

# Affective flexibility without perceptual awareness

Philipp Homan<sup>1</sup>, H. Lee Lau<sup>1</sup>, Ifat Levy<sup>2</sup>, Candace M. Raio<sup>3</sup>, Dominik R. Bach<sup>4,5</sup>,  
David Carmel<sup>6,7\*</sup>, Daniela Schiller<sup>1\*</sup>

<sup>1</sup>Department of Psychiatry, Department of Neuroscience, and Friedman Brain Institute,  
Icahn School of Medicine at Mount Sinai  
1470 Madison Ave, NY, NY 10029, USA

<sup>2</sup>Departments of Comparative Medicine, Neuroscience and Psychology  
Yale University, New Haven, CT, USA

<sup>3</sup>Department of Psychology, New York University  
6 Washington Place, NY, NY 10003, USA

<sup>4</sup>Clinical Psychiatry Research  
Department of Psychiatry, Psychotherapy, and Psychosomatics  
University of Zurich, Zurich, Switzerland

<sup>5</sup>Wellcome Centre for Human Neuroimaging,  
12 Queen Square, London WC1N 3BG, UK

<sup>6</sup>Psychology Department, University of Edinburgh,  
7 George Square, Edinburgh EH8 9JZ, UK

<sup>7</sup>School of Psychology, Victoria University of Wellington,  
Easterfield Building, Kelburn Parade, Wellington 6012, New Zealand

\*These authors contributed equally; address correspondence to  
[daniela.schiller@mssm.edu](mailto:daniela.schiller@mssm.edu) or [david.carmel@vuw.ac.nz](mailto:david.carmel@vuw.ac.nz)

## 1 **Abstract**

2 In an ever-changing environment, survival depends on learning which stimuli represent threat, and also  
3 on updating such associations when circumstances shift. Humans can acquire physiological responses to  
4 threat-associated stimuli even when they are unaware of them, but the role of awareness in updating  
5 threat contingencies remains unknown. This complex process – generating novel responses while  
6 simultaneously suppressing learned ones – relies on distinct neural mechanisms from initial learning,  
7 and has only been shown with awareness. Can it occur unconsciously? Here we show that it can.  
8 Participants underwent classical threat conditioning to visual stimuli that were suppressed from their  
9 awareness. One of two images was paired with an electric shock; halfway through the experiment,  
10 contingencies were reversed and the shock was paired with the other image. We found that physiological  
11 responses reflected changes in stimulus-threat pairings independently of stimulus awareness,  
12 demonstrating the sophistication of unconscious affective flexibility.

13

## 14 **Introduction**

15 Flexible responses to environmental threats are essential for adaptive behavior. Cues that predict threat  
16 constantly change - new threats may arise while old ones cease to pose a risk. When consciously  
17 perceiving such cues, we are able to flexibly update and shift threat responses from one cue to another  
18 (1-3). But can we update our reaction to stimuli that predict danger when we are not aware of them? It is  
19 known that threat-conditioned stimuli that are perceived without awareness can still elicit defensive  
20 physiological reactions (4-7), and that new threat associations can be formed through classical  
21 conditioning even without any awareness of the conditioned stimuli (8-10). Updating threat associations  
22 when contingencies change, however, is an entirely different matter: it involves a complex process of  
23 creating novel responses while simultaneously suppressing acquired ones. To date, such updating has  
24 only been shown in humans who were aware of the stimuli (2), and in animals under conditions where  
25 stimuli were fully available for perceptual processing (11); these studies have shown, furthermore, that  
26 the neural substrates of threat updating differ from those of the initial learning. It is thus unknown  
27 whether the sophisticated re-evaluation involved in such affective flexibility requires awareness, or can  
28 be accomplished without it. Here we show that it can, and furthermore, that stimulus awareness does not  
29 seem to play a substantial role in such affective flexibility.

30 To examine this, we employed the reversal paradigm, a laboratory model that requires flexible  
31 updating of threat contingencies (2). In an initial acquisition phase, participants encounter two  
32 conditioned stimuli (CSs) and learn that only one of them predicts an electric shock. Halfway through  
33 the experiment, with no warning, these contingencies flip, initiating the reversal phase: Participants must  
34 flexibly learn that the formerly safe CS now predicts the shock and that the old one no longer does. To

35 assess learning, participants' physiological arousal is recorded throughout the experiment, typically (and  
36 here) by measuring their skin conductance responses. Appropriate response reversal requires a  
37 sophisticated form of updating, in that one must learn to respond to a cue that now predicts threat while  
38 simultaneously inhibiting responses to the previously threatening cue that is now safe.

39 To see whether reversal of conditioned threat requires awareness, we had a large group of  
40 participants ( $N = 86$ ) undergo reversal learning with the CSs suppressed from awareness by continuous  
41 flash suppression (CFS), a technique commonly used to examine unconscious perception (10, 12-14):  
42 The CSs were visual images presented monocularly, while the other eye was shown a high-contrast,  
43 dynamic image (the CFS mask) at the corresponding retinal location (See Figure 1 for a description of  
44 the design and procedure).

45 CFS can suppress images from awareness for several seconds. However, it is also known that its  
46 effectiveness may vary across trials and individuals, and the suppressed stimulus may "break through"  
47 the suppression (15). Over the last decade, a growing body of work has raised concerns that the standard  
48 approach - removing from analysis data (participants and trials) in which breakthrough had occurred -  
49 may bias the findings (16, 17; See Supplementary Methods for further details of these issues.) Here, we  
50 adopt a number of methodological approaches to ensure our results are robust to these potential  
51 concerns.

52 Specifically, we remove no data and instead incorporate individual levels of reported stimulus  
53 awareness, as well as response patterns that might reflect residual awareness, into a regression model  
54 accounting for physiological responses. The model also adjusts for baseline anxiety (which has been  
55 previously shown to correlate with unconscious learning; (10)). Additionally, we use a Bayesian

56 approach to establish that a model in which participants were updating their learning provides a better  
57 account for the findings than a model in which they were simply (and independently of the stimulus)  
58 predicting the probability of a shock on the next trial (18). Finally, in order to verify that our procedure  
59 is able to induce reversal learning when participants are awareness of the stimuli, we ran a no-CFS  
60 group ( $N = 12$ ), in which participants also viewed the CSs monocularly (as the CFS group did), but were  
61 aware of them as no CFS masks were presented to their other eye.

62 We hypothesized that physiological responses to threat can be flexibly reversed without perceptual  
63 awareness. We find that reversal indeed occurs independently of CS awareness, and that there is strong  
64 evidence for the reversal of threat learning even in its complete absence.

## 65 **Results**

### 66 **Overall assessment of physiological reversal learning**

67 To assess the physiological arousal evoked by CSs, we used a model-based approach (19) to estimate  
68 the amplitude of anticipatory sudomotor nerve activity (SNA) from skin conductance data recorded  
69 during stimulus presentation. A variational Bayes approximation was employed to invert a forward  
70 model that describes how hidden SNA translates into observable SCRs (see Materials and Methods).  
71 Previous work has shown that this approach is more sensitive than conventional SCR peak-to-peak  
72 analysis (19-21). Figure 2A shows the time course of evoked SNA to Spiders A and B, separately for the  
73 CFS and no-CFS groups. In both groups, responses to Spider A relative to Spider B were larger during  
74 the acquisition phase and smaller during the reversal phase. To quantify the magnitude of physiological  
75 reversal learning, we calculated a reversal learning index for each participant (see Materials and

76 Methods). The reversal learning index was positive and significantly greater than zero for both the CFS  
77 and no-CFS groups (Figure 2B). A linear mixed model (see Materials and Methods for details) revealed  
78 a significant interaction of stage and spider in both groups (CFS:  $\beta = 0.27$ ,  $t(2935) = 4.23$ ,  $P = < 0.001$ ;  
79 no-CFS:  $\beta = 1.23$ ,  $t(2935) = 7.29$ ,  $P = < 0.001$ ). Note that a significant interaction is formally equivalent  
80 to a significant reversal learning index. On its own, however, it simply reveals a difference in the  
81 comparative magnitude of responses to the two CSs across the two halves of the experiment; follow-up  
82 tests show that this difference is indeed due to reversal: Spider A evoked greater responses than Spider  
83 B in the acquisition phase (CFS:  $t(341.9) = 3.0$ ,  $P = 0.003$ ; no-CFS:  $t(201.1) = 4.6$ ,  $P < 0.001$ ) and the  
84 pattern was reversed in the reversal phase (CFS:  $t(341.9) = 2.8$ ,  $P = 0.005$ ; no-CFS:  $t(341.9) = 3.6$ ,  $P =$   
85  $0.0003$ ). These results indicate that reversal learning was evident in both groups. Although Figure 2  
86 shows that it was more pronounced in the no-CFS group, we note that this difference is not  
87 straightforwardly interpretable because the no-CFS group (a control, intended to rule out an ineffective  
88 manipulation if no effect was found for the CFS group) was substantially smaller; furthermore, as  
89 addressed in detail below, suppression from awareness was very heterogenous in the CFS group.

90 As previous work has found a negative association between anxiety and threat acquisition with and  
91 without awareness (10), we also calculated correlations between the CFS group's baseline anxiety  
92 measures (STAIT, STAIS, FSQ) and the reversal learning index. Overall, reversal learning decreased  
93 significantly with increasing levels of state and trait anxiety, and to a lesser but non-significant extent  
94 for spider phobia (Figure 2C).

## 95 **Reversal learning and perceptual awareness**

96 The CFS manipulation reduced awareness of the CSs; as expected, however, it was differentially  
97 effective in doing so across participants, precluding an overall conclusion that all learning under CFS  
98 happened non-consciously. The CFS group showed significantly lower accuracy in response to the  
99 "which seen?" question ( $M = 0.46$ ,  $SD = 0.29$ ) compared to the no-CFS group ( $M = 0.86$ ,  $SD = 0.16$ ;  $t$   
100 ( $22.77$ ) =  $-7.24$ ,  $P < 0.001$ ), and accuracy in the CFS group was not significantly different from the 50%  
101 random-response level ( $t$  ( $85$ ) =  $-1.21$ ,  $P = 0.229$ ). The CFS group also showed lower confidence ( $M =$   
102  $1.73$ ,  $SD = 0.65$ ) than the no-CFS group ( $M = 2.83$ ,  $SD = 0.08$ ;  $t$  ( $95.38$ ) =  $-15.05$ ,  $P < 0.001$ ).

103 However, group differences in accuracy and confidence, and even random-level response accuracy,  
104 are not sufficient to establish an absence of perceptual awareness in the CFS group. Notably, average  
105 confidence of correct responses in this group was low but significantly greater than the minimum value  
106 of 1 ( $t$  ( $77$ ) =  $10.79$ ,  $P < 0.001$ ), suggesting that at least some participants were aware of some of the  
107 CSs; learning might thus have arisen from a subset of trials and/or participants where such awareness  
108 occurred. To address this, we quantified CS awareness by calculating an awareness index for each  
109 participant, ranging in possible values from 0 for no awareness to 1 for full awareness (see Materials and  
110 Methods). Although the awareness index of the CFS group ( $M = 0.28$ ,  $SD = 0.34$ ) was significantly  
111 lower than the no-CFS group's ( $M = 0.92$ ,  $SD = 0.18$ ;  $t$  ( $23.93$ ) =  $-10.19$ ,  $P < 0.001$ ), it was still  
112 significantly higher than zero ( $t$  ( $85$ ) =  $7.59$ ,  $P < 0.001$ ).

113 Therefore, in order to test our main hypothesis that the reversal of acquired threat responses can be  
114 achieved without perceptual awareness, we characterized the quantitative relation between the level of  
115 awareness and the magnitude of reversal learning in the CFS group. To control for possible artifacts of

116 regression to the mean (see Supplementary Methods), we first calculated the correlation between two  
117 independent estimates of the awareness index (16), one calculated from even-numbered trials, the other  
118 from odd-numbered trials. These measures were strongly correlated ( $r(84) = 0.96$ ,  $P < 0.001$ ; Figure  
119 3A); participants' awareness level in one set of trials was thus overwhelmingly predictive of their  
120 awareness in the other set. Thus, two independent measures of awareness showed very similar results  
121 which suggests that the overall awareness index was unlikely to be influenced by extreme values that  
122 were due to measurement-level noise. Such extreme values would have occurred in one but not the other  
123 measure and would have thereby attenuated the correlation between even and odd trials considerably.

124 Next, we examined the association between the awareness index and the reversal learning index,  
125 using values of both indices obtained separately from even (Figure 3B) and odd (Figure 3C) trials. As  
126 the color-coding of Figure 3 shows, the relation between individual participants' awareness and their  
127 reversal learning was highly consistent across these separate measurements. In light of this, we pooled  
128 the data from all trials and regressed the reversal learning index on the perceptual awareness index  
129 (Figure 3D). The parameter of interest was the intercept, which corresponds to the magnitude of reversal  
130 learning at zero perceptual awareness. The intercept was positive and significantly different from zero.  
131 Furthermore, the awareness index regressor did not contribute significantly to prediction of reversal  
132 learning; importantly, this finding was even stronger in models that accounted for STAIT scores and a  
133 binary factor indicating whether participants were tracking the stimuli with their responses (see  
134 Materials and Methods; Figure 3E and Table 1).



## 135 **Comparing learning and expectation-based accounts**

136 Well-controlled lab-based conditioning procedures require strict constraints that preclude complete  
137 randomization of the number and order of different CSs; this comes with a cost: participants are able to  
138 develop expectations with above-chance validity, based on the sequence of trials so far, about the  
139 likelihood of a shock on any upcoming trial (18). Even without any awareness of the CSs, a participant  
140 should have been able to distinguish two types of trials: reinforced (with shock) and non-reinforced (no-  
141 shock). In a study with two CSs and a 100% reinforcement rate like ours, such expectations would  
142 correspond to an anticipated pattern of alternating trial-types (shock/no-shock or vice versa), with an  
143 increase in shock anticipation after every no-shock trial. The question, therefore, was whether the  
144 physiological responses we had measured might simply reflect participants' pattern-based anticipation  
145 of shock, rather than learning of the contingencies associated with the CSs.

146 To answer this question, we used a Bayesian approach to compare the probability of our findings  
147 being accounted for by a classic Rescorla-Wagner learning model (22) and a trial-sequence model. We  
148 hypothesized that successful threat reversal without perceptual awareness should be better explained by  
149 the Rescorla-Wagner learning model, whereas simple pattern-based expectation would be better  
150 explained by the trial-sequence learning model. We used maximum likelihood estimation to assess the  
151 log likelihood and calculate the Bayesian Information Criterion (BIC) of each model (See Materials and  
152 Methods for details of each model and calculation of the BIC). A smaller BIC indicates a better model,  
153 and BIC values can thus be compared by calculating the difference between them and interpreting the  
154 resulting  $\Delta$  BIC as providing evidence against the higher BIC. The Rescorla-Wagner model (BIC:  
155 562.1) outperformed the pattern-based expectation model (BIC: 584.9), with the difference ( $\Delta$  BIC:

156 22.9) greater than 10, suggesting that the evidence against the trial switch model is very strong (23).  
157 Repeating this comparison for just the participants with zero mean awareness confirmed the lower BIC  
158 for the Rescorla-Wagner model (BIC: 114.3) compared to the pattern-based expectation model (BIC:  
159 125.7), with the difference again greater than 10 ( $\Delta$  BIC: 11.3; see also Figure S2). This model  
160 comparison provides convincing evidence that a classical Rescorla-Wagner learning model explains our  
161 findings better than an alternative expectation-based model.

## 162 **Discussion**

163 These results indicate that participants were able to update their defensive physiological responses  
164 independently of their awareness of threat-related cues. Previous studies have shown that new threat  
165 associations can be formed without perceptual awareness of the conditioned stimuli (5, 9-10). However,  
166 until now it was unknown whether the far more complex process of threat reversal - shifting reactions  
167 from a stimulus that no longer predicts danger to one that now does - can be accomplished without  
168 awareness. Our finding of reversal learning occurring independently of the level of perceptual awareness  
169 suggests that separate processes underlie affective flexibility and conscious processing (24). Conversely,  
170 the negative correlation between reversal learning and anxiety suggests that the various impairments  
171 caused by anxiety are not limited to systems underlying conscious processes.

172 Previous studies have pointed out the limitations of using accuracy and confidence measures to  
173 assess perceptual awareness, and suggested remedies including the calculation of metacognitive  
174 sensitivity measures (25), Bayesian statistics (26), or parametric variation of the experimental  
175 manipulation (27). The present study addresses an issue not covered in previous discussions, by showing

176 that a trial-wise analysis may reveal hints for incomplete suppression that analyses relying on average  
177 measures might easily miss. Future studies that rely on forced-choice questions for awareness  
178 assessment should thus examine response patterns across trials in addition to collecting aggregate  
179 measures.

180 Notably, a previous study (10) that used CFS to investigate acquisition of threat responses without  
181 awareness of the stimuli found that such acquisition can occur, but is rapidly forgotten. The present  
182 study again showed that such acquisition can occur (and, additionally, be reversed), but did not find the  
183 same rapid forgetting. The reasons for this are unclear, but we speculate that the difference may be due  
184 to specific aspects of the stimuli, design and procedure: our use of pictures of spiders (rather than faces)  
185 and a 100% (rather than 50%) reinforcement protocol may have altered the temporal characteristics of  
186 acquisition. Similarly, the temporal profile of reversal may change if the stimuli and reinforcement  
187 regime are different.

188 The present results add to a growing body of findings distinguishing functions that do and do not  
189 require awareness. Such distinctions are important in guiding research into the neural mechanisms of  
190 conscious and non-conscious processing. Previous research hints at the mechanism underlying the non-  
191 conscious affective flexibility reported here, although it remains to be elucidated: The ability to reverse  
192 conditioned responses depends on the integrity of circuitry spanning several neural regions, particularly  
193 the ventromedial prefrontal cortex (vmPFC) and its connections with the amygdala (1) where threat  
194 associations are formed (28). Consistent with this, it is known that patients with anxiety disorders often  
195 show rigid and inflexible threat responses in conjunction with prefrontal cortex dysfunction (29, 30).

196 Indeed, the real-life settings that people with anxiety disorders find challenging often require the  
197 updating and shifting of threat responses. Deficits in affective flexibility may thus explain the threat  
198 learning and extinction deficits seen in such disorders (31). Compared to healthy controls, patients are  
199 less able to distinguish between safe and unsafe stimuli in threat learning (when it is adaptive to do so),  
200 and distinguish between them to a greater extent during extinction (when it is non-adaptive). Threat  
201 learning without perceptual awareness is also negatively correlated with baseline state anxiety in healthy  
202 participants (10). Our new finding that baseline anxiety is negatively correlated with affective flexibility  
203 suggests a potential use for reversal learning as a model paradigm for investigating how anxiety  
204 modulates various processes in a variety of disorders, including, for example, posttraumatic stress  
205 disorder, in which there is an impairment of threat inhibition (32).

## 206 **Methods**

### 207 **Participants**

208 Ninety-eight healthy participants (mean age = 29.97; range 18-65) were assigned to one of the two  
209 groups: reversal learning with CFS (CFS group;  $N = 86$ , 48 female) or without CFS (no-CFS group;  $N =$   
210 12, 5 female). Assignment was random until each group reached a size of 12; subsequent participants  
211 were assigned to the CFS group. Measures of trait and state anxiety (Spielberger Trait-State Anxiety  
212 Inventory (33); STAIT and STAIS, respectively) and spider phobia (Fear of Spider Questionnaire; FSQ  
213 (34)) were taken prior to participation and did not differ between the groups (Table S1). The experiment  
214 was approved by the Institutional Review Board of the Icahn School of Medicine at Mount Sinai. All  
215 participants provided written informed consent and were financially compensated for their participation.

## 216 **Experimental procedure**

217 Participants viewed the stimuli monocularly, through a mirror stereoscope (StereoAids, Australia)  
218 placed at a distance of 45 cm from a 17-inch Dell monitor. The CSs (schematic low-contrast images of  
219 spiders), presented to the left eye only, were suppressed from awareness in the CFS group: while the left  
220 eye saw them, the right eye was presented with "Mondrians" - arrays of high contrast, multi-colored,  
221 randomly generated rectangles alternating at 10 Hz. Both the CSs and the CFS masks were flanked by  
222 identical textured black and white bars, to facilitate stable ocular vergence. The no-CFS group viewed  
223 identical CSs (also presented monocularly), but with no Mondrians presented to the other eye.

224 The experiment consisted of 16 acquisition trials followed by 16 reversal trials. One of two spider  
225 images was presented on each trial. The spider images were schematic and had similar low-level  
226 features. During acquisition, spider A always terminated with a shock and spider B never did. Reversal  
227 occurred halfway through the experiment: spider B now terminated with a shock and spider A did not.  
228 The spider stimuli were presented for 6 s each in pseudorandomized order. One of four possible trial  
229 orders was used for each participant. Orders were generated by imposing specific constraints on the trial  
230 order, such that the first trial was always reinforced and no more than two of the same trial type ever  
231 occurred consecutively.

232 Trial order and spider identity were counterbalanced across participants. To assess the effectiveness  
233 of the awareness manipulation (35), 1 s after the offset of every CS participants were shown the question  
234 "Which seen?" (1 = flower, 2 = spider; notably, flowers were never shown, meaning the question  
235 addressed detection rather than discrimination as it could be answered correctly even with a brief  
236 glimpse). This was followed by the question "How confident?" (1 = guess to 3 = sure; participants were

237 instructed to indicate how confident they were of the flower/spider answer they had just given). Both  
238 questions were presented binocularly (1.5 - 2 s each, during which responses had to be given by pressing  
239 number keys on a standard keyboard). The second question was followed by an 8 to 10 s inter-trial  
240 interval.

## 241 **Psychophysiological stimulation and measurement**

242 Mild electric shocks were delivered using a Grass Medical Instruments SD9 stimulator and stimulating  
243 bar electrode attached to the participant's right wrist. Shocks (200 ms; 50 pulse/s) were delivered at a  
244 level determined individually by each participant as "uncomfortable but not painful" (maximum of  
245 60V), during a work-up procedure prior to the experiment.

246 Skin conductance responses (SCR) were measured with Ag-AgCl electrodes, filled with standard  
247 isotonic NaCl electrolyte gel, and attached to the middle phalanges of the second and third fingers of the  
248 left hand. SCR signals were sampled continuously at a rate of 200 Hz, amplified and recorded with a  
249 MP150 BIOPAC Systems skin conductance module connected to a PC.

## 250 **Analysis of physiological responses**

### 251 **Model-based analysis**

252 We estimated SNA from SCR data with a model-based variational Bayes approximation (19), inverting  
253 a forward model that describes how (hidden) SNA translates into (observable) SCR. A unit increase in  
254 SNA corresponds to an increase in SCR of 1 micro Siemens. The model assumes that the observed SCR  
255 can be decomposed into different components including anticipation, evocation, and spontaneous

256 fluctuations, each of which are generated by bursts of SNA driven by changes in sympathetic arousal.  
257 The generative (forward) model thus describes how sympathetic arousal, the physiological measure that  
258 is taken as an index of the psychological process of threat, translates into sudomotor nerve bursts which  
259 then generate the observable SCR (19). Using Bayesian inference, the forward model can then be  
260 reversed in order to estimate the most likely underlying SNA given the observed SCR:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)},$$

261 (1)

262 where the most likely parameter vector  $\Theta$  (corresponding to the SNA) given the observed outcome  $y$   
263 (corresponding to the SCR) is given by the prior estimate of  $\Theta$  weighted by the likelihood of  $y$  given  $\Theta$ .  
264 Solving this equation involves integration over the model evidence  $p(y)$  which is analytically hard to  
265 compute (and possibly intractable). This can be resolved by replacing this integration problem by an  
266 optimization problem, which can be approximated with Variational Bayes procedures (36), where the  
267 log of the model evidence can be framed as the sum of the Kullback-Leibler divergence and the Free  
268 Energy. By maximizing the Free Energy the Kullback-Leibler divergence is minimized, and a lower  
269 bound to the log model evidence can be derived iteratively.

270 The SNA estimates were computed using previously developed software package PsPM (19)  
271 implemented in MATLAB R2016b (The Mathworks Inc, Natick, MA, USA). The statistical analyses  
272 were conducted with the R software (37) (R version 3.4.2 (2017-09-28)) and the libraries lme4 (38) and

273 lsmeans (39). Welch's t-tests were used instead of two sample t-tests when groups had unequal  
274 variances.

## 275 **Reversal Learning Index**

276 An estimate of SNA was obtained for each trial. We expected Spider A to evoke greater SNA than  
277 Spider B during the acquisition phase, and Spider B to evoke greater SNA than Spider A during the  
278 reversal phase. The strength of reversal learning can thus be quantified by calculating, separately for the  
279 acquisition and reversal phases, the difference between the average SNA evoked by each spider. To  
280 quantify the degree of reversal (which is formally equivalent to the interaction of phase and stimulus),  
281 the reversal learning index was calculated by subtracting the difference between mean SNAs evoked by  
282 each spider during reversal from the difference during acquisition (the larger the index, the greater the  
283 magnitude of reversal learning):

$$\begin{aligned} \text{Reversal learning index} &= \Delta\text{Acquisition} - \Delta\text{Reversal} \\ \Delta\text{Acquisition} &= [\text{mean}(\text{Spider A}) - \text{mean}(\text{Spider B})]_{\text{Acquisition}} \\ \Delta\text{Reversal} &= [\text{mean}(\text{Spider A}) - \text{mean}(\text{Spider B})]_{\text{Reversal}} \end{aligned}$$

284 (2)

285 To formally test for group differences in the strength of reversal learning, we computed a linear  
286 mixed model using the lme4 library in R. We used the skin conductance response (converted to a model-  
287 based measure of sudomotor nerve activity, SNA) as the dependent variable and entered group (CFS,



288 no-CFS), stage (acquisition, reversal), and spider (spider A, spider B) as well as a continuous variable  
289 for trial (to account for habituation) as predictors. The random structure of the model included an  
290 intercept and slopes for stage and spider.

## 291 **Assessments of perceptual awareness**

### 292 **Perceptual awareness index**

293 To characterize participants' reported awareness of CSs, each trial was assigned a perceptual awareness  
294 score, defined by a combination of detection and confidence responses: Correct answers with a  
295 confidence rating of 1 (guess) and incorrect answers irrespective of confidence were assigned an  
296 awareness score of 0; correct answers with a confidence rating of 2 (medium) were assigned a score of  
297 0.5, and correct answers with a confidence rating of 3 (high) were assigned an awareness score of 1. A  
298 perceptual awareness index was calculated for each participant by averaging awareness scores across all  
299 trials.

### 300 **Stimulus-response association patterns ("tracking")**

301 We also assessed response patterns across trials, to see whether participants were able to track stimuli  
302 with their responses, accurately discriminating the images despite not being able to label them. We  
303 plotted individual trial-by-trial responses to the question "Which seen?", overlaid on the trial-by-trial  
304 presentation of spiders (spider A, spider B; Figure S1A). We then calculated the number of consecutive  
305 "hits", defined as the number of consecutive trials where these two time-courses were either identical or  
306 consistently in opposition, suggesting that there was a possible association between the stimulus and the

307 response during those trials. The probability of such consecutive hits occurring by chance alone can be  
308 derived as follows:

309 Let  $p = 0.5$  be the probability of a hit,  $k$  the number of consecutive hits,  $n$  the number of trials left,  $i$   
310 the number of consecutive hits already observed; the chance of observing  $k$  consecutive hits for the  
311 remaining  $n$  trials can then be formulated as a recursive problem:

$$f_{p,k}(i, n) = pf_{p,k}(i + 1, n - 1) + (1 - p)f_{p,k}(0, n - 1),$$

312 (3)

313 which can be solved analytically with dynamic programming or recursion. Trivially,  $f_{p,k}(k, n) = 1$  for  $n$   
314  $\geq 0$  since  $k$  consecutive hits have already been observed, and  $f_{p,k}(i, n) = 0$  for  $k - i > n$  since there are  
315 not enough trials left to observe  $k$  consecutive hits.

316 For example, assuming we want to know how likely it is to observe  $k = 8$  consecutive hits within  $n =$   
317 32 trials given  $p = 0.5$ , i.e.,  $f_{0.5,8}(0, 32)$ , we find that this yields a probability of 0.050.

318 Alternatively, the probability can be derived by simulation for all possible numbers of consecutive  
319 hits within 32 trials (i.e., from 1 to 31). For each possible number, we thus also simulated  $10^5$  draws of a  
320 binomial distribution and calculated the average probability of that number of hits being consecutive. As  
321 can be seen in Figure S1B, the result for 8 consecutive hits (0.04991) was very close to the analytical  
322 solution. Fifteen participants showed evidence of tracking the spiders or the shocks with their responses  
323 (8 or more consecutive hits); notably, 3 of these participants appeared to have a perceptual awareness

324 index of zero. We thus adjusted our subsequent analysis with an additional binary covariate, indicating  
325 whether participants did or did not show 8 or more consecutive hits.

## 326 **Comparing learning and expectation-based models**

327 The Rescorla-Wagner model (22) describes how the prediction for each trial is updated according to a  
328 prediction error and learning rate:

$$\begin{aligned} V_{n+1}(x_n) &= V_n(x_n) + \alpha \delta_n \\ \delta_n &= r_n - V_n(x_n), \end{aligned}$$

329 (4)

330 where  $x_n$  is the conditioned stimulus on trial  $n$  (Spider A or Spider B), and  $\delta_n$  is the punishment  
331 prediction error that measures the difference between the expected and the actual shock ( $r_n$ ) on trial  $n$ .  
332 The learning rate  $\alpha$  for the value update is a constant free parameter. The value for the CS not observed  
333 on trial  $n$  remains unchanged. To derive the best fits for the Rescorla-Wagner model, we assumed that  
334  $V_0 = 0.5$ , reflecting the assumption that getting a shock or not was equally likely for the first trial.

335 For the alternative trial-sequence learning model, we assumed that a participant expecting a strict  
336 sequence of alternating trial types (shock/no shock or vice versa) would update this expectation  
337 according to the actually encountered trial types and a constant learning rate:

$$\begin{aligned}V'_{n+1} &= V'_n + \alpha' \delta'_n \\ \delta'_n &= r'_n - V'_n \\ \tau_n &= |(r'_{n-1} - 1)|,\end{aligned}$$

338 (5)

339 where  $V'_{n+1}$  is the expected trial type switch at trial  $n+1$  (if  $V'_{n+1}$  is larger than 0.5, a trial switch is  
340 expected),  $\alpha'$  is the learning rate, and  $\delta'_n$  is the prediction error. The prediction error corresponds to the  
341 difference between the actual trial type switch for trial  $n$  ( $r'_n$ ; coded as one for a trial type switch and  
342 zero for an equal trial type) and the expectation for trial  $n$ . A changing trial type for trial  $n$  was tracked  
343 by  $\tau_n$ , which was one if the preceding trial was zero and zero if the preceding trial type was one. To map  
344 these expectations onto expected values, we assumed that

$$V_{n+1} = \begin{cases} V'_{n+1} \cdot \tau_n + (1 - V'_{n+1})(1 - \tau_n), & \text{if } V' > 0.5 \\ V'_{n+1}, & \text{otherwise,} \end{cases}$$

345 (6)

346 where the expected value for trial  $n+1$  was calculated according to whether a trial type switch was  
347 expected ( $V' > 0.5$ ) or not.

348 We performed a formal model comparison between the conventional Rescorla-Wagner model and  
349 the trial switch model for our data set (Figure S2), using maximum likelihood estimation and non-linear  
350 optimization (implemented with the `fmincon` function in MATLAB R2016b (The Mathworks Inc,

351 Natick, MA, USA). Using the log likelihood, we calculated the Bayesian Information Criterion (BIC) to  
352 compare the two models as follows:

$$\text{BIC} = \log(n) k - 2 \cdot \log(\hat{L}),$$

353 (7)

354 where  $n$  is the number of data points,  $k$  is the number of regressors, and  $\hat{K}$  is the maximized value of the  
355 likelihood function.

356

## 357 **References**

- 358 1. Morris JS, Dolan RJ (2004) Dissociable amygdala and orbitofrontal responses during reversal fear  
359 conditioning. *Neuroimage* 22(1):372–80.
- 360 2. Schiller D, et al. (2008) From fear to safety and back: Reversal of fear in the human brain. *Journal*  
361 *of Neuroscience* 28(45):11517–25.
- 362 3. Fleming SM, Dolan RJ, Frith CD (2012) Metacognition: computation, biology and function.  
363 *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*  
364 367(1594):1280–6.
- 365 4. Critchley HD, Mathias CJ, Dolan RJ (2002) Fear conditioning in humans: the influence of  
366 awareness and autonomic arousal on functional neuroanatomy. *Neuron* 33(4):653–63.
- 367 5. Morris JS, Ohman A, Dolan RJ (1998) Conscious and unconscious emotional learning in the  
368 human amygdala. *Nature* 393(6684):467–70.
- 369 6. Ohman A, Soares JJ (1994) "unconscious anxiety": phobic responses to masked stimuli. *Journal*  
370 *of Abnormal Psychology* 103(2):231–40.
- 371 7. Whalen PJ, et al. (1998) Masked presentations of emotional facial expressions modulate amygdala  
372 activity without explicit knowledge. *Journal of Neuroscience* 18(1):411–8.
- 373 8. Katkin ES, Wiens S, Ohman A (2001) Nonconscious fear conditioning, visceral perception, and  
374 the development of gut feelings. *Psychol Sci* 12(5):366–70.
- 375 9. Manns JR, Clark RE, Squire LR (2002) Standard delay eyeblink classical conditioning is  
376 independent of awareness. *Journal of Experimental Psychology: Animal Behavior Processes*  
377 28(1):32–7.

- 378 10. Raio CM, Carmel D, Carrasco M, Phelps EA (2012) Nonconscious fear is quickly acquired but  
379 swiftly forgotten. *Current Biology* 22(12):R477–9.
- 380 11. Izquirdo A, Brigman JL, Radke AK, Rudebeck PH, Holmes A (2017) The neural basis of reversal  
381 learning: An updated perspective. *Neuroscience* 345:12–26.
- 382 12. Tsuchiya N, Koch C (2005) Continuous flash suppression reduces negative afterimages. *Nature*  
383 *Neuroscience* 8(8):1096–101.
- 384 13. Stein T, Hebart MN, Sterzer P (2011) Breaking continuous flash suppression: A new measure of  
385 unconscious processing during interocular suppression? *Frontiers in Human Neuroscience* 5:167.
- 386 14. Carmel D, Arcaro M, Kastner S, Hasson U (2010) How to create and use binocular rivalry. *J Vis*  
387 *Exp* (45).
- 388 15. Gayet S, Stein T (2017) Between-subject variability in the breaking continuous flash suppression  
389 paradigm: Potential causes, consequences, and solutions. *Frontiers in Psychology* 8:437.
- 390 16. Shanks DR (2016) Regressive research: The pitfalls of post hoc data selection in the study of  
391 unconscious mental processes. *Psychonomic Bulletin & Review* 24:752–775.
- 392 17. Stein T, Sterzer P (2014) Unconscious processing under interocular suppression: getting the right  
393 measure. *Frontiers in Psychology*, 5, 387.
- 394 18. Wiens S, Katkin ES, Ohman A (2003) Effects of trial order and differential conditioning on  
395 acquisition of differential shock expectancy and skin conductance conditioning to masked stimuli.  
396 *Psychophysiology* 40(6):989–97.
- 397 19. Bach DR, Daunizeau J, Friston KJ, Dolan RJ (2010) Dynamic causal modelling of anticipatory  
398 skin conductance responses. *Biological Psychology* 85(1):163–70.

- 399 20. Bach DR (2014) A head-to-head comparison of scralyze and ledalab, two model-based methods  
400 for skin conductance analysis. *Biological Psychology* 103:63–8.
- 401 21. Staib M, Castegnetti G, Bach DR (2015) Optimising a model-based approach to inferring fear  
402 learning from skin conductance responses. *Journal of Neuroscience Methods* 255:131–8.
- 403 22. Rescorla R, Wagner A (1972) *A theory of Pavlovian conditioning: Variations in the effectiveness*  
404 *of reinforcement and nonreinforcement*, eds. Black A, Prokasy W. (Appleton-Century-Crofts, New  
405 York), pp. 64–99.
- 406 23. Raftery AE (1995) Bayesian model selection in social research. *Sociological methodology* pp.  
407 111–163.
- 408 24. Lau H, Rosenthal D (2011) Empirical support for higher-order theories of conscious awareness.  
409 *Trends in Cognitive Sciences* 15(8):365 – 373.
- 410 25. Fleming SM, Lau HC (2014) How to measure metacognition. *Frontiers in Human Neuroscience*  
411 8:443.
- 412 26. Dienes Z (2015) *How Bayesian statistics are needed to determine whether mental states are*  
413 *unconscious*, ed. Overgaard M. (Oxford University Press, Oxford), pp. 199–221.
- 414 27. Schmidt T (2015) Invisible stimuli, implicit thresholds: Why invisibility judgments cannot be  
415 interpreted in isolation. *Adv Cogn Psychol* 11(2):31–41.
- 416 28. Roy M, Shohamy D, Wager TD (2012) Ventromedial prefrontal-subcortical systems and the  
417 generation of affective meaning. *Trends in Cognitive Science* 16(3):147–56.
- 418 29. Ressler KJ, Mayberg HS (2007) Targeting abnormal neural circuits in mood and anxiety disorders:  
419 from the laboratory to the clinic. *Nature Neuroscience* 10(9):1116–24.



- 420 30. Rauch SL, Shin LM, Phelps EA (2006) Neurocircuitry models of posttraumatic stress disorder and  
421 extinction: human neuroimaging research—past, present, and future. *Biological Psychiatry*  
422 60(4):376–82.
- 423 31. Duits P, et al. (2015) Updated meta-analysis of classical fear conditioning in the anxiety disorders.  
424 *Depression and Anxiety* 32(4):239–253.
- 425 32. Jovanovic T, Norrholm SD (2011) Neural mechanisms of impaired fear inhibition in posttraumatic  
426 stress disorder. *Frontiers in Behavioral Neuroscience* 5:44.
- 427 33. Spielberger CD (1983) State-trait anxiety inventory for adults. *PsycTESTS Dataset*.
- 428 34. Szymanski J, O’Donohue W (1995) Fear of spiders questionnaire. *PsycTESTS Dataset*.
- 429 35. Merikle PM, Smilek D, Eastwood JD (2001) Perception without awareness: perspectives from  
430 cognitive psychology. *Cognition* 79(1-2):115–34.
- 431 36. Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W (2006) Variational free energy  
432 and the laplace approximation. *NeuroImage* 34(1):220–234.
- 433 37. R Core Team (2016) *R: A Language and Environment for Statistical Computing* (R Foundation for  
434 Statistical Computing, Vienna, Austria).
- 435 38. Bates DM (2005) Fitting linear mixed models in R. *R News* 5:27–30.
- 436 39. Lenth RV (2016) Least-squares means: The R package lsmeans. *Journal of Statistical Software*  
437 69(1):1–33.

## 438 **Acknowledgments**

439 We thank Patrik Vuilleumier who created and shared the spider stimuli. This work was supported in part  
440 through the computational resources and staff expertise provided by Scientific Computing at the Icahn  
441 School of Medicine at Mount Sinai. Funding was provided by NIMH grant MH105515 and a  
442 Klingenstein-Simons Fellowship Award in the Neurosciences to D.S.; ERC Advanced Grant XSPECT-  
443 DLV-692739 to D.C. (Co-I); and Swiss National Science Foundation grant SNF 161077 to P.H.

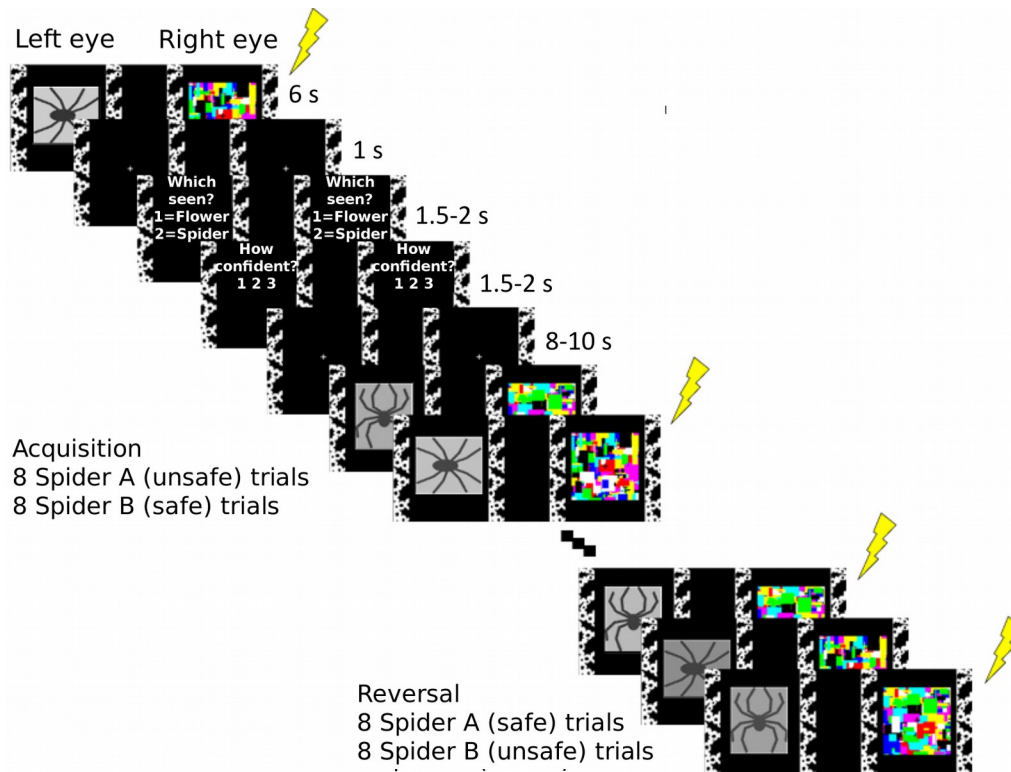
## 444 **Author Contributions**

445 PH carried out the computational modeling and statistical analysis, interpreted the results, and drafted  
446 the manuscript. HL prepared materials, collected the data and critically revised the manuscript. IL  
447 contributed to the conception of the study, the computational modeling, the interpretation of the results,  
448 and critically revised the manuscript. CMR contributed to the interpretation of the results and critically  
449 revised the manuscript. DRB contributed to the computational modeling and critically revised the  
450 manuscript. DS conceived, designed and coordinated the study, contributed to data analysis and  
451 interpretation, and critically revised the manuscript. DC contributed to the conception of the study, data  
452 analysis and interpretation of results, and drafted the manuscript in its final form. All authors gave final  
453 approval for publication.

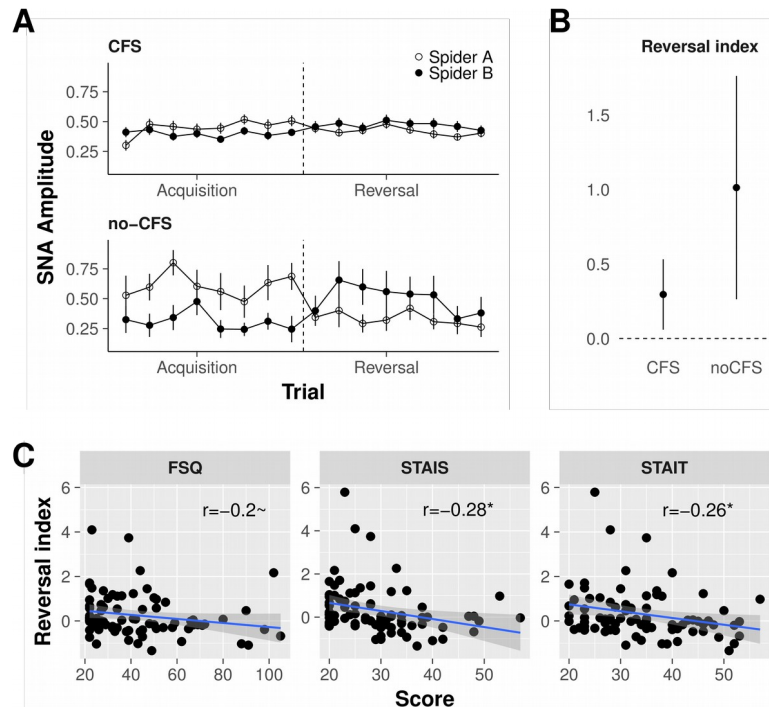
## 454 **Conflict of Interests**

455 All authors declare no conflicts of interest with regard to the current study.

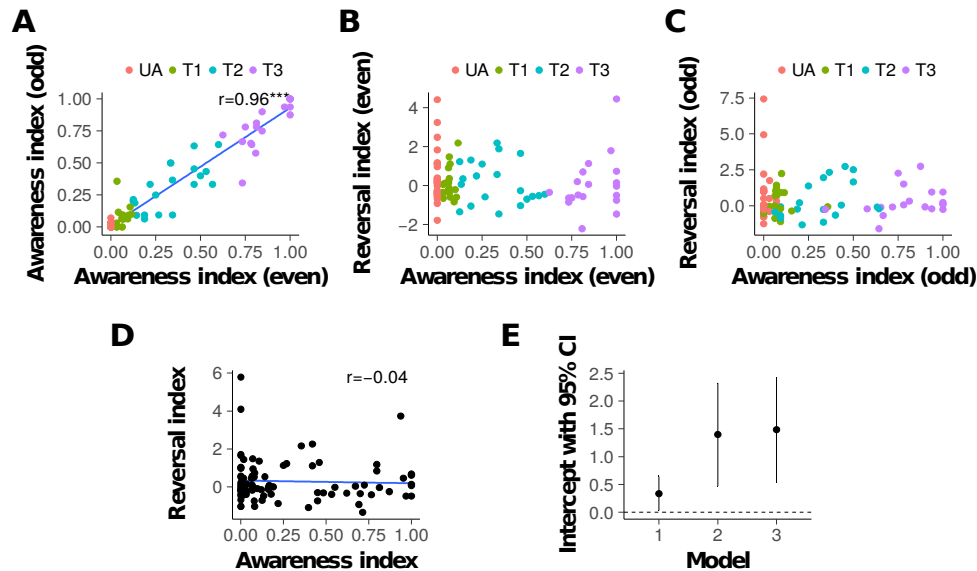
## Figures and Tables



**Figure 1: Schematic description of experimental design and procedure.** In each trial of the acquisition phase, participants were presented with one of two stimuli (schematic pictures of spiders, presented monocularly for 6 sec and suppressed from awareness by a CFS mask shown to the other eye). One image (spider A) always terminated with a mild electric shock to the wrist, whereas the other (spider B) never did. Halfway through the experiment, with no warning, the contingencies flipped and the reversal phase began: the formerly safe stimulus (spider B) now predicted the shock, and the old threat-associated one (spider A) was now safe. Each spider was shown 8 times in each phase. Trial order was pseudorandomized (see Materials and Methods) and spider identity (A and B) was counterbalanced across participants. To assess the success of the awareness manipulation, participants answered the questions "Which seen?" (1=flower, 2=spider) and "How confident?" (1=guess to 3=sure), presented binocularly (1.5 - 2 s each), beginning 1 s after the offset of every CS, and followed by an 8-10 s inter-trial interval (the questions are only shown here for the first depicted trial, but were repeated in all trials). Participants who underwent the same procedure without CFS were shown identical CSs, but the CFS mask was absent.



**Figure 2: Physiological reversal learning.** **A.** Time courses reveal reversal of threat responses with and without continuous flash suppression. Data points represent trial-wise mean responses to spider A (the CS+ during acquisition) and spider B (the CS- during acquisition). Both groups showed reversal learning, as indicated by greater responses to Spider A during the acquisition phase and greater responses to Spider B during the reversal phase. Error bars represent standard errors. **B.** **Mean reversal learning index for each group.** Error bars represent 95% confidence intervals, indicating that the interaction of stage and stimulus and thus the magnitude of reversal learning in both groups was significantly greater than zero. **C.** **Heightened anxiety is associated with impaired reversal learning under CFS.** A negative correlation between baseline anxiety measures and the strength of threat reversal learning is evident for state and trait anxiety. Blue lines show linear fits of each score to the reversal index, and ribbons around lines indicate bootstrapped 95% confidence intervals around the estimate. *Abbreviations:* STAIS/STAIT, state/trait anxiety subscale of the Spielberger State-Trait Anxiety Inventory; FSQ, Fear of Spider Questionnaire, ~,  $P < .1$ ; \*,  $P < .05$ .



**Figure 3: Characterizing the relation between perceptual awareness and reversal learning in the CFS group.** **A. Correlation between the awareness index of even and odd-numbered trials.** Each data point represents an individual participant. The strong positive correlation between these independent measures of awareness demonstrates that individual participants' awareness ratings - even those with extreme values of zero or one - are unlikely to be due to measurement noise. For illustrative purposes, the color scheme marks all participants with an awareness index of 0 in even trials in red (UA, unaware,  $N = 27$ ) and classifies the rest of the CFS group in 3 tertiles (T1-T3). Note that some data points overlap. **B. Reversal learning plotted against perceptual awareness for individual participants, for data obtained from even-numbered trials.** The color scheme is the same as in Panel A. **C. Reversal learning plotted against perceptual awareness for individual participants, for data obtained from odd-numbered trials.** Individual participants are marked with the same color as in the previous panels; the overall distribution of participants is highly similar across panels. This suggests that two independent measures of awareness (even and odd trials) showed very similar results, indicating that the overall awareness index was unlikely to be influenced by extreme values that were due to measurement noise. **D. Reversal learning as a function of perceptual awareness in the CFS group, using data pooled from all trials.** The intercept, indicating the magnitude of reversal learning in the absence of awareness, is positive and significantly different from zero. **E. Reversal Index intercepts and their 95% confidence intervals in a series of regression models.** Model 1 depicts the intercept (the value of the reversal index when the awareness index equals zero) shown in Panel D. Model 2 shows the intercept when the regression model includes STAIT scores in addition to the perceptual awareness index. Model 3 regresses the reversal index onto the perceptual awareness index, STAIT and tracking scores. (Excluding the potential outlier in the top left corner of panel D weakens significance of the intercept in model 1,  $P = 0.07$ ; the intercepts of model 2 and 3 remain significant after removal of this outlier). Blue lines show linear fits, and ribbons around lines indicate bootstrapped 95% confidence intervals around the estimate.

<b>Model</b>	<b>Predictor</b>	<b>Beta</b>	<b>SE</b>	<b><i>t</i></b>	<b><i>P</i></b>
1	Intercept	0.3	0.2	2.1	0.035
1	Awareness index	-0.1	0.4	-0.4	0.692
2	Intercept	1.4	0.5	3	0.004
2	STAIT	0	0	-2.3	0.024
2	Awareness index	-0.2	0.4	-0.5	0.596
3	Intercept	1.5	0.5	3.1	0.003
3	STAIT	0	0	-2.4	0.021
3	Tracking score	-0.3	0.3	-1	0.318
3	Awareness index	-0.2	0.4	-0.5	0.597

**Table 1: Regression coefficients for all awareness index models.** Reversal learning was the dependent variable in all models. Model 1 included an intercept and the perceptual awareness index; model 2 additionally included STAIT scores; model 3 additionally included STAIT and tracking scores.