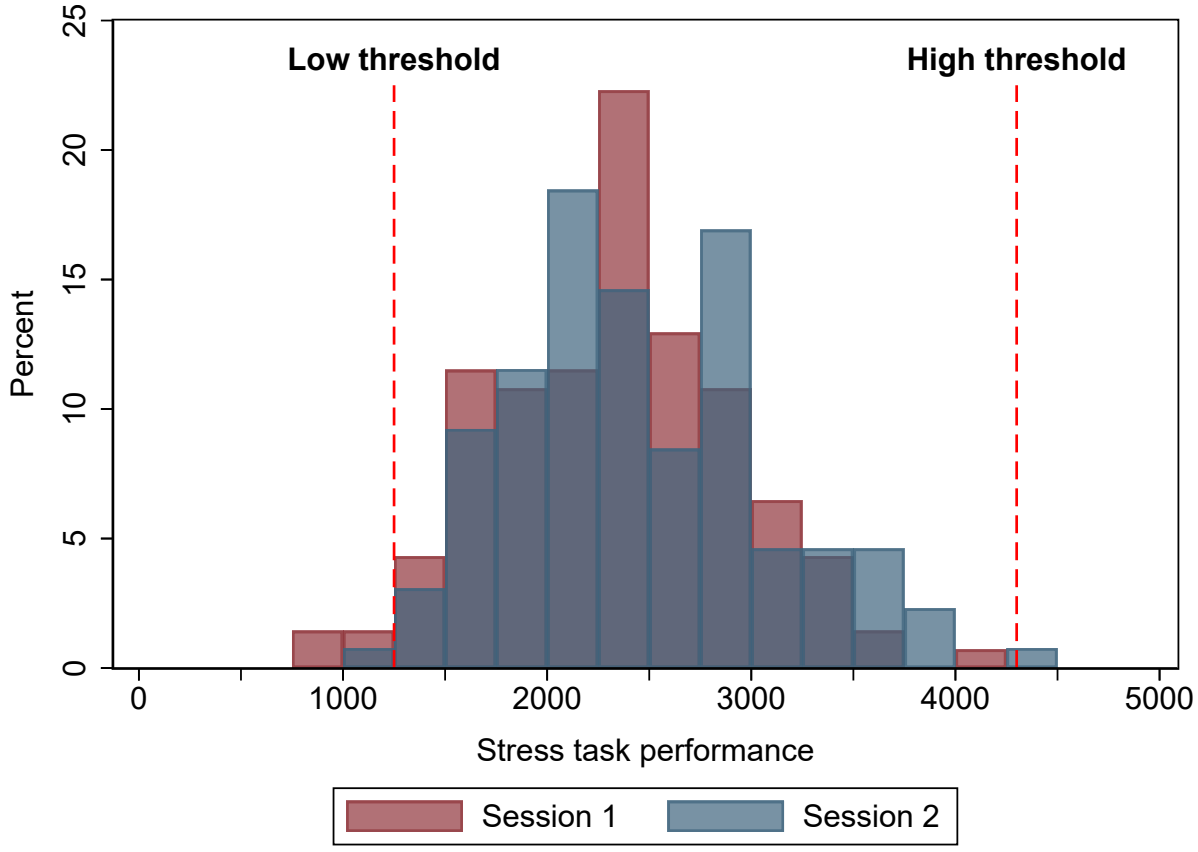


Table G.4: Success expectations, asked at the end of the second session

	No feedback condition: Success expectation	Feedback conditions: Recalled success expectation
No deduction	0.660	0.637
	(0.070)	(0.054)
Deduction	0.702	0.435
	(0.067)	(0.054)
Difference:	0.043	−0.202**
	(0.097)	(0.076)
N	94	165

Note: The above table summarizes participants' expectation of being above the success threshold in the stress task, asked at the end of the second session. In the feedback condition, participants were asked to recall their expectations prior to receiving feedback. Participants with a performance level above the high or below the low threshold were excluded from the sample for simplicity. Standard errors are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

Table G.5: Expected token loss in the stress task, asked at the end of the second session

	(1)	(2)	(3)	Differences:		
	No feedback	Success	Failure	(2) - (1)	(3) - (1)	(3) - (2)
S2: Stress - Session 2	2001.489	1703.684	2527.439	−297.805	525.950*	823.755***
	(156.728)	(149.943)	(184.311)	(216.903)	(241.939)	(237.600)
S1: Stress - Session 1	1909.574	1301.163	1739.409	−608.412**	−170.165	438.246
	(140.409)	(169.623)	(181.998)	(220.197)	(229.865)	(248.788)
Difference:	91.915	402.521	788.030**	310.607	696.115*	385.509
Session 2 - Session 1	(210.424)	(226.396)	(259.025)	(309.085)	(333.725)	(344.019)
N	94	81	85			

Note: Participants with a performance level above the high or below the low threshold were excluded from the sample for simplicity. Standard errors are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

G.2 Attrition

Participants in the experiment were asked to complete two online sessions. However, not all participants returned for the second session. Of the 317 participants who completed the first session, 269 (84.86%) completed the second session, 12 (3.79%) provided invalid or incomplete responses in the second session and 36 (11.36%) did not return for the second session.

This type of sample attrition, a common occurrence in longitudinal studies, can result in biased estimates if the attrition is non-random. We hence examine whether attrition differs between experimental conditions and whether it is related to any variables in the first session.

As shown in Table G.6, we observe a statistically significant difference in attrition between participants assigned to the stress task and the control task in the first session. While 89.66% of those assigned to the control task in the first session provided a valid response in the second session, only 80.81% of those assigned to the stress task did. This difference in attrition rates

suggests attrition may not be random. It is therefore necessary to correct for potential attrition bias in the analysis.

To further examine whether attrition between the two sessions is random, we compare data from the first session of the experiment between the attrited and returning participants in Tables G.7 and G.8. Within both the stress and the control groups, we find no statistically significant difference in the task perceptions and measures of psychological stress between participants who returned for the second session and participants who did not. This indicates that the reason for higher attrition in the stress group is not that those who perceived the task as particularly stressful or difficult did not return for the second session.

We furthermore observe an association between attrition and some of the outcome decisions, underlining the importance of correcting for potential attrition bias in the analysis. We develop a Heckman selection model to correct for potential attrition bias, details of the model can be found in Online Appendix E.2. We explore the (quasi-random) delay of reminder emails as an exclusion restriction, as it influences the likelihood of attrition and is unrelated to experimental conditions or participant characteristics. Results reported in Online Appendix G.3 show the findings of our validation study to be robust to correcting for non-random attrition.

For future research, we suggest collecting the demographics for the full sample at the end of the first session, as specified in the protocol. This enables more detailed attrition balance checks and a potential correction for attrition bias using inverse probability weighting. We furthermore suggest randomizing the delay of the reminder emails to be able to employ it as an exclusion restriction in a Heckman selection model.

Table G.6: Attrition

	Experimental condition in session 1						
	Total		S1: Stress		S2: Control		S1 vs S2
	N	%	N	%	N	%	P-value
Session 1:	317	100.00	172	100.00	145	100.00	
Session 2:							
- no response	36	11.36	26	15.12	10	6.90	0.02
- incomplete / invalid response	12	3.79	7	4.07	5	3.45	0.77
- complete / partial response	269	84.86	139	80.81	130	89.66	0.03

Table G.7: Comparison of first session characteristics by attrition status

	Mean difference: attrited - returning		
	Total	Session 1: Stress	Session 1: Control
Task perception:			
relaxing	−0.03	0.25	−0.09
	(0.21)	(0.25)	(0.34)
easy	−0.21	0.27	−0.34
	(0.20)	(0.23)	(0.25)
stressful	0.22	−0.06	0.11
	(0.21)	(0.25)	(0.28)
difficult	0.21	−0.14	0.12
	(0.19)	(0.22)	(0.21)
enjoyable	−0.17	−0.03	−0.15
	(0.18)	(0.22)	(0.31)
successful	−0.13	0.06	−0.07
	(0.17)	(0.19)	(0.29)
tiring	0.20	0.04	0.19
	(0.19)	(0.24)	(0.28)
State anxiety:			
task response (T2 - T1)	3.47	1.41	1.81
	(1.82)	(2.35)	(1.27)
Altruism task:			
Average donation	7.95	13.74*	−1.83
	(4.63)	(5.36)	(8.99)
Risk taking task:			
Indifference point	6.26	6.38	8.00
	(3.26)	(3.52)	(6.97)
Intertemporal substitution task:			
Delta (long-term discount rate)	−0.03	−0.03	−0.04
	(0.03)	(0.03)	(0.05)
Beta (present bias)	−0.07	−0.01	−0.18*
	(0.05)	(0.06)	(0.07)
Effort task:			
Number of sliders	0.23	−0.58	2.26
	(1.45)	(1.86)	(2.25)

Note: The mean differences reported are between attrited and returning participants for the second session. Mean differences are based on a t-test, not assuming equal variances. Standard errors are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

Table G.8: Comparison of first session timing by attrition status

	Mean difference: attrited - returning
Study day - first session	2.91
	(2.15)
Month of first session:	
July	0.02
	(0.06)
Day of first session:	
Sunday	0.02
	(0.04)
Monday	0.14*
	(0.07)
Tuesday	0.16*
	(0.07)
Wednesday	−0.04
	(0.04)
Thursday	−0.14*
	(0.07)
Friday	−0.10**
	(0.04)
Saturday	−0.05***
	(0.01)
Time of first session:	
8:00 - 11:59	0.02
	(0.07)
12:00 - 14:59	−0.08
	(0.06)
15:00 - 18:59	−0.01
	(0.07)
19:00 - 22:59	0.08
	(0.07)
23:00 - 07:59	−0.01
	(0.03)

Note: The mean differences reported are between attrited and returning participants for the second session. Mean differences are based on a t-test, not assuming equal variances. Standard errors are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

G.3 Attrition model

As discussed above, attrition between the first and second session was higher among participants assigned to the stress task in the first session. To correct for potential attrition bias in the validation results, we estimate the Heckman selection model in equations (E.2) and (E.3) for the validation outcomes. The corresponding estimation results are reported in Tables G.9 and G.10, which compare results with and without the Heckman correction for attrition. The attrition-corrected results do not differ substantially from the non-corrected results.

The attrition correction model estimated here imposes the exclusion restriction that the delay between the automatic invite and first reminder email for the second session affects attrition in the selection equation, but does not directly affect the outcome variables in the main equation. Without exclusion restrictions, Heckman models rely exclusively on the assumption of a normal distribution to correct for selection, so it is important to examine the strength of the exclusion restriction and evaluate potential threats to its validity.

In the estimated selection equation, a longer delay of the reminder email predicts a significantly higher probability of participants returning for the second session. This suggests that reminders with a short delay were less effective than those with a longer delay, potentially be-

Table G.9: Validation results - Fixed effects / first difference estimations with and without a Heckman correction for attrition

	Task: stressful		Task: successful		STAI Response	
	(1) FD/FE	(2) FD Heckman	(3) FD/FE	(4) FD Heckman	(5) FD/FE	(6) FD Heckman
Main equation:						
Stress	1.250***	1.262***	−0.891***	−0.941***	11.943***	11.809***
	(0.145)	(0.148)	(0.130)	(0.137)	(1.276)	(1.310)
Failure	0.194	0.182	−0.055	−0.002	10.000***	10.114***
	(0.293)	(0.291)	(0.263)	(0.256)	(2.540)	(2.515)
Stress * Failure	0.314	0.325	−0.427	−0.472	−9.260*	−9.367**
	(0.419)	(0.414)	(0.378)	(0.372)	(3.607)	(3.561)
Success	−0.452	−0.453	0.285	0.290	1.295	1.323
	(0.298)	(0.293)	(0.268)	(0.260)	(2.567)	(2.524)
Stress * Success	0.673	0.674	0.119	0.115	−6.123	−6.147
	(0.430)	(0.423)	(0.385)	(0.379)	(3.715)	(3.656)
Controls for:						
Second session	Yes	Yes	Yes	Yes	Yes	Yes
Performance bracket	Yes	Yes	Yes	Yes	Yes	Yes
Selection equation:						
Delay - first reminder		0.245**		0.244**		0.243**
		(0.075)		(0.075)		(0.076)
Stress		0.162		0.161		0.167
		(0.092)		(0.092)		(0.093)
Constant		0.401		0.402		0.396
		(0.208)		(0.208)		(0.210)
Observations (selected)	512	512	510	510	502	502
Participants (selected)	256	256	255	255	251	251
Participants (not selected)		48		48		48

Note: Standard errors are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

Table G.10: Validation results for success / failure - Estimations with and without a Heckman correction for attrition

	Perceived performance		Expected token loss	
	(1) OLS	(2) Heckman	(3) OLS	(4) Heckman
Main equation:				
Stress	−0.596**	−0.618**	91.915	86.792
	(0.209)	(0.210)	(224.260)	(221.487)
Failure	−0.300	−0.340	−206.732	−216.034
	(0.212)	(0.209)	(227.017)	(225.038)
Stress * Failure	−0.291	−0.257	771.924*	779.628*
	(0.301)	(0.295)	(322.424)	(318.186)
Success	0.193	0.187	−608.412**	−609.785**
	(0.214)	(0.210)	(229.417)	(225.955)
Stress * Success	−0.040	−0.035	310.607	311.726
	(0.308)	(0.302)	(329.970)	(324.924)
Constant	2.574***	2.327***	1909.574***	1853.083***
	(0.148)	(0.216)	(158.576)	(222.900)
Controls for:				
Performance bracket	Yes	Yes	Yes	Yes
Selection equation:				
Delay - first reminder		0.281***		0.281***
		(0.061)		(0.061)
Stress		−0.067		−0.067
		(0.145)		(0.145)
Constant		−0.061		−0.061
		(0.181)		(0.181)
Observations (selected)	266	266	266	266
Participants (selected)	266	266	266	266
Participants (not selected)		96		96

Note: Standard errors are shown in parentheses. Significance levels are indicated as follows: * p<0.05, ** p<0.01, *** p<0.001

cause participants were annoyed about too frequent emails or because reminders became more helpful as the intention to complete the study began to slip from participants' minds.

In our validation study, the reminder timing was not explicitly randomized. Instead, the experimental team manually sent periodical email reminders to all participants eligible for the second session who had not yet completed it. The combination of automatic initial email invites and manually sent email reminders creates variation in the time gap between the initial invite and the first reminder. Based on the date of participants' first session and the dates on which manual reminders were sent, we calculated the potential delay of the first reminder in days, even for participants who completed the second session prior to receiving the reminder. To ensure that the relationship between email delay and attrition is not driven by an omitted variable, we look closer at the delay variable and check for any covariates among the variables of the study.

The majority of participants receive the reminder email between 1 to 4 days after the initial email, with only 6% of participants receiving the reminder with a delay of more than 4 days (see Figure G.3). The delay is calculated based on participants' completion date of the first session, therefore all participants who completed the first session on the same day were subject to the same delay. Figure G.4 illustrates how the delay varies over the study period. It is therefore not surprising that the delay variable is related to the weekday of the first session.

We furthermore find the delay variable to be related to state anxiety at baseline during the first session (see Tables G.11 and G.12). Within the selected sample, for which demographic variables are available, we also observe differences in out-of-state and international student rates as well as previous experiment participation between those with shorter and longer reminder delays (see Table G.13).

We cannot examine the robustness of the selection equation to including these demographic variables as they are not available for some of the attrited participants. However, our estimation results of the selection equation are robust to controlling for the weekday of the first session and for baseline state anxiety during the first session.

Further analysis shows that the significant relationship between reminder delay and attrition was not driven by outliers, and predictions based on the Probit model accurately capture the distribution of conditional attrition probabilities. This is illustrated in Figure G.5.

A weak exclusion restriction in the selection equation can result in multicollinearity problems when estimating a Heckman selection model. Following the recommendation in Certo *et al.* (2016), we therefore examine the correlations between the Inverse Mills Ratio and the regressors of interest in our main equation. With correlations (in absolute values) between 0.01 and 0.50, multicollinearity does not appear to be a major concern (see Tables G.14 and G.15).

These additional analyses do not identify any problems with the selection equation and the imposed exclusion restriction, thus supporting the use of the Heckman correction model estimated here.

Figure G.3: Distribution of the email reminder delays

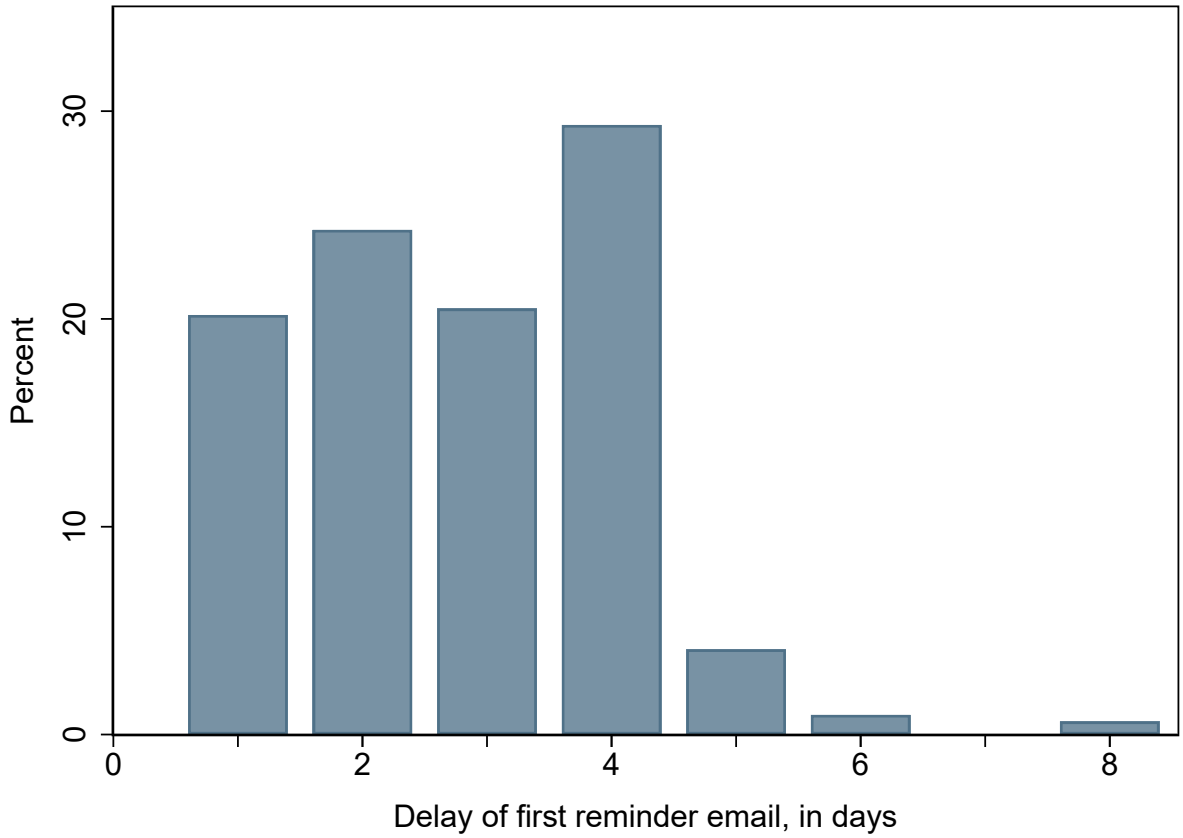


Figure G.4: Email reminder delays across the study period

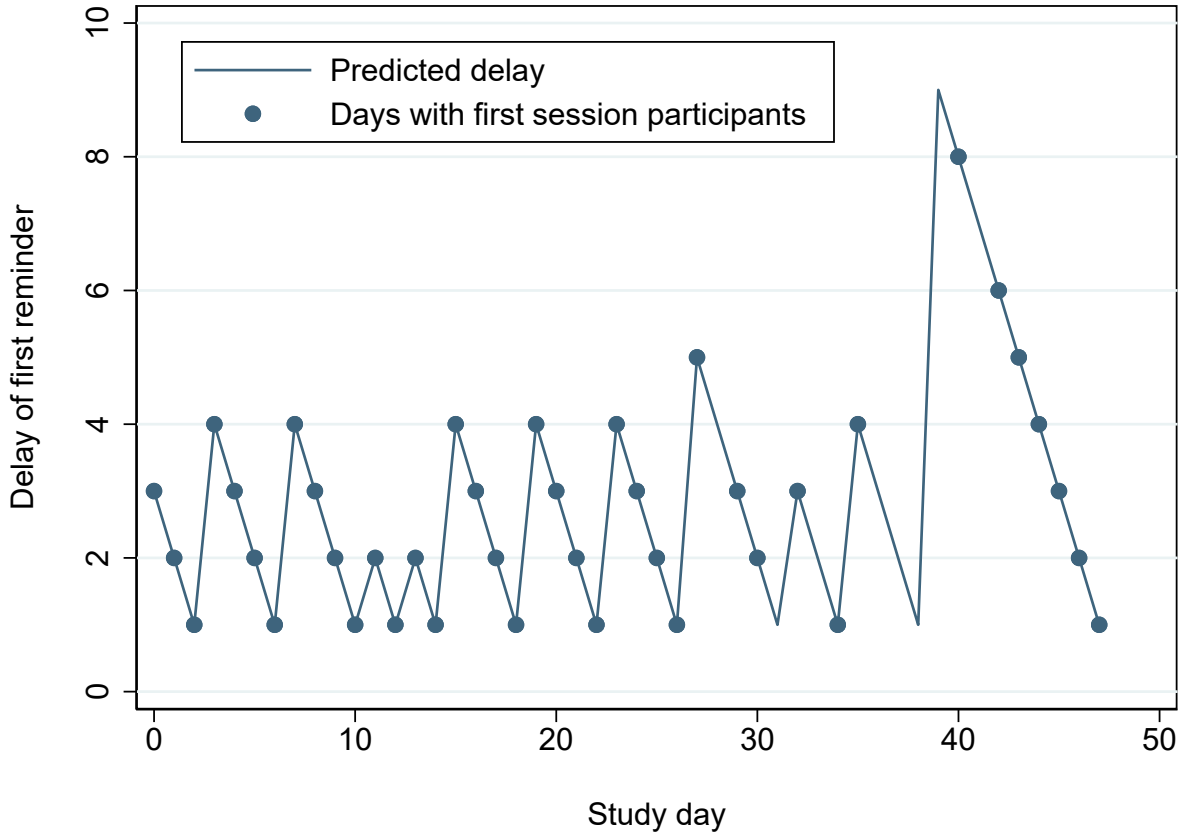


Figure G.5: Email reminder delay and attrition

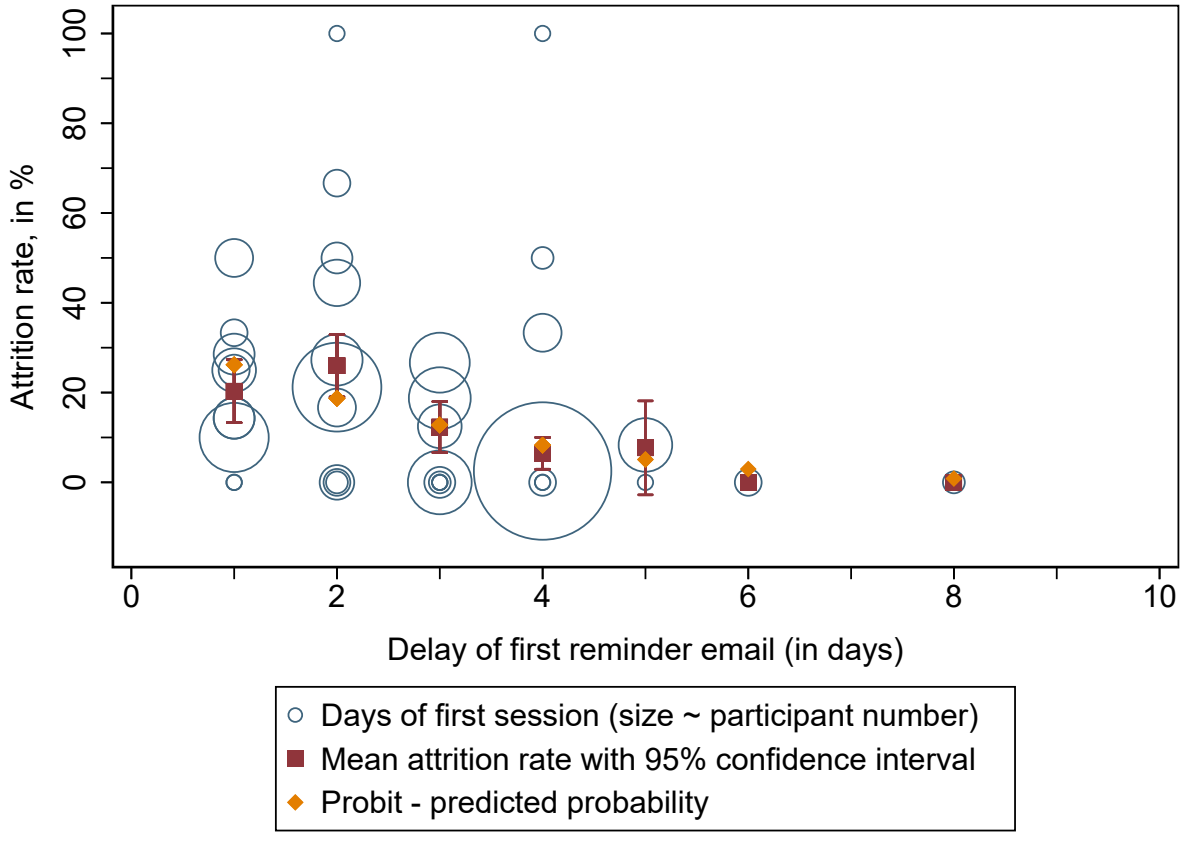


Table G.11: Balance of first session variables across email reminder delays - Part 1

	Email reminder delay		Difference
	1-3 days	4+ days	P-Value
Stress - 1st session	0.57	0.49	0.14
	(0.50)	(0.50)	
Day of first session:			
Sunday	0.06	0.02	0.03
Monday	0.23	0.00	0.00
Tuesday	0.24	0.07	0.00
Wednesday	0.17	0.01	0.00
Thursday	0.13	0.75	0.00
Friday	0.13	0.12	0.81
Saturday	0.04	0.04	0.74
Time of first session:			
8:00 - 11:59	0.21	0.27	0.27
12:00 - 14:59	0.24	0.28	0.43
15:00 - 18:59	0.28	0.23	0.27
19:00 - 22:59	0.21	0.19	0.60
23:00 - 07:59	0.05	0.04	0.46
State anxiety:			
T1	36.37	39.64	0.03
	(12.51)	(12.13)	
T2	40.33	42.72	0.15
	(13.82)	(13.97)	
T3	41.53	43.64	0.13
	(12.51)	(11.02)	
task response (T2 - T1)	3.71	3.18	0.65
	(10.49)	(9.38)	
Task perception:			
relaxing	2.68	2.83	0.30
	(1.23)	(1.15)	
easy	3.36	3.56	0.20
	(1.26)	(1.33)	
stressful	2.94	2.88	0.72
	(1.31)	(1.31)	
difficult	2.66	2.51	0.33
	(1.27)	(1.25)	
enjoyable	2.88	2.85	0.87
	(1.13)	(1.19)	
successful	3.35	3.39	0.80
	(1.02)	(1.15)	
tiring	2.48	2.44	0.75
	(1.17)	(1.11)	
N	206	111	317

Table G.12: Balance of first session variables across email reminder delays - Part 2

	Email reminder delay		Difference
	1-3 days	4+ days	P-Value
Altruism task:			
Average donation	27.75	27.41	0.91
	(23.70)	(23.16)	
Risk taking task:			
Indifference point	42.47	39.74	0.25
	(20.74)	(19.23)	
Intertemporal substitution task:			
Delta (long-term discount rate)	0.89	0.90	0.77
	(0.14)	(0.11)	
Beta (present bias)	1.03	1.09	0.35
	(0.31)	(0.59)	
Effort task:			
Number of sliders	28.66	30.26	0.13
	(9.38)	(8.75)	
Incentive comprehension question:			
attempts until correct answer	1.91	1.53	0.13
	(2.57)	(1.78)	
Task comprehension questions:			
donation	0.26	0.27	0.98
	(0.44)	(0.44)	
risk taking	0.96	0.94	0.46
	(0.20)	(0.25)	
intertemporal substitution	0.97	0.96	0.91
	(0.18)	(0.19)	
N	206	111	317

Table G.13: Balance of demographic variables across email reminder delays in the selected sample

	Email reminder delay		Difference
	1-3 days	4+ days	P-Value
Age	23.73	23.09	0.41
	(6.49)	(6.12)	
Gender identity:			
female	0.65	0.68	0.56
male	0.35	0.29	0.33
other/unknown	0.01	0.03	0.20
Undergraduate student	0.72	0.73	0.83
	(0.45)	(0.45)	
Year of undergrad studies:			
freshman	0.02	0.01	0.54
	(0.14)	(0.10)	
sophomore	0.07	0.14	0.12
	(0.26)	(0.34)	
junior	0.22	0.17	0.30
	(0.41)	(0.37)	
senior	0.41	0.42	0.87
	(0.49)	(0.50)	
Subject area:			
business	0.34	0.35	0.81
science/engineering	0.40	0.33	0.23
arts/humanities/social sciences	0.14	0.20	0.24
health	0.09	0.08	0.71
other/unknown	0.02	0.04	0.51
Out of state	0.35	0.45	0.09
	(0.48)	(0.50)	
International student	0.14	0.23	0.06
	(0.35)	(0.42)	
Living on campus	0.16	0.14	0.53
	(0.37)	(0.34)	
Monthly income:			
$0 - $599	0.25	0.22	0.61
$600 - $999	0.16	0.18	0.66
$1000 - $1999	0.33	0.36	0.66
$2000 and higher	0.26	0.24	0.71
Employment during term time:			
none	0.36	0.36	0.99
part-time	0.48	0.51	0.65
full-time	0.16	0.13	0.53
Employed during summer break	0.49	0.45	0.57
	(0.50)	(0.50)	
Exam accommodations	0.06	0.06	0.92
	(0.23)	(0.24)	
Previous experiment participation	0.62	0.72	0.09
	(0.49)	(0.45)	
N	165	104	269

Table G.14: Correlations between Inverse Mills Ratio and regressors of interest

	(1) Task: stressful IMR	(2) Task: successful IMR	(3) STAI Response IMR
Stress	−0.479	−0.475	−0.501
Failure	0.125	0.132	0.092
Stress * Failure	−0.199	−0.191	−0.223
Success	−0.047	−0.050	−0.026
Stress * Success	−0.215	−0.217	−0.222
Observations	256	255	251
Participants	256	255	251

Note: The above matrix displays pairwise correlation coefficients between the Inverse Mills Ratio (IMR) and the regressors of interest. They were computed using the selected sample only.

Table G.15: Correlations between Inverse Mills Ratio and regressors of interest

	(1) Perceived performance IMR	(2) Expected token loss IMR
stress	0.042	0.042
Failure	0.101	0.101
Stress * Failure	0.039	0.039
Success	−0.041	−0.041
Stress * Success	0.007	0.007
Observations	266	266
Participants	266	266

Note: The above matrix displays pairwise correlation coefficients between the Inverse Mills Ratio (IMR) and the regressors of interest. They were computed using the selected sample only.