Biological Psychiatry: CNNI

Archival Report

Mathematical Characterization of Changes in Fear During Exposure Therapy

Ana Portêlo, Youssef Shiban, and Tiago V. Maia

ABSTRACT

BACKGROUND: During exposure therapy, patients report increases in fear that generally decrease within and across exposure sessions. Our main aim was to characterize these changes in fear ratings mathematically; a secondary aim was to test whether the resulting model would help to predict treatment outcome.

METHODS: We applied tools of computational psychiatry to a previously published dataset in which 30 women with spider phobia were randomly assigned to virtual-reality exposures in a single context or in multiple contexts (n = 15 each). Patients provided fear ratings every minute during exposures. We characterized fear decrease within exposures and return of fear between exposures using a set of mathematical models; we selected the best model using Bayesian techniques. In the multiple-contexts group, we tested the predictions of the best model in a separate, test exposure, and we investigated the ability of model parameters to predict treatment outcome.

RESULTS: The best model characterized fear decrease within exposures in both groups as an exponential decay with constant decay rate across exposures. The best model for each group had only two parameters but captured with remarkable accuracy the patterns of fear change, both at the group level and for individual subjects. The best model also made remarkably accurate predictions for the test exposure. One of the model's parameters helped predict treatment outcome.

CONCLUSIONS: Individual patterns of fear change during exposure therapy can be characterized mathematically. This mathematical characterization helps predict treatment outcome.

https://doi.org/10.1016/j.bpsc.2021.01.005

Exposure therapy is effective for specific phobias and other anxiety disorders (1). During exposure therapy, patients experience fear that usually decreases within and across exposures (2). Often, some fear returns between exposures—for example, through renewal, which occurs when the exposure context changes (3,4). The patterns of fear change during exposure therapy have usually been described qualitatively (5–7). Some work has tried to characterize these patterns quantitatively, using hierarchical linear models (8,9), but such patterns are patently nonlinear. An accurate mathematical characterization of these patterns could have clinical utility, especially if it supports treatment-outcome predictions. Such characterization would be especially compelling if it relates to theory-driven mechanistic models. Here, we use computational-psychiatry tools (10–13), applied to a published dataset (4), to 1) develop such mathematical characterization, using fear ratings of patients with spider phobia during exposure therapy; 2) show that the resulting mathematical model supports treatment-outcome predictions; and 3) show that this data-driven model relates to theory-driven models.

We focused on developing a model that jointly describes fear decrease within exposures and return of fear between exposures. We also considered possible changes in fear-decrease rate across exposures because both habituation and extinction become faster with repetition (3,14). Similarly, we considered possible changes in return-of-fear rates with repeated exposures because, for example, extinction in multiple contexts decreases renewal-induced return of fear (4,15–18), repeated habituation slows spontaneous recovery (14,19), and repeated dishabituation itself habituates (14). We therefore investigated four component processes: fear decrease within exposures, possible changes in fear-decrease rate across exposures, return of fear between exposures, and possible changes in return-of-fear rate across exposures.

We first investigated these processes using classical statistical analyses. We used the results of those analyses to ground model development. We defined a set of plausible mathematical formulations for each component process, and we created integrated models by considering all possible combinations of those formulations. We then selected the integrated model that best described the full pattern of fear within and across exposures. We tested the predictions of this model in a new, test exposure. We also investigated the usefulness of model parameters, obtained after only two exposures, to predict individual treatment outcome. Finally, we considered the relation of the selected model to theory-driven models.

DATASET

An ideal model would work in the complex world of real-life clinical practice. Attempting to fully tackle such complexity, however, could hamper progress; we therefore sought more

SEE COMMENTARY ON PAGE 1040

controlled conditions for this work. We found an excellent testbed in a published virtual-reality study of exposure therapy for spider phobia (4), with three advantages for our purposes: the fear-eliciting stimulus—a wiggling but otherwise stationary virtual-reality spider—was equal for all patients and remained constant within each exposure; patients reported fear (on a scale of 0–100) every minute, thereby ensuring sufficiently dense sampling for model identification; and in one condition, context changed on each exposure, thereby eliciting return of fear through renewal (18,20).

Detailed information about patient characteristics is provided in the previously published article that analyzed this dataset using simple statistical analyses (4). Briefly, patients were 30 treatment-naïve women with spider phobia and no comorbidities. The study was approved by the Ethics Committee of the University of Würzburg. Patients provided written informed consent.

Exposure therapy consisted of four training exposures followed by a test exposure (Figure 1). Patients were randomly assigned to one of two groups: single context (n = 15), in which the spider was presented in the same room color in all training exposures, or multiple contexts (n = 15), in which the room color was different on each training exposure. In both groups, the spider in the test exposure was presented in a room with a color not used during training. Patients completed the German version (21) of the Fear of Spiders Questionnaire (FSQ) (22) before and after treatment.

ANALYSES USING CLASSICAL STATISTICS

We first investigated the four component processes using classical statistics. (We use "classical statistics" to refer to standard statistical approaches, as opposed to custom-made models.) Specifically, we used mixed analyses of variance

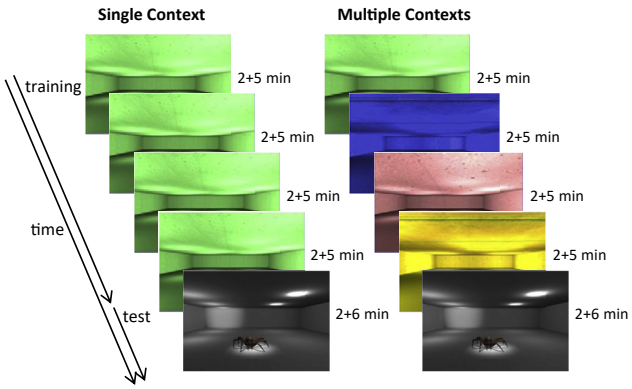

Figure 1. Virtual-reality exposure therapy protocol. Patients with spider phobia were randomly assigned to one of two groups: single context (left; n = 15) or multiple contexts (right; n = 15). Exposures consisted of seeing a wiggling spider in a colored room on a virtual-reality headset. In both groups, patients underwent four training exposures (5 min each) followed by a test exposure (6 min). The only difference between the groups was that the training exposures in the single-context group all used the same room color, whereas in the multiple-contexts group they each used a different color. The test exposure had the same room color for both groups, which had not been seen during training in either group. Each exposure was preceded by a 2-minute presentation of the corresponding room without the spider. Patients were instructed not to avoid looking at the spider.

(Supplement). The main findings from those analyses were that 1) fear decreased within exposures in both groups; 2) fear decrease diminished across exposures, with no evidence that this reduction differed between the groups; 3) the groups differed significantly in return of fear, with evidence for return of fear in the multiple-contexts group but not in the single-context group; and 4) there was no statistically significant evidence that return of fear changed across exposure intervals, even in the multiple-contexts group (Figure 2; Supplement).

The analyses using classical statistics help test hypotheses about component processes but have multiple limitations. We therefore conducted model-based analyses that address these limitations (Table S1).

METHODS AND MATERIALS

Candidate Models

The analyses using classical statistics confirmed that 1) fear decreased within exposures, 2) the amount of fear decrease diminished across exposures, and 3) the multiple-contexts, but not the single-context, group exhibited return of fear. We therefore incorporated these three processes in our models. Specifically, each model consisted of functions D, S, and R representing, respectively, fear decrease within exposures, possible changes in the steepness of D across exposures, and return of fear between exposures. The classical statistical analyses showed no evidence for a change in return of fear across exposure intervals during the training period, which we used to fit our models, so we did not model such change.

We defined several plausible mathematical formulations for D, S, and R; we then created the set of candidate models by considering all possible combinations of those formulations (Figure 3; Supplement). We performed model selection using these integrated models, to consider all fear ratings simultaneously. Given the treatment protocols—eliciting renewal and therefore return of fear in the multiple-contexts but not in the single-context group—and the results of the classical statistical analyses, we expected R to differ between the groups. We describe next the formulations we tested for D, S, and R.

Fear Decrease Within Exposures. We represented fear decrease within exposures by a function $D(t)$ (Figure 3) that gave the estimated fear for each time t within the exposure (other than the initial rating for the exposure, which we address below). We tested several functions: constant (D^{Const}), linear (D^{Lin}), exponential (D^{Exp}), power (D^{Pow}), linear on the logarithm (D^{Ln}), linear on the square root ($D^{LinSqrt}$), and exponential on the square root ($D^{ExpSqrt}$) (Supplement). We expected that the exponential function might be best because, empirically, habituation (3,14) and possibly extinction (3) often produce exponential decays. Moreover, both theoretical models of habituation (23,24) and simple reinforcement-learning or conditioning models, such as the Rescorla-Wagner model, applied to extinction (25,26) produce exponential decays. (Reinforcement-learning and conditioning models are often applied on a trial-by-trial—in this case, session-by-session—basis, but they could model gradual extinction within each session if applied repeatedly

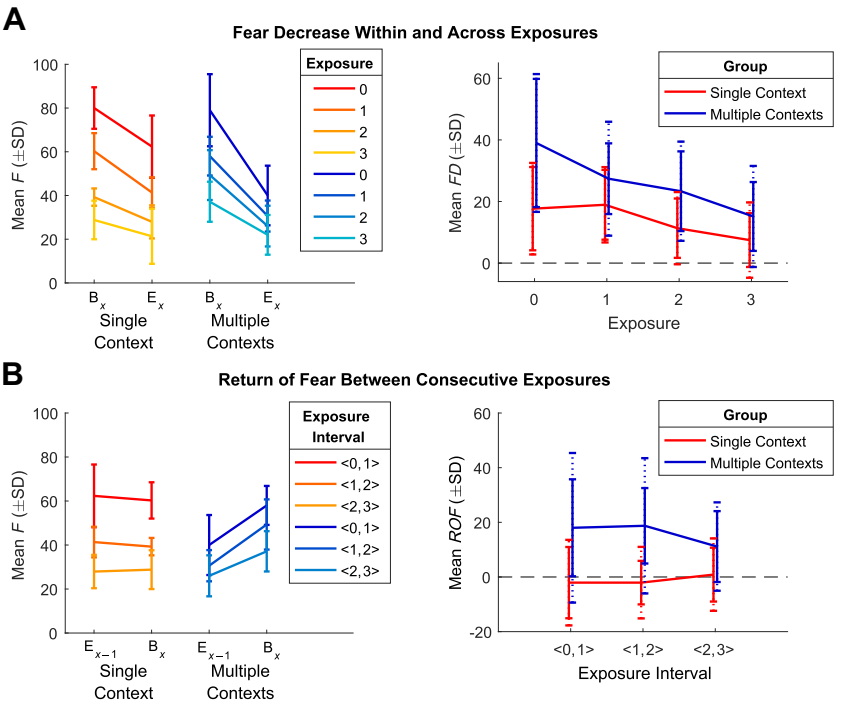

Figure 2. Fear ratings of patients in the single-context (n = 15) and multiple-contexts (n = 15) groups. Dashed error bars represent SDs; solid error bars represent SDs with the variability related to between-subject differences removed [by mean-normalizing each subject's scores and correcting for the bias introduced by that normalization (46)]. Plots on the left show only the latter to avoid excessive clutter. **(A)** Fear decrease within and across exposures. (Left panel) Mean (± SD) fear ratings (F) at the beginning (B_x) and end (E_x) of each exposure (x). Fear decreases both within exposures (compare values at the beginning vs. end of each exposure) and across exposures (compare differently colored lines). (Right panel) Mean (± SD) fear decrease (FD) for each exposure, where FD for exposure x is defined as $F(B_x) - F(E_x)$. Both groups show FD within exposures ($FD > 0$), with the amount of FD diminishing across exposures. **(B)** Return of fear (ROF) between consecutive exposures. (Left panel) Mean F (± SD) at the end of each exposure (E_{x-1}) and at the beginning of the subsequent exposure (B_x) for each of the 3 intervals between consecutive exposures ($<x-1, x>$). (Right panel) Mean (± SD) ROF between consecutive exposures, where ROF for exposure interval $<x-1, x>$ is defined as $F(B_x) - F(E_{x-1})$. The multiple-contexts group exhibits return of fear ($ROF > 0$) but the single-context group does not ($ROF \approx 0$).

within the session.) The Supplement describes the theoretical rationale for the other functions.

Change in Fear-Decrease Steepness Across Exposures. The classical statistical analyses found that the amount of fear decrease diminished across exposures. Those analyses, however, could not adjudicate whether this reduction reflected a parametric change in fear-decrease steepness or instead resulted from nonlinear fear decrease with a constant steepness parameter—e.g., exponentially decaying fear with a constant decay rate (Supplement). We therefore considered two possible formulations for function $S(x)$, which gives the steepness parameter for exposure x: the steepness parameter could be constant (S^{Const}) or vary linearly with exposure number (S^{Lin}) (Figure 3; Supplement). We did not test nonlinear formulations for $S(x)$ to avoid a combinatorial explosion in the number of models.

Return of Fear. We represented return of fear between exposures by a function $R(<x-1, x>)$ of the interval between exposures $x - 1$ and x, which gave the estimated return of fear for that interval (Figure 3; Supplement). The classical statistical analyses found evidence for return of fear in the multiple-contexts group but not in the single-context group, with no significant evidence that return of fear changed across exposure intervals. We therefore considered formulations of R with no return of fear (R^{No}) and with constant return of fear across exposure intervals (R^{Const}). We also considered a formulation in which return of fear was proportional to the estimated fear decrease in the preceding exposure (R^{Prop}) (Supplement). This formulation, inspired by habituation models in which spontaneous recovery is proportional to the difference between the initial and habituated responses (23,24), captured the intuition that return of fear might partially, and proportionally, undo the preceding fear decrease.

Initial Fear. The initial fear for exposure x (with $x > 0$) was obtained by adding $R(<x-1, x>)$ to the estimated fear at the end of exposure $x - 1$ (Figure 3; Supplement). Preliminary analyses (not shown) determined that setting the initial fear for the first exposure to its observed value produced better models than considering it a free parameter.

Model Selection

We used the VBA (Variational Bayesian Analysis) toolbox (27) for Bayesian model selection. We grouped models into families of related models (28). We used between-groups random-effects Bayesian model selection to calculate the posterior probability (PP) that the model (or family) frequencies were equal for the two groups (27–29). Depending on whether the PP was high or low, we performed random-effects Bayesian model selection on pooled data from the two groups or separately for each group, respectively. In either case, we identified the best model (or family) using the exceedance probability (EP): the probability that a model (or family) is better than any other (28). The Supplement provides additional details about the model fitting and selection.

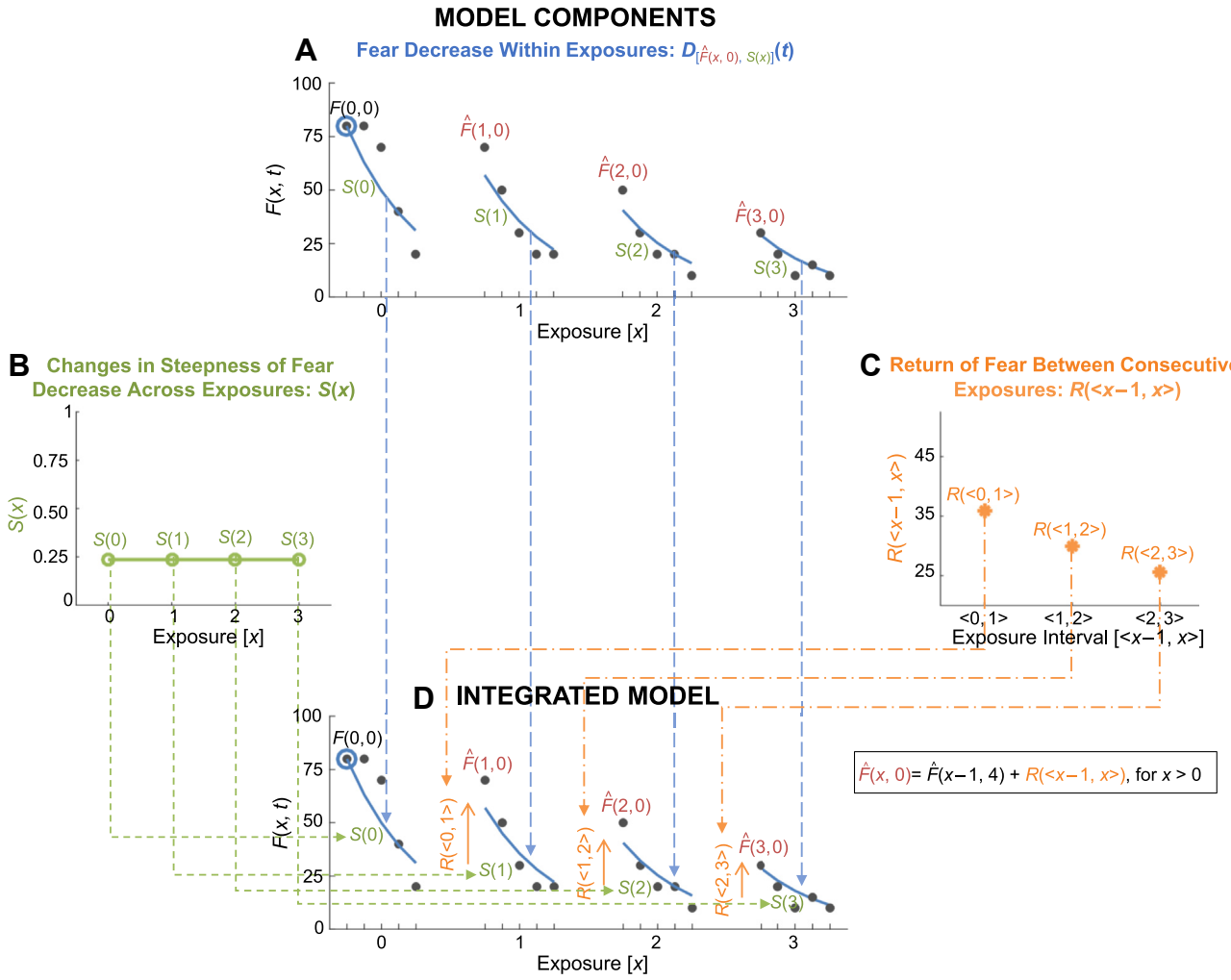

Figure 3. Model structure. We developed models capable of accounting simultaneously for each subject's full pattern of fear ratings, $F(x, t)$, for each time, t, within each exposure, x. These models captured **(A)** fear decrease within exposures, **(B)** possible changes in the fear-decrease steepness across exposures, and **(C)** return of fear between exposures, **(D)** bringing these component processes together in an integrated model. For illustrative purposes, the figure shows the fear ratings from an individual subject from the multiple-contexts group (subject #19), with the corresponding model fit. **(A)** Fear decrease within exposures. We modeled fear decrease within exposures by a function $D(t)$ of time t within the exposure. This function was characterized by two parameters that could vary with exposure (x): the estimated fear at the beginning of the exposure [$\hat{F}(x, 0)$] and a parameter that determined the fear-decrease steepness for the exposure [$S(x)$]. We represent the dependence of $D(t)$ on these parameters by writing them in subscripted square brackets: $D_{[\hat{F}(x,\,0),\,S(x)]}(t)$. Although $\hat{F}(x, 0)$ and $S(x)$ are "parameters" in the sense that they characterize D, they are not free parameters; instead, they are determined by an interaction between the modeled processes. For this reason, we do not represent them with Greek letters, which we reserve for free parameters. Panel **(A)** shows the fear ratings from subject #19 (dots) and the corresponding fit using an exponentially decaying function (solid lines), which we found was the best formulation for D (as shown in the Results). For each exposure, x, this exponentially decaying function D is parameterized by $\hat{F}(x, 0)$ and $S(x)$. **(B)** Changes in fear-decrease steepness across exposures. We allowed the steepness of D to change across exposures according to a function $S(x)$. Panel **(B)** shows the estimated values of $S(x)$ for each exposure x for subject #19. Those values were constant, which we found was the best formulation for S (as shown in the Results). **(C)** Return of fear. We modeled return of fear between exposures $x - 1$ and x by a function $R(<x-1, x>)$. Panel **(C)** shows the estimated values of $R(<x-1, x>)$ for each exposure interval, $<x-1, x>$, for subject #19. Those values were proportional to the estimated fear decrease in exposure $x - 1$, which we found was the best formulation for R (as shown in the Results). **(D)** Integrated model. The integrated model was defined by combining functions D, S, and R. Fear within each exposure x was estimated by application of $D_{[\hat{F}(x,\,0),\,S(x)]}(t)$ for all time points, t, other than the first (i.e., for all $t > 0$). For all exposures other than the first (i.e., for all $x > 0$), the estimated initial fear rating, $\hat{F}(x, 0)$, was obtained by summing the estimated fear at the end of the previous exposure, $\hat{F}(x - 1, 4)$, and the estimated return of fear between $x - 1$ and x, $R(<x, x-1>)$. In the special case of the first exposure, the estimated initial fear rating was set to the observed initial fear rating: $\hat{F}(0, 0) = F(0, 0)$. Panel **(D)** shows the same data and fits as panel **(A)** does, but it illustrates the interaction between the three functions: 1) the fits use the exponentially decaying function D [incoming dashed arrows from panel **(A)**]; 2) the values of $S(x)$ that parameterize D on each exposure are given by their own function $S(x)$ [incoming dashed arrows from panel **(B)**]; 3) and the values of the estimated initial fear ratings, $\hat{F}(x, 0)$, for each exposure other than the first, are given by summing the estimated return of fear, given by function R [incoming dashed arrows from panel **(C)**], to the estimated fear at the end of the previous exposure, $\hat{F}(x-1, 4)$ (see equation in rectangle).

Model Validation: Test Exposure

We tested our model using the test exposure. We therefore used a holdout-set approach, fitting and selecting the model using a data subset (the training exposures) and testing it using a different subset (the test exposure). We used only the multiple-contexts group for these analyses because it underwent the same protocol during the training and test exposures: changing the room color in every exposure. In the single-context group, the protocol changed between the training exposures, which used a constant color, and the test exposure, which used a different color (Figure 1); thus, the model derived using the training exposures could not be expected to accurately predict the test exposure.

Predicting Treatment Outcome

We used linear regression to test whether model parameters obtained for individual patients could predict treatment outcome, defined as the change in FSQ score from pre- to posttreatment ($\Delta FSQ = FSQ_{Pre} - FSQ_{Post}$). We estimated the model parameters using only the first two exposures because one wants to predict treatment outcome early in treatment. We focused on predicting changes in FSQ (ΔFSQ), not posttreatment FSQ scores, because pre- and posttreatment FSQ scores correlated strongly ($r = .87$; 95% confidence interval [CI], 0.61 to 0.96; $t_{11} = 5.83$, $p < .001$) (Figure S1); moreover, the model describes fear changes, so conceptually it aligns more closely with score changes. We used only the multiple-contexts group for treatment-outcome prediction. In the single-context group, the model was developed under constant exposure conditions, so it could not be expected to accurately predict generalization to different contexts, as is likely required when filling out the FSQ.

RESULTS

Fear Decrease Within Exposures

Function D fundamentally determines model behavior, so we initially grouped models into families based on their D function (Equations SE3–SE9 in the Supplement). The two groups had the same frequencies of D functions (PP = 0.999). The family in which D was exponential (D^{Exp}) (Equation SE5 in the Supplement) was the best in the combined sample with both groups (EP = 0.991) (Figure 4A). All subsequent results therefore refer to models from this family.

Change in Fear-Decrease Steepness Across Exposures

We next grouped models into families based on their S functions (Equations SE10–SE11 in the Supplement), considering only models in which D was exponential (D^{Exp}). There was evidence that the two groups had the same frequencies of S functions (PP = 0.856). The family in which the steepness of D^{Exp} was constant across exposures (S^{Const}) (Equation SE10 in the Supplement) was the best in the combined sample with both groups (EP = 0.982) (Figure 4B). All subsequent results therefore refer to models from this family. Univariate Wald tests confirmed that the parameter determining the fear-decrease steepness, λ, was greater than 0 in each group (single-context group: $W_1 = 16.84$, $p < .001$; mean = 0.075; 95% CI, 0.038 to 0.112; multiple-contexts group: $W_1 = 13.63$, $p < .001$; mean = 0.197; 95% CI, 0.130 to 0.263), confirming that D^{Exp} represented exponentially decaying fear (Equation SE5 in the Supplement).

Return of Fear

As we hypothesized, the frequencies of the R functions differed between the groups (1 − PP = 0.927), given the selection of D^{Exp} with S^{Const}. We therefore conducted random-effects Bayesian model selection for R in each group separately. The single-context and multiple-contexts groups were best characterized by R^{Const} (EP = 0.996) (Figure 4C) and R^{Prop} (EP = 0.970) (Figure 4D), respectively.

We tested whether, in each group, the parameter estimate for the selected R function differed significantly from zero, using a univariate Wald test. The constant return of fear (ρ) (Equation SE13 in the Supplement) for the single-context group did not differ significantly from zero ($W_1 = -0.008$, $p = .994$; mean = −0.99; 95% CI, −7.34 to 5.36). The proportionality constant (α) (Equation SE14 in the Supplement) for the multiple-contexts group differed significantly from 0 ($W_1 = 5.73$, $p < .001$); moreover, as expected (Supplement), it fell between 0 and 1 (mean = 0.556; 95% CI, 0.384 to 0.729).

Relation Between Fear Decrease and Return of Fear

The parameters for the single-context group (λ and ρ) correlated strongly and significantly ($r = .714$, $t_{13} = 3.68$, $p = .003$), but the parameters for the multiple-contexts group (λ and α) did not ($r = -.002$, $t_{13} = -0.005$, $p = .996$) (Figure S2).

Goodness of Fit and Accuracy of Predictions

Model fits and predictions for the training and test exposures, respectively, were remarkably accurate, at both the group and individual-subject levels (Figure 5). This accuracy was confirmed by quantitative analyses of the fits and predictions and by graphical analyses of the residuals (Supplement, Table S2, and Figure S3).

Predicting Treatment Outcome

Data visualization showed that ΔFSQ related positively to λ but did not seem to relate to α (Figure S4). Thus, and because we had few datapoints, we focused on predicting ΔFSQ using only λ, with simple linear regression. Exploratory analyses (not shown) showed that the fit and predictive accuracy were worse if using both λ and α.

The λ coefficient in the simple linear regression of ΔFSQ on λ was significant ($b_\lambda = 44.04$, $t_{11} = 2.39$, $p = .036$). Moreover, the regression's R^2 was reasonably high for behavioral studies ($R^2 = .341$, adjusted $R^2 = .282$). The predicted R^2, obtained using leave-one-out cross-validation, confirmed that λ was a useful, even if not strong, predictor of ΔFSQ (predicted $R^2 = .128$) (see also Figure S5). The Supplement shows that ΔFSQ was better predicted by λ than by variables not derived from the model. In fact, predicting ΔFSQ was not possible without the model.

DISCUSSION

Summary and Implications

Fear Decrease Within Exposures. Fear decayed exponentially within exposures in both groups. This finding is consistent with the exponential decay found in habituation (14),

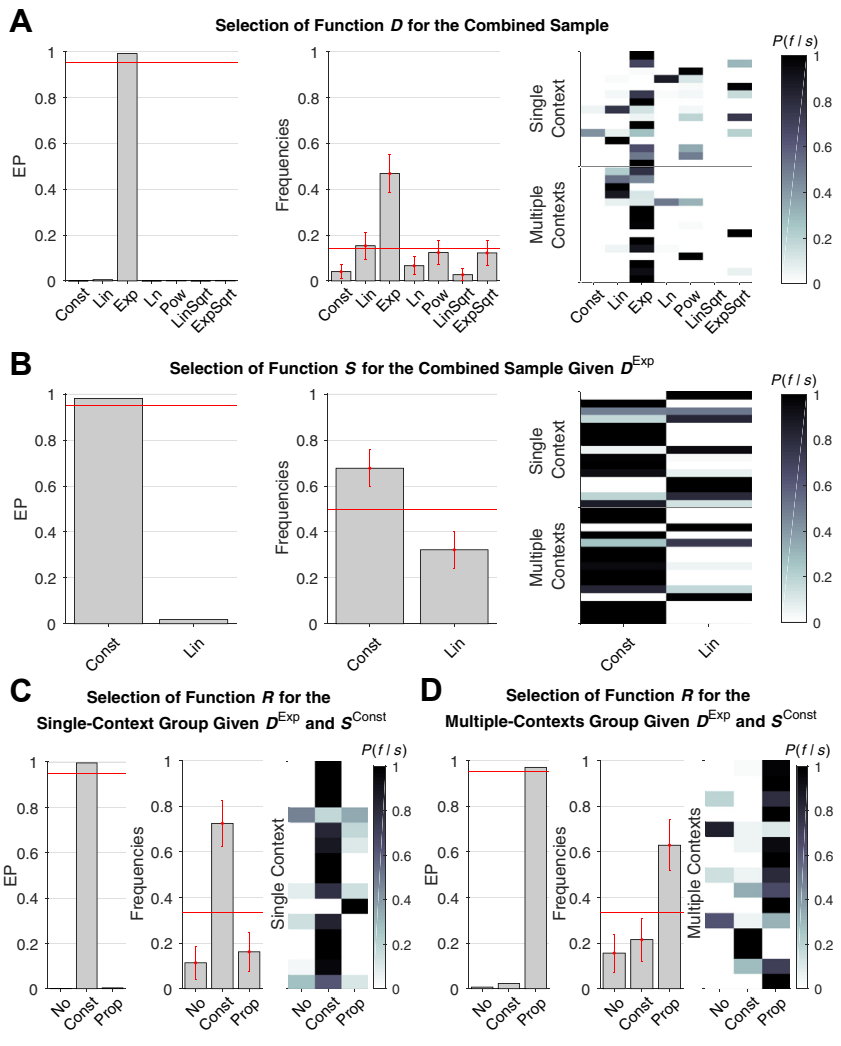

Figure 4. Results of the Bayesian model-selection analyses. Each panel contains three graphs. The graph on the left represents the exceedance probability (EP) for each family or model; a family or model whose EP is above the red line is significantly more likely than any of the other families or models. The middle graph represents the frequency distribution of each family or model. The right graph represents the probability distribution over the various families or models for each individual subject; in panels **(A)** and **(B)**, the first 15 lines represent the subjects from the single-context group and the last 15 lines represent the subjects from the multiple-contexts group; in panels **(C)** and **(D)**, only subjects from the group represented (single context and multiple contexts, respectively) are shown. **(A)** Selection of the function that best characterizes fear decrease within exposures (function *D*; see Supplement) in the pooled sample with both groups (with each function defining a family). The 7 candidate functions tested were constant (Const), linear (Lin), exponential (Exp), power (Pow), linear on the logarithm (Ln), linear on the square root (LinSqrt), and exponential on the square root (ExpSqrt) (Equations SE3–SE9 in the Supplement, respectively). The Exp function was confidently selected as the best function (left panel); most subjects in each group were best characterized by this function (right panel). **(B)** Selection of the function that best characterizes possible changes in the fear-decrease steepness across exposures (function *S*; see Supplement) in the pooled sample with both groups (with each function defining a family), given the selection of the Exp function for *D*. The two candidate functions tested were constant (Const) and linear (Lin) (Equations SE10 and SE11 in the Supplement, respectively). The Const function was confidently selected as the best function (left panel); most subjects in each group were best characterized by this function (right panel). **(C)** Selection of the function that best characterizes the return of fear between consecutive exposures (function *R*; see Supplement) for the single-context group, given the selection of the Exp and Const functions for *D* and *S*, respectively. The three candidate functions tested were no return of fear (No), constant return of fear in all exposure intervals (Const), and return of fear proportional to the estimated fear decrease in the preceding exposure (Prop) (Equations SE12–SE14 in the Supplement, respectively). The Const function was confidently selected as the best function (left panel), characterizing most subjects from the single-context group (middle and right panels). **(D)** Selection of the function that best characterizes return of fear between consecutive exposures (function *R*; see Supplement) for the multiple-contexts group, given the selection of the Exp and Const functions for *D* and *S*, respectively. The candidate functions tested were the same as in panel **(C)**. The Prop function was confidently selected as the best function (left panel), characterizing most subjects from the multiple-contexts group (middle and right panels).

extinction (3), and their theoretical models (23–26). The exponential decay means that the fear-decrease rate—the derivative of fear with respect to time—is proportional to the fear level (Equation SE17 in the Supplement; see also text in the Supplement). Clinically, the counterintuitive implication is that stronger fear decreases faster than weaker fear. A related implication is that fear decreases faster early in exposure therapy—and, within each exposure, early in the exposure—when it is strongest.

The other mathematical formulations of fear decrease received little support. The lack of support for the linear function highlights the inadequacy of linear models to capture fear change within exposures—a conclusion that is also obvious visually (Figure 5). The lack of support for the power function is reminiscent of an argument concerning skill learning. Originally, skill learning, assessed through reaction-time decreases, was suggested to follow a power function—the power law of practice (30). Subsequent work, however, suggested that this finding could be an artifact of averaging across subjects, with individual subjects following instead an exponential function (31,32). This argument relates to our findings because, by using mixed models, we fit our models at the individual-subject level (with subjects nested in a population). It seems interesting that behavior changes exponentially in domains as distinct as fear decrease and skill learning.

Change in Fear-Decrease Steepness Across Exposures. The exponential decay rate was constant across exposures in both groups. Thus, the finding that fear decrease diminishes across exposures, obtained in the classical

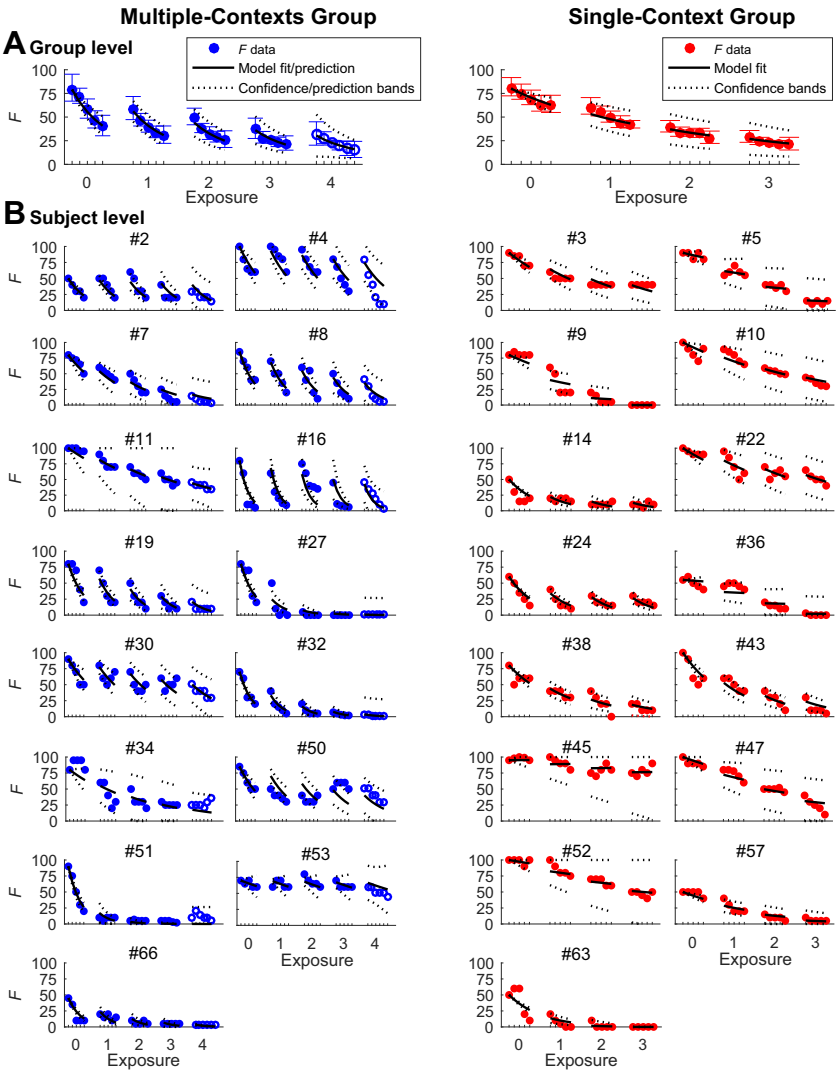

Figure 5. Fear ratings (*F*) (circles) and model fits/predictions (solid lines) for the multiple-contexts (blue; left) and single-context (red; right) groups. We show model fits for the training exposures (exposures 0–3) for each group. We show model predictions for the test exposure (exposure 4) for the multiple-contexts group only. We do not show model predictions for the test exposure for the single-context group because for that group the protocol during training was different from that in the test exposure (the room color was constant during training but changed in the test exposure), so we did not expect the model developed for the training exposures in the single-context group to provide good predictions for the test exposure. **(A)** Group results for each group (multiple-contexts group: blue, left; single-context group: red, right): Mean of *F* (± SD) for each group (circles and error bars), mean of the model fits for the training exposures for each group (solid lines for exposures 0–3), and mean of the model predictions for the test exposure for the multiple-contexts group (solid line for exposure 4). The error bars show the SDs with the variability related to between-subject differences removed [by mean-normalizing each subject's scores and correcting for the bias introduced by that normalization (46)]. The dashed lines show confidence bands obtained by simulations that resampled the multivariate parameter distribution using the parameter variance–covariance matrix (so-called population prediction intervals) (47). **(B)** Results for each individual subject from each group (multiple-contexts group: blue, left; single-context group: red, right): *F* (circles), model fits for the training exposures for each group (solid lines for exposures 0–3), with corresponding 95% confidence bands (dashed lines for exposures 0–3), and model predictions for the test exposure (solid line for exposure 4) for the multiple-contexts group only, with corresponding 95% simultaneous prediction bands (dashed lines for exposure 4). In cases in which the upper band went above 100 (subjects #11, #45, #47, and #52), we capped it at 100 (the maximum value of the rating scale). Within each group, we ordered subjects by their subject number (indicated following the #). The fits are extremely accurate, both **(A)** at the group level and **(B)** for individual subjects. The remarkable accuracy of the fits is especially notable because the best model for each group only has two free parameters (see Supplement).

statistical analyses, does not imply a parametric change in the fear-decrease process. An exponential decay with constant decay rate produces the reduction in fear decrease across exposures because later exposures have smaller initial fear ratings (Figure S6): exponentially decaying fear implies that fear decrease is proportional to the fear level, so if (initial) fear is smaller, fear decrease becomes smaller.

Potentiation of habituation and what might similarly be called potentiation of extinction (3,14) suggest that the decay rate might even increase across exposures. We did not find evidence for such increase. This null result might be due to the small number of exposures: the function that allows the decay rate to change (S^{Lin}) has two parameters, whereas the function with a constant decay rate (S^{Const}) has one; overcoming the penalty for the extra parameter might require more exposures. Of course, by a similar argument, we also cannot exclude a decrease in decay rate across exposures; however, there is little theoretical justification to expect it.

Return of Fear. Given the treatment protocols, we expected that the groups would differ in *R* because the multiple-contexts group, but not the single-context group, underwent renewal (because of the context changes). We found that the groups were indeed characterized by different *R* functions: return of fear was constant for the single-context group but proportional to the (estimated) fear decrease in the preceding exposure for the multiple-contexts group.

For the single-context group, the classical statistical analyses showed no evidence for return of fear. The model-based analyses, however, found that constant return of fear (R^{Const}) was better than no return of fear (R^{No}). This apparent discrepancy is resolved by noting that ρ, the parameter for

R^{Const} (Equation SE13 in the Supplement), has both fixed- and random-effects components. Consistent with the classical statistical analyses, at the group level, ρ did not differ significantly from 0. Thus, R^{Const} was selected not because, at the group level, ρ differed significantly from 0, but because the random-effects component of ρ captured meaningful variability in return of fear across subjects (Figure S7).

For the multiple-contexts group, the classical statistical analyses showed evidence for return of fear but not for changes in return of fear across exposure intervals. The lack of change in return of fear across exposure intervals seems inconsistent with the finding that renewal decreases with repeated exposures in different contexts (15,17). However, if we also included in the classical statistical analyses the interval between the last training exposure and the test exposure, return of fear did seem to decrease across exposure intervals (Figure S8). In the model-based analyses, we could not investigate this issue directly because of the small number of exposure intervals, as *R* formulations explicitly modeling changes in return of fear across exposure intervals would require at least an additional parameter. The finding that the multiple-contexts group was best characterized by R^{Prop}, however, hints at a decrease in return of fear across exposure intervals: as shown in the classical statistical analyses, fear decrease diminishes across exposures (Figure 2A; Supplement), so, if return of fear is proportional to fear decrease in the preceding exposure, it will decrease across exposure intervals.

Relation Between Fear Decrease and Return of Fear. In the single-context group, the positive correlation between the fear-decrease steepness (λ) and return of fear (ρ) indicates that patients whose fear decreased more had more return of fear. In the multiple-contexts group, the selection of R^{Prop} as the return-of-fear function already captures this positive relation between fear decrease and return of fear (Equation SE14 in the Supplement); thus, the fear-decrease steepness (λ) and the proportion of fear that returned (α) did not correlate. This lack of correlation shows that the model for the multiple-contexts group adequately orthogonalized the fear-decrease and return-of-fear processes.

Treatment-Outcome Prediction. A model parameter estimated after two exposures helped to predict treatment outcome, measured by the change in FSQ score from pre- to posttreatment. Such predictions, moreover, were not attainable without the model. Although the predictive accuracy was insufficient to support real-life decisions, our aim with this part of the work was to demonstrate that model-derived parameters have potential clinical usefulness, producing better predictions than are attainable without a model. Moreover, the model parameter might be useful for real-life treatment-outcome prediction if used together with other features.

This part of the work also bears on whether within-exposure fear decrease is relevant for treatment outcome. We found that within-exposure fear decrease measured directly from the data did not predict treatment outcome (Supplement); however, λ, which characterizes such fear decrease more precisely, did. Reports that within-exposure fear decrease is irrelevant for treatment outcome (33) may therefore have resulted from the use of simple measures of fear decrease and might have been different had a model been used. We make this point tentatively, however, for two reasons. First, treatment in our dataset occurred in a single day; as discussed below, short- and long-term treatment effects may differ. Second, participants whose fear decreased more during the exposures might be more convinced that their fear of spiders had decreased, and therefore have greater reductions in FSQ scores, irrespective of whether that effect generalized to real life. Weighing for this possibility, we did not find a relation between the model parameters and changes from pre- to posttreatment in proximity to a real spider in a behavior avoidance test (analyses not shown).

Other Applications. During exposure therapy, therapists have patients record fear ratings (or subjective units of distress); therapists often analyze these records to guide treatment decisions. We envision an app that summarizes a patient's ratings by estimating our model's two parameters. Given the precision with which our model characterizes individual fear patterns, we hypothesize that the two parameters may capture virtually all relevant information in the full ratings time series. Thus, therapists may be able to consult only those two numbers instead of having to analyze the full time series. Of course, this idea must be tested.

Relation to Theoretical Models

Computational psychiatry encompasses data- and theory-driven approaches (13), which can be combined (12). We used a hybrid data- and theory-driven approach by using data to select the best among a set of theoretical models (13). The winning model couples exponentially decaying fear within exposures, which corresponds to the solution of a first-order homogeneous linear differential equation, with discrete jumps between exposures, making it a hybrid dynamical model (Supplement). In the model, fear within exposures decreases following a differential equation in which the instantaneous fear-decrease rate is proportional to the fear level (Equation SE17 in the Supplement), which is consistent with theoretical habituation models (23,24,34). Simple reinforcement-learning or conditioning models typically work with discrete, not continuous, time, so they correspond to difference, not differential, equations; still, such models applied to extinction produce exponentially decaying fear (25,26). The winning model's formulation of fear decrease within exposures is therefore consistent with theoretical ideas concerning habituation and extinction.

Limitations and Extensions

Sample Considerations and Limitations. The dataset we used had small samples (n = 15) in each group. In studies with findings only at the group level, small samples can induce spurious results (35). Our model, however, provided excellent fits and predictions for individual subjects (Figure 5). The probability of a finding that applies to most subjects in a sample of size 15 not generalizing to larger samples is minuscule. Moreover, the applicability of our findings to individual patients is more pertinent to the development of personalized psychiatry than the typical group findings, which often fail to apply to individual patients. In addition, we did not focus primarily on between-group comparisons, so we did not use only 15 datapoints per group. We modeled fear changes throughout exposure therapy: for the training exposures alone, we had 20 fear ratings per subject (4 exposures × 5 ratings per

exposure) and therefore 300 per group (20 × 15). Having 20 ratings per subject and a model with two parameters per subject obeys the rule of thumb of having 10 datapoints per parameter (36). The main limitation of our dataset's sample was that it included only adult women.

Short- Versus Long-term Effects. An important limitation of the dataset we used is that all exposures occurred within a day. Short- versus long-term habituation and extinction differ (37–39); moreover, some anxiety disorders may involve disturbances in extinction recall (40). Future work should test whether the model we identified applies to longer treatment protocols. In forgetting, for example, short-term forgetting may follow an exponential function, as we found for fear decrease, but longer-term forgetting may follow a power function (41). In any case, we see the core contribution of our work to be less about our specific results than about our approach and workflow, which can be applied across treatment protocols and disorders.

Extension to Naturalistic Treatment Protocols. We used a dataset from a highly controlled virtual-reality study to facilitate model development. The disadvantage of this approach is that it limits ecological validity. Real-life exposure therapy is more complex than the protocol in our dataset: fear-eliciting stimuli vary within sessions (e.g., gradually approaching a spider) and across sessions (e.g., approaching different spiders); exposure contexts can and should vary widely to support generalization; and intervals between exposures may vary. Addressing these complexities will require representations that support generalization across stimuli and contexts (42–45) and a return-of-fear formulation that captures spontaneous recovery across varying intervals (23).

Unifying the Findings of the Two Groups. The two groups were characterized by the same fear-decrease function but different return-of-fear functions. A simple extension of our work would characterize both groups with a single model. Specifically, return of fear might occur only when context changes—thereby occurring for the multiple-contexts group but not for the single-context group during the training exposures. In addition, return of fear might decrease with consecutive context changes, as suggested by prior findings (15–17) and by our analyses showing that return of fear might decrease in the multiple-contexts group (Figure S8). The decrease in return of fear with consecutive context changes would explain why return of fear between the last training exposure and the test exposure is smaller in the multiple-contexts group than in the single-context group (Figure S9).

We did not implement this extension because we did not find significant evidence that return of fear decreased across exposure intervals, even in the multiple-contexts group, when analyzing only the training period (which we used for model fitting and selection). A dataset with more exposures and greater variation in the number of context changes would be better to test such a model.

Conclusions

We found that a simple model characterizes the fear changes of individual patients during exposure therapy with remarkable precision. The model combines exponentially decaying fear within exposures with fear increases between exposures. The model has only two parameters, thereby solving the curse of dimensionality, and hence providing a promising basis for clinically relevant individual-level predictions (12). We illustrated this potential by predicting treatment outcome using a model parameter estimated early in treatment.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by Fundação para a Ciência e a Tecnologia, Portugal (Ph.D. fellowship to AP; Grant No. SFRH/BD/52223/2013).

The authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Instituto de Medicina Molecular (AP, TVM), Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal; and the Department of Psychology (Clinical Psychology and Psychotherapy Research) (YS), PFH Private University of Applied Sciences, Göttingen, Germany.

Address correspondence to Tiago V. Maia, Ph.D., at tiago.v.maia@gmail.com.

Received Aug 19, 2020; revised and accepted Jan 13, 2021.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.bpsc.2021.01.005.

REFERENCES

1. Abramowitz JS, Deacon BJ, Whiteside SP (2019): Exposure Therapy for Anxiety: Principles and Practice, 2nd ed. New York: The Guilford Press.
2. Benito KG, Walther M (2015): Therapeutic process during exposure: Habituation model. J Obsessive Compuls Relat Disord 6:147–157.
3. McSweeney FK, Swindell S (2002): Common processes may contribute to extinction and habituation. J Gen Psychol 129:364–400.
4. Shiban Y, Pauli P, Mühlberger A (2013): Effect of multiple context exposure on renewal in spider phobia. Behav Res Ther 51:68–74.
5. Marks IM (1975): Behavioral treatments of phobic and obsessive compulsive disorders: A critical appraisal. In: Hersen M, Eisler RM, Miller PM, editors. Progress in Behavior Modification, 1st ed. New York: Academic Press, 75–168.
6. Foa EB, Kozak MJ (1986): Emotional processing of fear: Exposure to corrective information. Psychol Bull 99:20–35.
7. Bjork RA, Bjork EL (2006): Optimizing treatment and instruction: Implications of a new theory of disuse. In: Nilsson L-G, Ohta N, editors. Memory and Society: Psychological Perspectives. Psychology Press, 116–140.
8. Sripada RK, Rauch SAM (2015): Between-session and within-session habituation in prolonged exposure therapy for posttraumatic stress disorder: A hierarchical linear modeling approach. J Anxiety Disord 30:81–87.
9. Culver NC, Vervliet B, Craske MG (2015): Compound extinction: Using the Rescorla–Wagner model to maximize exposure therapy effects for anxiety disorders. Clin Psychol Sci 3:335–348.
10. Pine DS (2016): Clinical advances from a computational approach to anxiety. Biol Psychiatry 82:385–387.
11. Maia TV, Frank MJ (2011): From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci 14:154–162.
12. Huys QJM, Maia TV, Frank MJ (2016): Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci 19:404–413.
13. Maia TV (2015): Introduction to the series on computational psychiatry. Clin Psychol Sci 3:374–377.
14. Rankin CH, Abrams T, Barry RJ, Bhatnagar S, Clayton DF, Colombo J, *et al.* (2009): Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. Neurobiol Learn Mem 92:135–138.

15. Balooch SB, Neumann DL, Boschen MJ (2012): Extinction treatment in multiple contexts attenuates ABC renewal in humans. Behav Res Ther 50:604–609.
16. Andreatta M, Leombruni E, Glotzbach-Schoon E, Pauli P, Mühlberger A (2015): Generalization of contextual fear in humans. Behav Ther 46:583–596.
17. Thomas BL, Vurbic D, Novak C (2009): Extensive extinction in multiple contexts eliminates the renewal of conditioned fear in rats. Learn Motiv 40:147–159.
18. Mystkowski JL, Craske MG, Echiverri AM (2002): Treatment context and return of fear in spider phobia. Behav Ther 33:399–416.
19. Waters WF, McDonald DG (1976): Repeated habituation and over-habituation of the orienting response. Psychophysiology 13:231–235.
20. Bouton ME, Winterbauer NE, Todd TP (2012): Relapse processes after the extinction of instrumental learning: Renewal, resurgence, and reacquisition. Behav Processes 90:130–141.
21. Rinck M, Becker ES, Pössel P (2003): Fragebogen zur Angst vor Spinnen (FAS) [Questionnaire for Fear of Spiders (FSQ)]. In: Hoyer J, Margraf J, editors. Angst-diagnostik Grundlagen und Testverfahren. Berlin: Springer, 435–438.
22. Szymanski J, O'Donohue W (1995): Fear of Spiders Questionnaire. J Behav Ther Exp Psychiatry 26:31–34.
23. Wang D (1993): A neural model of synaptic plasticity underlying short-term and long-term habituation. Adapt Behav 2:111–129.
24. Stanley JC (1976): Computer simulation of a model of habituation. Nature 261:146–148.
25. Culver NC (2013): Extinction-based processes for enhancing the effectiveness of exposure therapy. Doctoral dissertation, University of California, Los Angeles, California.
26. Dayan P, Abbott LF (2001): Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Cambridge, MA: MIT Press.
27. Daunizeau J, Adam V, Rigoux L (2014): VBA: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. PLoS Comput Biol 10:e1003441.
28. Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014): Bayesian model selection for group studies – Revisited. Neuroimage 84:971–985.
29. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009): Bayesian model selection for group studies. Neuroimage 46:1004–1017.
30. Newell A, Rosenbloom PS (1981): Mechanisms of skill acquisition and the law of practice. In: Anderson JR, editor. Cognitive Skills and Their Acquisition. Hillsdale, NJ: Lawrence Erlbaum Associates, 1–55.
31. Heathcote A, Brown S (2000): The power law repealed: The case for an exponential law of practice. Psychon Bull Rev 7:185–207.
32. Newell KM, Mayer-Kress G, Liu Y-T (2006): Human learning: Power laws or multiple characteristic time scales? Tutor Quant Methods Psychol 2:66–76.
33. Craske MG, Kircanski K, Zelikowsky M, Mystkowski JL, Chowdhury N, Baker AS (2008): Optimizing inhibitory learning during exposure therapy. Behav Res Ther 46:5–27.
34. Lara R, Arbib MA (1985): A model of the neural mechanisms responsible for pattern recognition and stimulus specific habituation in toads. Biol Cybern 51:223–237.
35. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR (2013): Power failure: Why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376.
36. Steyerberg EW, Harrell FE (2003): Statistical models for prognostication. In: Max MB, Lynn J, editors. Interactive Textbook on Clinical Symptom Research. Available at: https://web.archive.org/web/20041031140843/http://painconsortium.nih.gov/symptomresearch/chapter_8/sec8/cess8pg2.htm. Accessed September 15, 2019.
37. Kandel E, Siegelbaum SA (2013): Cellular mechanisms of implicit memory storage and the biological basis of individuality. In: Kandel E, Schwartz JH, Jessell TM, Siegelbaum SA, Hudspeth AJ, editors. Principles of Neural Science, 5th ed. New York: McGraw-Hill, 1461–1486.
38. Plendl W, Wotjak CT (2010): Dissociation of within- and between-session extinction of conditioned fear. J Neurosci 30:4990–4998.
39. Gillihan SJ, Foa EB (2011): Fear extincion and emotional processing theory: A critical review. In: Schachtman TR, Reilly SS, editors. Associative Learning and Conditioning Theory: Human and Non-Human Applications. New York: Oxford University Press, 27–43.
40. Milad MR, Rosenbaum BL, Simon NM (2014): Neuroscience of fear extinction: Implications for assessment and treatment of fear-based and anxiety related disorders. Behav Res Ther 62:17–23.
41. Averell L, Heathcote A (2011): The form of the forgetting curve and the fate of memories. J Math Psychol 55:25–35.
42. Gershman SJ, Blei DM, Niv Y (2010): Context, learning, and extinction. Psychol Rev 117:197–209.
43. Pearce JM (1987): A model for stimulus generalization in Pavlovian conditioning. Psychol Rev 94:61–73.
44. Rudy JW, O'Reilly RC (2001): Conjunctive representations, the hippocampus, and contextual fear conditioning. Cogn Affect Behav Neurosci 1:66–82.
45. Dunsmoor JE, Niv Y, Daw N, Phelps EA (2015): Rethinking extinction. Neuron 88:47–63.
46. Morey RD (2008): Confidence intervals from normalized data: A correction to Cousineau (2005). Tutor Quant Methods Psychol 4:61–64.
47. Bolker BM (2008): Ecological Models and Data in R. Princeton, NJ: Princeton University Press.