Appendix A. Supplementary Methods

A.1. Participants

One hundred twenty healthy males participated in this study. Due to subject withdrawal (n=2) and technical issues (n=1), three participants were excluded from the current analyses (delayed-stress (n=1), no-stress (n=2)). All analyses were performed on data from the remaining 117 participants. Participants in the experimental groups did not differ in age, BMI, education, marital status, profession, disturbed sleep rhythm, alcohol and tobacco use (Table A.1). There were differences in recreational drug use (F(2)=3.448, p=.035). Bonferroni post-hoc tests revealed that drug use was more prevalent in the no-stress group, compared to the immediate-stress group (p=.033). Participants were recruited via online platforms and flyers at the university campus. Participants received €48 for their participation.

Table A.1 Demographics

		Delayed-stress group (n=35)	Immediate-stress group (n=42)	No-stress group (n=40)[b]	Statistics [a]
Age (mean, SD)		23.90 (4.89)	26.66 (9.53)	23.93 (3.48)	F(2)=2.303, p=0.105
Body Mass Index (mean, SD)		22.97 (2.65)	23.08 (2.79)	22.72 (1.68)	F(2)=0.225, p=0.799
Education (n)	High school	23	24	19	F(2)=0.859, p=0.426
	University of applied sciences	2	5	4	
	University	10	12	15	
	Other	-	1	1	
Marital status (n)	Single	20	26	23	F(2)=0.107, p=0.899
	Non-single	15	16	16	
Profession (n)	Student	30	26	30	F(2)=2.731, p=0.069
	Employed	1	7	5	
	Non-employed	3	5	1	
	Other	1	4	3	
Substance use (n)	Alcohol	33	38	33	F(2)=0.945, p=0.392
	Tobacco	5	7	8	F(2)=0.253, p=0.777
	Recreational drugs	14	13	23	F(2)=3.448, p=0.035*
Disturbed day-night or sleep rhythm (n)		3	1	3	F(2)=0.778, p=0.462

[a] Results of the one-way ANOVA test with experimental group as between-subject factor; [b] Demographic information of one subject was lost due to technical problems; * p < 0.05

A.2 Measures for stress (re)activity

Saliva samples were stored at -80 degrees Celsius before they were delivered in batches to the University Medical Center Utrecht LKHC laboratory for biochemical analysis. A Beckman-Coulter AU5811 chemistry analyzer was used to measure sAA (Beckman-Coulter Inc, Brea, CA). An in-house competitive radio-immunoassay employing a polyclonal anticortisol-antibody (K7348), with [1,2-3H(N)]-Hydrocortisone (PerkinElmer NET396250UC) as a tracer, was performed to measure cortisol without extraction (lower detection limit 1.0 nmol/l).

A.3 Fear Generalization Task

A.3.1 Task Stimuli

The US was applied via a bar electrode on the dominant inner forearm. US-intensity did not differ between experimental groups (F(2)=0.462, p=0.632), nor did subjective shock-ratings (F(2)=1.054, p=0.352). Experimental groups did not differ in their ability to identify the CUE and CTX stimuli combinations that predicted the US (shock-prediction) after the experiment (F(2)=1.532, p=0.223). Subjective sound-ratings did not differ between experimental groups at the start of the acquisition phase (F(2)=0.823, p=0.442), at the end of the acquisition phase (F(2)=0.585, p=0.559), at the start of the surprise test phase (F(2)=1.259, p=0.288), or at the end of the test phase (F(2)=0.628, p=0.536).

A.3.2 Task phases

The NAh-probes prior to the acquisition phase were delivered with random 9s, 11, or 13s intervals (no more than two repetitions of identical intervals in successive trials). Before the acquisition phase, six additional context-habituation trials were presented after noise-habituation (i.e. all three CTX stimuli were shown twice for 10s, inter-trial-interval (ITI): 8-10s), to reduce reactivity to novelty (data not further analyzed. Over the course of these trials, only 1 startle probe was used during ITI (ITI-probe). For each subject the acquisition phase started with a safe trial, followed by a threat trial. Thereafter the trial type order was randomized in blocks of two trials consisting of one CTX- and one CTX+. Consequently, no more than two of the same consecutive trial types could be presented. In each trial, the CTX was present for 15s. The CUE appeared randomly after 6 to 9s as partial overlay of the CTX for 5 s (maximal two successive trials with the same onset time). One second before CUE off-set a startle probe was delivered to measure cue conditioning (i.e. CUE-probe). In each 4-trial block, two additional startle probes (1 CTX- and 1 CTX+) were offered 3s prior to CUE presentation to measure context conditioning (i.e. CTX-probe). ITI was jittered (8-10s, maximal two similar in consecutive trials) and one ITI-probe was delivered randomly per four trials (halfway ITI duration). After twenty-four hours, the sequence of noise alone habituation trials (Nah-probe) was repeated and immediately followed by the surprise test phase of the FGT. The presentation order was fixed for the first three trials (G-CTX, CTX-, CTX+), but shuffled in blocks of 3 for the subsequent trials. As in the acquisition phase, CTX's were shown for 15s and CUE's for 5s (starting 6-9s after CTX onset). The CUE-probe was also presented 1s before CUE off-set, but no US was applied throughout the test phase. Per block of six trials, three context-probes were delivered in shuffled sequence (3s prior to CUE-onset and equally divided over CTX+, CTX- and G-CTX). Also, one ITI-probe was presented during each 6-trial block (halfway ITI duration) and ITI was variable (8-10s).

A.3.3 Measurement and Preprocessing of the Fear-potentiated startle (FPS)

FPS eyeblink responses were measured with electromyography (EMG), via a Biopac MP150 system (Biopac Systems Inc, RRID:SCR_014829) (sample rate 1000 Hz) and a pair of 4mm Ag-AgCL electrodes, placed over the orbicularis oculi muscle of the left eye[1,2]. EMG data was recorded in μV and off-line pre-processed according to published guidelines[1,2]. In brief, the signal was stopband (50 Hz) and bandpass (28-500 Hz) filtered using a 4th order Butterworth filter, rectified and smoothed (time constant = 10 ms), using a custom-built Matlab script (MATLAB, RRID:SCR_001622). For each startle probe, the resulting signal was segmented into a baseline period (50 ms pre-probe), a response onset window (20 – 120 ms post-probe) and a response peak window (20 – 150 ms post-probe). Response magnitudes were defined as a baseline-to-peak difference, i.e. mean activity during the baseline period was subtracted from the highest value in the response peak window. To reduce movement and spontaneous blinking artifacts, trials were considered invalid and scored as missing values if there was excessive activity (more than two standard deviations above the subject's mean baseline activity) in the pre-probe period[3], this led to 2.28% missing values. When there was less then 10% increase in standard deviation during a trial (compared to mean baseline SD of that subject), trials were scored as null-responses[3]. In total, there were 2.14% null-responses in the dataset.

A.4 Experimental procedures

Participants visited the lab twice (24 hours apart) and all experimental procedures were performed in the afternoon, when cortisol levels are relatively low due to circadian rhythmicity[4,5]. The first visit commenced with the collection of informed consent and the baseline measures. Subsequently participants were subjected to the first experimental intervention (which ended 160 minutes prior to fear acquisition). After approximately 120 minutes participants were exposed to the second experimental intervention that ended 30 minutes prior to the acquisition phase of the FGT. The next day participants came back to complete the surprise test phase of the FGT. At the end of day two, after completion of the experiment, participants were debriefed about the study aims and experimental groups. Importantly, the acquisition and test phase of the FGT were conducted in the same experimental room, that was only used for this task. Of note, the waiting period/questionnaire collection, placebo-TSST, and TSST all took place in different rooms. Because the study was part of a larger project, participants performed other tasks between T10-T11 and T13-T14 in a different experimental room (Figure 3).

A.5 Statistical analysis

One-way ANOVA's were performed in SPSS, version 25[6], to compare demographics, subjective shock- and sound-ratings, and shock-prediction between experimental groups. Significant differences were followed by Bonferroni-corrected post-hoc tests. The effect of experimental group on shock-intensity was tested with a one-way ANOVA in R[7].

Two sAA data-points (of a total of 1400) and one cortisol data-point (of 1400) were missing and excluded from the analyses. One participant (no-stress group) was excluded from the cortisol analysis, due to high values that were most likely caused by corticosteroid use (Cooks distance for this participant was .71, well above the cutoff of .40 for influential points [8]). No participants were excluded from sAA analysis.

For the FPS responses to NAh-, CUE-, and CTX-probes in the FGT, individual data-points (i.e. s) were imputed, if 1/3 (or more) data-points for that probe category were present for a subject. For ITI-probes, means were calculated based on imputed data-points (if 2/3 (or more) data-points were present), or imputed directly (if 1/3 (or more) data-points were missing). The imputation was done using the predictive mean matching algorithm of the Multivariate Imputation by Chained Equations (MICE) package in R[9], with 100 multiple imputations using 50 iterations. The miceadds-package was also used[10]. Imputation reduced the level of missing values to 7.13%. In the FPS analysis, potential influential cases, based on cook's distance, were treated conservatively (i.e. no participants were excluded from the main FPS analyses). Sensitivity analyses were performed to check if the 'influential' participants changed the results significantly. This was only the case for the mean FPS responses to ITI-probes, were exclusion of three influential cases (based on Cook's distance) revealed a significant influence of group (Dm=3.425, rm=.047, df1=2, df2=93129.681, p=.033). Since these participants were no influential cases on the other analyses (CUE, CTX, NAh), we did not follow up this effect.

Endocrine and FGT data were also analyzed with R, by means of linear mixed-effect models (LMMs). LMM assumptions were checked using the influence.ME[8] and moments[11] packages for R. The lme4[12], LMERConvienceFunction[13] and lmerTest[14] packages for R were used to fit and test the LMMs. P-values were obtained by Wald tests[15] of the full model where the effect in question was compared against the model without the effect (α=0.05). The emmeans package[16] for R was used to 1) perform Tukey adjusted post-hoc pairwise comparisons on the endocrine data and 2) calculate Estimated Marginal Means (EMM) to follow-up significant effects on FPS responses, within each imputed dataset. EMM parameters were subsequently pooled according to Rubins rules[17,18]. To date, there are no statistical methods implemented in the major software packages to pool the results from post-hoc tests on EMMs of imputed datasets[19]. Therefore, differences in EMMs could not be tested statistically in the current study. Note: Trialnumber and Trialtype were included as categorial factors in the reported FGT analyses, to test the influence of experimental group on each trial and stimulus type. It has been proposed that entering these factors as continuous variables offers the opportunity to study generalization gradients in more detail[20,21]. LMM analyses with Trialnumber and Trialtype as continuous variables revealed similar results in our dataset and are therefore not reported. Figures were made with the ggplot2[22] and ggpubr[23] packages for R.

References Supplementary Methods

1. Blumenthal TD, Cuthbert BN, Filion DL, et al. Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology* 2005; 42: 1–15.
2. Lonsdorf TB, Menz MM, Andreatta M, et al. Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neurosci Biobehav Rev* 2017; 77: 247–285.
3. Klumpers F, Morgan B, Terburg D, et al. Impaired acquisition of classically conditioned fear-potentiated startle reflexes in humans with focal bilateral basolateral amygdala damage. *Soc Cogn Affect Neurosci* 2015; 10: 1161–1168.
4. Dickerson SS, Kemeny ME. Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research. *Psychol Bull* 2004; 130: 355–391.
5. Kudielka BM, Schommer NC, Hellhammer DH, et al. Acute HPA axis responses, heart rate, and mood changes to psychosocial stress (TSST) in humans at different times of day. *Psychoneuroendocrinology* 2004; 29: 983–992.
6. IBM. IBM SPSS Advanced Statistics 25. *Ibm*. Epub ahead of print 2017. DOI: 10.1080/02331889108802322.
7. R core team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*, http://www.r-project.org/ (2018).
8. Nieuwenhuis R, Te Grotenhuis M, Pelzer B. influence.ME: Tools for Detecting Influential Data in Mixed Effects Models. *R J* 2012; 4: 38–47.
9. Van Buuren S, Groothuis-Oudshoorn K. Multivariate Imputation by Chained Equations. *J Stat Softw*. Epub ahead of print 2011. DOI: 10.1177/0962280206074463.
10. Robitzsch A, Grund S, Henke T. miceadds: Some additional multiple imputation functions, especially for mice. R package version 2.14-26.
11. Komsta L, Novomestky F. moments: Moments, cumulants, skewness, kurtosis and related tests. R package version 0.14.
12. Bates D, Mächler M, Bolker B, et al. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015; 67: 1–48.
13. Tremblay A, Ransijn J. LMERConvenienceFunctions: Model Selection and Post-hoc Analysis for (G)LMER Models. R package version 2.10.
14. Kuznetsova A, Brockhoff P, Christensen R. lmerTest Package: Tests in Linear Mixed Effects Models. *J Stat Softw* 2017; 82: 1–26.
15. Meng XL, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*. Epub ahead of print 1992. DOI: 10.1093/biomet/79.1.103.
16. Lenth R. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.2.3.
17. Rubin DB. Inference and missing data. *Biometrika*. Epub ahead of print 1976. DOI: 10.1093/biomet/63.3.581.
18. Marshall A, Altman DG, Holder RL, et al. Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. *BMC Med Res Methodol*. Epub ahead of print 2009. DOI: 10.1186/1471-2288-9-57.
19. van Ginkel JR, Kroonenberg PM. Analysis of Variance of Multiply Imputed Data. *Multivariate Behav Res* 2014; 49: 78–91.
20. Vanbrabant K, Boddez Y, Verduyn P, et al. A new approach for modeling generalization gradients: a case for hierarchical models. *Front Psychol* 2015; 6: 652.
21. McGlade AL, Zbozinek TD, Treanor M, et al. Pilot for novel context generalization paradigm. *J Behav Ther Exp Psychiatry* 2019; 62: 49–56.
22. Wickham H. ggplot2: Elegant Graphics for Data Analysis.
23. Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.1.8.