RESEARCH ARTICLE

Methodological implications of sample size and extinction gradient on the robustness of fear conditioning across different analytic strategies

Luke J. Ney[1], Patrick A. F. Laing[2], Trevor Steward[3], Daniel V. Zuj[4]*, Simon Dymond[4,5], Ben Harrison[2], Bronwyn Graham[6], Kim L. Felmingham[3]

1 School of Psychological Sciences, University of Tasmania, Tasmania, Australia, **2** Melbourne Neuropsychiatry Centre, Department of Psychiatry, University of Melbourne & Melbourne Health, Victoria, Australia, **3** School of Psychological Sciences, University of Melbourne, Victoria, Australia, **4** School of Psychology, Swansea University, Wales, United Kingdom, **5** Department of Psychology, Reykjavik University, Reykjavik, Iceland, **6** School of Psychology, University of New South Wales, New South Wales, Australia

* d.v.zuj@swansea.ac.uk

Abstract

Fear conditioning paradigms are critical to understanding anxiety-related disorders, but studies use an inconsistent array of methods to quantify the same underlying learning process. We previously demonstrated that selection of trials from different stages of experimental phases and inconsistent use of average compared to trial-by-trial analysis can deliver significantly divergent outcomes, regardless of whether the data is analysed with extinction as a single effect, as a learning process over the course of the experiment, or in relation to acquisition learning. Since small sample sizes are attributed as sources of poor replicability in psychological science, in this study we aimed to investigate if changes in sample size influences the divergences that occur when different kinds of fear conditioning analyses are used. We analysed a large data set of fear acquisition and extinction learning (N = 379), measured via skin conductance responses (SCRs), which was resampled with replacement to create a wide range of bootstrapped databases (*N* = 30, *N* = 60, *N* = 120, *N* = 180, *N* = 240, *N* = 360, *N* = 480, *N* = 600, *N* = 720, *N* = 840, *N* = 960, *N* = 1080, *N* = 1200, *N* = 1500, *N* = 1750, *N* = 2000) and tested whether use of different analyses continued to produce deviating outcomes. We found that sample size did not significantly influence the effects of inconsistent analytic strategy when no group-level effect was included but found strategy-dependent effects when group-level effects were simulated. These findings suggest that confounds incurred by inconsistent analyses remain stable in the face of sample size variation, but only under specific circumstances with overall robustness strongly hinging on the relationship between experimental design and choice of analyses. This supports the view that such variations reflect a more fundamental confound in psychological science—the measurement of a single process by multiple methods.

OPEN ACCESS

Citation: Ney LJ, Laing PAF, Steward T, Zuj DV, Dymond S, Harrison B, et al. (2022) Methodological implications of sample size and extinction gradient on the robustness of fear conditioning across different analytic strategies. PLoS ONE 17(5): e0268814. https://doi.org/10.1371/journal.pone.0268814

Editor: Marta Andreatta, Erasmus University Rotterdam: Erasmus Universiteit Rotterdam, GERMANY

Received: June 17, 2021

Accepted: May 10, 2022

Published: May 24, 2022

Copyright: © 2022 Ney et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data are available on OSF: https://osf.io/8asxw/.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Fear conditioning paradigms are critical to understanding and improving treatment for several psychiatric disorders, including post-traumatic stress disorder (PTSD) and anxiety [1, 2]. Fear extinction occurs when a previously conditioned fear stimulus (conditioned stimulus, CS+) is repeatedly presented without aversive reinforcement, causing new safety information to compete with pre-existing fear memory [3–5]. Patients with anxiety-related disorders show deficits in extinction learning, which is believed to facilitate disease progression and maintenance [6, 7]. The rate of an individual's fear extinction learning can be estimated by the decrease in threat response to the unreinforced CS+ when compared to the safety signal (CS-), typically indexed via various physiological measures [8], with skin conductance responses (SCRs) being most commonly used. Extinction learning has been subject to extensive research on its neurobiological basis [9–19], and serves as a highly informative framework for investigating pharmacological and psychological adjuncts to exposure therapy for PTSD, and deficits associated with treatment outcomes [20–27].

The replicability crisis has inspired a growing movement dedicated to improving the quality of research practices in psychological science [28–32]. These issues of replicability extend to research on human fear conditioning [8]. Importantly, inconsistent research practices in fear conditioning might explain the contradictory and null outcomes identified across recent large-scale studies and meta-analyses [7, 33–36]. These limitations have been identified for several methodological domains including, but not limited to, study design [37, 38], pre-processing of psychophysiological data [39–42], and statistical analysis strategies [43–46]. It is increasingly clear that issues such as these undermine the replicability of fear conditioning research, and the subsequent translation of experimental findings to clinical outcomes.

Our previous report [54] was concerned with the effect of analytic strategy on robustness. Simply put, 'robustness' in psychology refers to the ability for a result to be consistent across multiple arbitrary statistical specifications [28]. In our case, arbitrary specifications associated with inconsistency in analytical strategies and we demonstrated stark divergence of effect sizes when different statistical methods were used to index extinction [47]. Specifically, a large data set was resampled to create 40 data sets of $N = 60$ rows with three groups per sample. Different statistical strategies, all intending to measure extinction, were compared against each other across the 40 data sets, but varied with respect to the numbers of trials included the stages of the phases the trials were drawn from, and whether the data was analysed trial-by-trial or averaged. We tested the effect of these variations on robustness of studies that compared acquisition learning to extinction learning, change in responding during extinction (e.g., early to late extinction), and where extinction was treated as a single effect estimate. We showed that the rank order of these strategies varied significantly depending on the data set, which illustrates less than desirable robustness of these statistical tests [47]. However, solutions to the issue of inconsistent analytic strategy remain unexplored.

In the current study, we aimed to investigate one plausible solution—increased sample size. Increasing sample size will increase the power of a study—that is, the ability to detect a specific effect size within a sample. It has been observed that many fear conditioning studies may be underpowered due to low samples [39] and it is possible that improving the precision of physiological measures through more advanced pre-processing could be sufficient to improve robustness of fear conditioning and extinction outcomes [39–41, 48]. By increasing power, we increase the probability of detecting the effect, and it is possible that heterogeneity of outcomes can be caused by underpowered studies that do not accurately capture this effect. However, heterogeneous statistical analyses have been reported to produce misleading or false results independent of power considerations [28, 49].

To investigate if larger samples could address the analytical issue we previously identified, we bootstrapped data from existing data sets, obtaining rank orderings of previously used statistical methods for indexing fear acquisition and extinction [47]. In our previous study, each resampled data set had a sample size of $N = 60$ rows, broken into three groups during analysis. Here, we resampled from our real data ($N = 379$) with replacement to create bootstrapped samples of $N = 30$, 60, 120, 180, 240, 360, 480, 600, 720, 840, 960, 1080, 1200, 1500, 1750, and 2000 observations, with each row being equivalent to one subject. These numbers were chosen to cover a broad range of plausible sample sizes used in human fear conditioning research. We performed two experiments—in the first, group allocation was randomised, and no group-level effect was anticipated. In the second, we added a group-level effect to our bootstrapped data. The group level effect was varied across three conditions, which were roughly based on [50] with one group who had high responding during acquisition and rapid extinction, another group who had lower responding during acquisition and rapid extinction, and another group who did not extinguish the CS+ response. This work represents a significant contribution above our previous study because (a) we create much larger samples spanning a wide range of simulated sample sizes; and (b) we test the results of this and our previous work against the presence of a simulated group-level effect. Testing our hypotheses with the inclusion of simulated group-level effects is a significant contribution because most fear conditioning studies will observe group differences and our original analyses were likely not representative of these studies; hence, the effect of heterogeneity of analytical methods in studies with groups effects is unknown. In this extension of our previous work, we therefore aimed to identify possible boundary conditions of an originally bleak report of the robustness of statistical analysis pipelines for fear conditioning research.

We hypothesised that larger samples would not improve robustness of rank ordering between analytic strategies in either condition, because we believe that the issue of analytical heterogeneity is a fundamental violation of replicability that cannot be solved by increasing power alone. We hypothesised that the type of simulated effect would vary the robustness of different statistical strategies because some strategies are used to examine different stages of learning during fear conditioning tasks.

Methods

The current manuscript uses secondary data analysis strategies on existing datasets, and did not require further ethical approval. The original studies received ethical approval from the University of Tasmania Social Sciences Human Research Ethics Committee. The fear acquisition and extinction procedures, as well the data set, for this study are identical to those of our previous study [47]. Briefly, six data sets gathered over seven years were resampled with replacement to form new samples. Participants reported no significant physical illnesses, no history of head trauma or loss of consciousness, no current or significant historical use of illicit substances, and no heavy alcohol use or dependence. Of the 379 participants included in this dataset, $N = 51$ (13.46%) had a diagnosis of PTSD (clinician diagnosed) or had a score above 40 on the PCL-IV or above 30 on the PCL-5 [51, 52]. No other psychiatric diagnoses were permitted in any of the studies. PTSD cases were retained in the sample in order to remain consistent with the previous study [47]. Since the predictor variable in these studies are the analytical strategies themselves, it is unlikely that systemic variability in participant characteristics would affect results [47].

Fear conditioning paradigm and equipment

As in our previous report [47], data was obtained from five trials of acquisition learning and ten trials of extinction learning (split into early and late extinction phases of five trials each, which were separated by an instruction screen) across a total of 379 participants across the six studies. Acquisition and extinction phases were also separated by an instruction screen, which in all cases read "In the following phase, you may or may not receive shocks. Please press any key to continue". For each trial, a CS+ (a coloured circle) and a CS- (a different coloured circle) were presented on a computer screen for 12s with intertrial intervals of 12-21s (*M* = 16s). In all studies, skin conductance was recorded from the first and third fingers of the left hand in micro-Siemens (μS) using a 22 mVrms, 75 Hz constant-voltage coupler (ADInstruments). A stimulus isolator (ADInstruments) was placed on the right hand and delivered a 500ms electric shock immediately following the CS+ offset during acquisition learning. No shocks were delivered during the extinction learning phase. Skin conductance responses (SCRs) were scored using a custom-coded peak scoring method which subtracts the average skin level 2s prior to CS onset from the peak conductance occurring 0.9-5s following CS onset, which scores the first interval response, and it should be noted that studies score skin conductance responding differently [53]. A bidirectional Butterworth filter was applied to the raw SCR trace to reduce noise.

Resampling procedure

Data was bootstrapped (i.e., resampled with replacement) using rows of participant data [54]. Using bootstrapping, it is possible to validate the accuracy of statistical techniques across a range of sample sizes, and this has been done in previous literature assessing the effect of sample size on correlation, factor analysis, principal components analysis, prognostic modelling, and other statistical techniques [55–58]. Data was resampled by row such that all CS+ or CS- responses from a particular phase (e.g., acquisition) were resampled together. New data sets of *N* = 30, *N* = 60, *N* = 120, *N* = 180, *N* = 240, *N* = 360, *N* = 480, *N* = 600, *N* = 720, *N* = 840, *N* = 960, *N* = 1080, *N* = 1200, *N* = 1500, *N* = 1750, and *N* = 2000 rows were created and a 'Group' variable consisting of equal but random allocation of belonging to the number 1, 2, or 3. Therefore, no group-level effects were expected in this analysis. Sample sizes were chosen to cover a wide range of possible study power in the simulated datasets. These sample sizes were determined arbitrarily due to current debate concerning accurate power determination of fear conditioning research using skin conductance responding [39]. Three groups were used because in our field of research (PTSD) it is typical to examine a PTSD group against both a trauma-exposed control and a non-trauma exposed control group [59].

For the second experiment, scores were modified for the third Group upon bootstrapping such that a higher but gradually decreasing CS+ response (relative to CS- response) was expected in each phase. Data was produced that resembled the three fear conditioning trajectories reported by [50]. These trajectories were replicated in our own clinical fear conditioning data (manuscript in preparation), and group-level simulated effects were created in the data from the current report based on the difference between each of the three trajectories and our bootstrapped data that did not have a simulated group-level effect. The modifications to produce the simulated effects are described below. Scores for Group 3's CS+ were modified to be 1, 0.8, 0.6, 0.4 and 0 standard deviations higher than their bootstrapped values during trials 1–5 of acquisition; 2, 1.5, 1, 0.8, 0.5 standard deviations higher than their bootstrapped values during trials 1–5 of early extinction; and 1, 0.8, 0.5, 0.2, and 0 standard deviations higher than their bootstrapped values during trials 1–5 of late extinction. Two other distinct group-level effects were simulated for Group 3, with CS+ modified to be 0, 0.3, 0.3, 0.3, and 0.3 standard deviations higher during acquisition, 1, 1, 1, 1, and 1 standard deviations higher during early extinction, and 1.5, 1, 0.5, 0.1, and 0 standard deviations during late extinction higher than the average data for Group 3a; and 0, 0, 0.3, 0.3, and 0.3 standard deviations higher during acquisition, 2, 1.5, 1, 0.5, and 0.3 standard deviations higher during early extinction, and 1.5, 1, 0.5, 0.1, and 0 standard deviations during late extinction higher than the average data for Group 3b. These simulated effects were achieved by adding the same value (e.g., 2 standard deviations above the mean score for trial 1) to all scores individually within that group. Therefore, all analyses in this study were conducted three times with Group 3 consisting of one of the three sets of simulated effects. An illustration of an example of this data is provided in Fig 1. Simulated effects were roughly based on the findings of Galatzer-Levy et al. (2017) [56], who identified three distinct trajectories during acquisition and extinction phases in fear conditioning data. In Fig 1, Group 3 is the trajectory that shows high differential acquisition and rapid extinction, Group 3a is the trajectory that shows moderate differential acquisition and rapid extinction, and Group 3b is the group that does not show extinction.

Types of analytical strategies included in comparisons

Further analyses were conducted using base R. Analytic strategies were identical to those used previously [47] and are summarised in Table 1. As described previously, some strategies averaged trials or subtracted CS- from CS+ scores, whereas others did not. These details are described in Table 1. The goal of these strategies was to either: (1) determine the change in SCRs from acquisition to extinction learning (CON-EXT); or (2) determine a static measure of extinction learning (EXT) or (3) determine the change in SCRs across the extinction learning phase (EXT-EXT). Since the goals of the strategies differed in these ways, we divided strategies into each of these categories and compared outcomes only within each category.

Data analysis

For each strategy, we compared the highest order group-level interaction via its computed partial eta squared (ηp^2) effect size. For each sample size, bootstrapped (1,000 times) Kendall non-parametric ranked order correlation coefficients (${}_{T}b$) and associated 95% bootstrapped confidence intervals were computed between analytical strategies of each of the three categories, based on the ηp^2 effect sizes generated. Therefore, each sample size (e.g. 30 "participants") was resampled 100 times to generate a rank order (${}_{T}b$) of ηp^2 across the different analysis strategies, and this procedure was bootstrapped 1,000 times to generate mean ${}_{T}b$ and 95% confidence intervals. The mean ${}_{T}b$ and its associated confidence intervals were the average correlation between one strategy and each of the other strategies separately (e.g., creating three mean ${}_{T}b$ values for Strategy 1 of the acquisition—extinction category). This entire procedure was completed using a custom R script that is available from the authors upon request. The data was compiled and is reported in the Supplementary Material up until $N = 960$. Data beyond this size is not reported due to excessive amount of the data reported in the manuscript and because the results at $N>960$ were almost identical to those obtained at $N = 960$. Using the average ${}_{T}b$ effect size of each strategy, we tested whether the rank order coefficients improved with increased sample size using Pearson's coefficient (r). This was completed for both the first (no group-level effect) and second (simulated group-level effect) experiments.

During data compilation, it was evident that there were large decreases in effect sizes with increased sample size ($p < .001$). As an exploratory analysis, effect sizes averaged across sample sizes for each category of analytical strategy were compared using Pearson's correlations (r). To ensure that the effects observed in this exploratory test were not due to variability caused by our resampling process (where CS+ or CS- scores for each participant from only one phase were resampled), we resampled using the full data from each participant to create data sets of $N = 30$, $N = 60$, $N = 120$, $N = 240$ rows, with three equally sized groups randomly allocated amongst these rows. Samples were not created that were larger than the number of actual

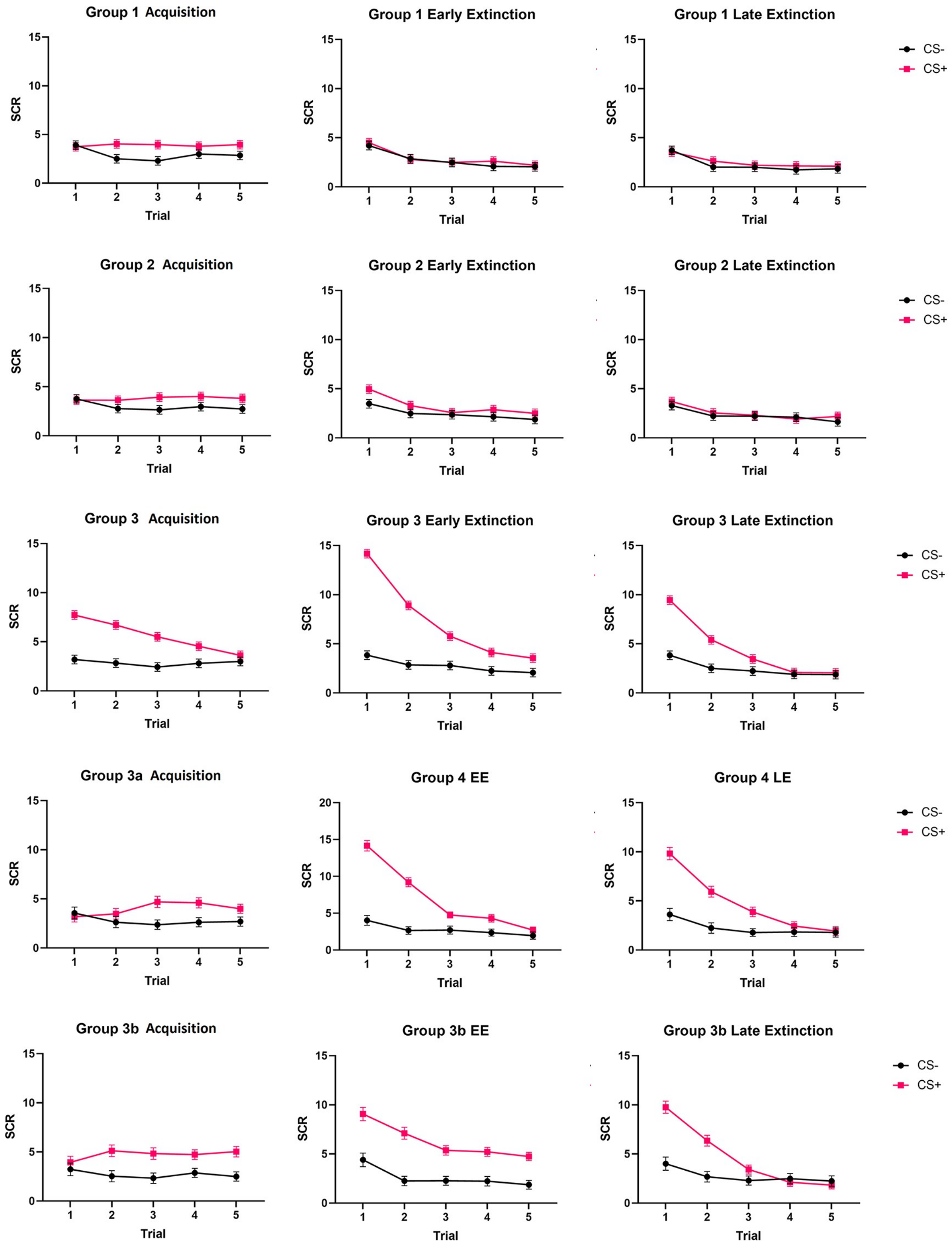

Fig 1. Example of simulated group-level effects in bootstrapped data (N = 960). SCR = Skin conductance response. Groups 3-3b have simulated effects of differing gradients to reflect possible differences in physiological expression of acquisition and extinction between participants and between studies. Error bars are 95% Confidence Intervals.

https://doi.org/10.1371/journal.pone.0268814.g001

Table 1. Description of different strategies for measuring extinction learning using skin conductance responses (Ney et al., 2020).

Analytic strategy	Strategy #	# of Trials	Trials Included	Trial Analysis	Stimuli Analysis	Analysis	Study
ACQ—EXT	Strategy 1	8 (ACQ), 16 (EXT)	All (ACQ), last 2 (EXT)	Average	Diff	Phase×group	[60]
	Strategy 2	5 (ACQ), 10 (EXT)	Maximum Response (ACQ), Last 2 (EXT)	Average	Diff	Phase×group	[61]
	Strategy 3	8 (ACQ), 7 (EXT)	All (ACQ), last 3 (EXT)	Average	Diff	Phase×group	[62]
	Strategy 4	20 (ACQ), 20 (EXT)	Last half (ACQ), First half (EXT)	Average, using paired t-test contrasts^	Diff	Phase×group	[63]
EXT	Strategy 1	16	Last three-quarters	Average	CS+, CS-	Group×stim	[64]
	Strategy 2	5	All	Trial-by-trial	CS+, CS-	Trial×Group×Stim	[65]
	Strategy 3	16	Last half	Average	CS+, CS-	Group×stim	[66]
	Strategy 4	10	Last trial	One trial	Diff	Group	[67]
	Strategy 5	10	Last 2	Average	CS+, CS-	Group×stim	[68]
	Strategy 6	5	All	Running average[#]	Diff	Trial×Group	[69]
	Strategy 7	8	First 2	Trial-by-trial	Diff	Trial×Group	
$\text{EXT}_{\text{early}}$-$\text{EXT}_{\text{late}}$	Strategy 1	6	First half, second half	Average	CS+, CS-	Phase×Group×Stim	[70]
	Strategy 2	14	First half, second half	Average	Diff	Phase×Group	[71, 72]
	Strategy 3	16	First quarter, last quarter	Average	CS+	Phase×Group	[73]
	Strategy 4	32, 16	First half, second half	Average	CS+	Phase×Group	[74, 75]

ACQ = Acquisition, EXT = Extinction, Diff = Differential, CS+ = Conditioned stimulus to the aversive unconditioned stimulus, CS- = Conditioned stimulus as a safety signal, Stim = stimulus type (CS+ v. CS-).

^This study was the only study to use a test other than ANOVA.

[#] Running average response was calculated with trials one and two averaged as a single response, trials two and three averaged, and so on.

https://doi.org/10.1371/journal.pone.0268814.t001

participants to avoid repeating participant data in the same sample. Again, the ηp^2 effect sizes from each category of analytical strategy were averaged and compared across sample size.

Results

The overall data from the original sample ($N = 379$) is reported in S1 Fig. The main index that was used as an outcome in the present study was the rank order of effect sizes produced by different analytical (i.e., statistical) approaches when applied to the same dataset. To ensure that this result was robust, datasets were bootstrapped so that the analysis was repeated many times. If a low rank order effect is produced, this implies that application of different analytical approaches to the same datasets produces inconsistent effect sizes relative to the other approaches. A high rank order effect suggests that application of different approaches to the same datasets produces consistent effect sizes relative to the other approaches, which implies robustness. To assess the robustness of each analytical method within each bootstrapped dataset, Kendall's rank correlation coefficient values ($_{\tau}b$) and corresponding 95% confidence intervals were computed for each of the three sets of analyses with sample size set to $N = 30$, $N = 60$, $N = 120$, $N = 180$, $N = 240$, $N = 360$, $N = 480$, $N = 600$, $N = 720$, $N = 840$, $N = 960$, $N = 1080$, $N = 1200$, $N = 1500$, $N = 1750$, and $N = 2000$ rows. Complete statistics from an exemplar of these analyses are reported in S1–S42 Tables and are summarised in Fig 2. We also entered the

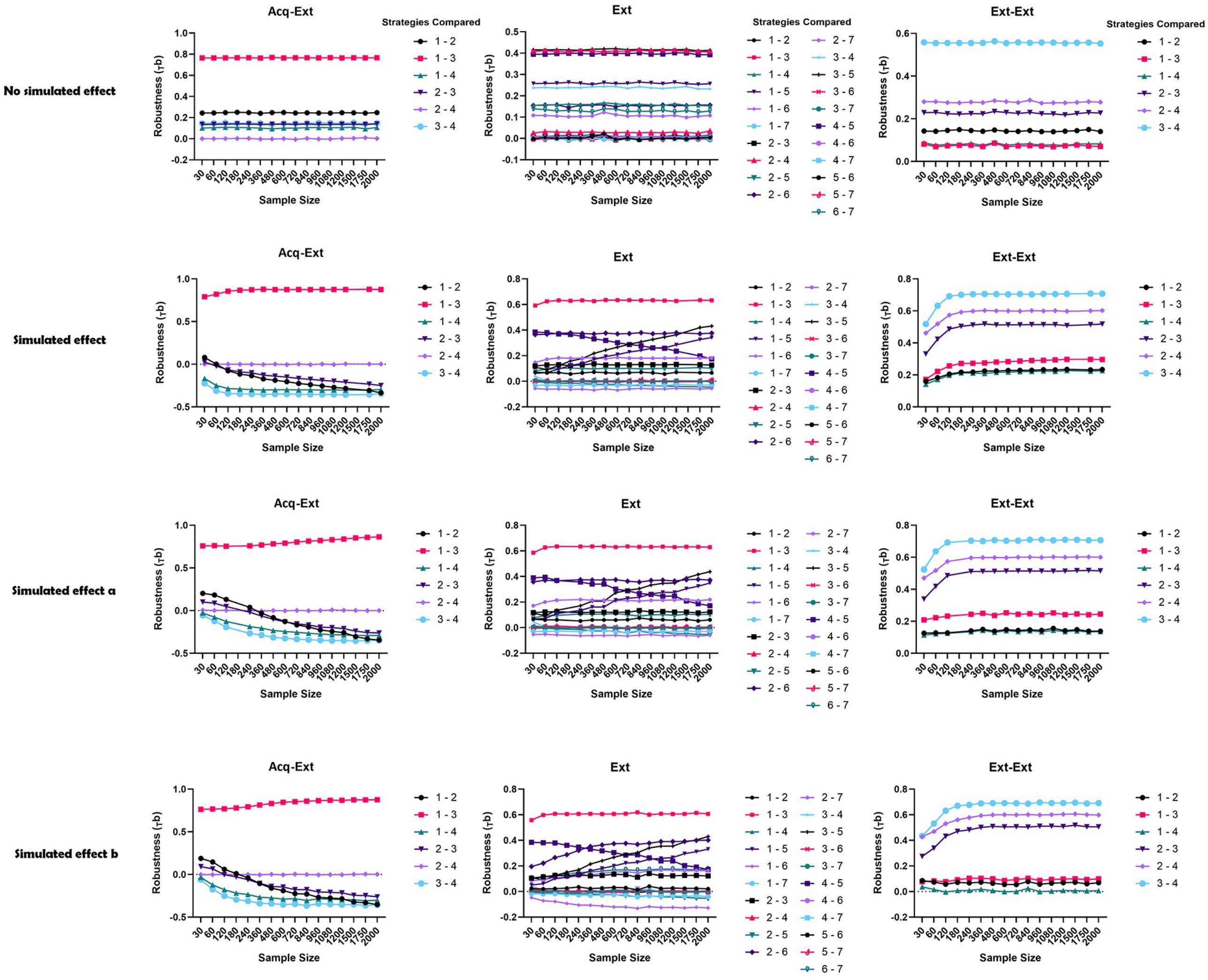

Fig 2. Effect of sample size on average Kendall's rank order effect size ($_Tb$) between statistical strategies attempting to elicit the same construct from different data sets. Higher $_Tb$ implies higher robustness. Top panel is data without simulated group-level effect, second panel simulates rapid decreasing differential conditioning during acquisition, third panel simulated gradual decrease in differential conditioning during acquisition, fourth panel simulated no change in differential conditioning during acquisition or early extinction.

https://doi.org/10.1371/journal.pone.0268814.g002

rank order for each analytical strategy compared to every other strategy into Pearson correlation models across each sample size. This data is visualised in Fig 2 and reported in Table 2.

Overall, findings for non-simulated effect datasets are congruent with our previous findings [54], which was conducted with a sample size of $N = 60$. There were no significant trends in the data for the no-effect data (summarised in Table 2), which suggests that increasing sample size did not improve robustness caused by variability in analytical strategies used to assess similar constructs in the same data.

ACQ-EXT

Strategies 1 and 3, which compared acquisition to extinction, produced high correlative values across sample sizes, whereas Strategies 2 and 4 were not similar to any Strategies (S22–S28 Tables). This finding replicated our findings from our previous report at $N = 60$.

Table 2. Pearson's correlation coefficient and significance of the relationship between sample size and rank order between different statistical strategies used to index static extinction (EXT), change in extinction (EXT-EXT) and acquisition to extinction (ACQ-EXT) during fear learning paradigms.

		Strategies Compared (EXT)																				
		1–2	1–3	1–4	1–5	1–6	1–7	2–3	2–4	2–5	2–6	2–7	3–4	3–5	3–6	3–7	4–5	4–6	4–7	5–6	5–7	6–7
		Strategies Compared (ACQ-EXT, EXT-EXT)																				
		1–2	1–3	1–4	2–3	2–4	3–4															
No simulated effect																						
ACQ-EXT	*r*	-.096	.221	-.023	.063	.348	-.030															
	p	.723	.411	.933	.817	.187	.913															
EXT	*r*	-.150	-.060	-.396	-.220	.050	.010	-.157	.087	-.031	.171	-.166	-.356	-.325	-.009	-.282	-.188	.049	-.589*	.229	-.069	-.467
	p	.579	.825	.128	.414	.855	.971	.561	.749	.910	.528	.539	.176	.219	.972	.290	.485	.858	.016	.394	.800	.068
EXT-EXT	*r*	-.071	-.265	-.028	-.088	-.078	-.209															
	p	.794	.321	.917	.746	.775	.437															
Simulated effect																						
ACQ-EXT	*r*	-.869**	.529*	-.455	-.901**	.146	-.431															
	p	< .001	.043	.089	< .001	.602	.109															
EXT	*r*	-.032	.391	-.507	.955**	.346	.217	.234	-.043	-.947**	.331	.353	-.359	.961**	.215	.034	-.990**	-.010	.094	-.104	.189	.515*
	p	.910	.150	.054	< .001	.206	.438	.400	.880	< .001	.228	.197	.188	< .001	.442	.905	< .001	.971	.740	.712	.500	.049
EXT-EXT	*r*	.661**	.662**	.641*	.482	.499	.448															
	p	.007	.007	.010	.069	.058	.094															
Simulated effect a																						
ACQ-EXT	*r*	-.932**	.981**	-.810**	-.919**	.009	-.736**															
	p	< .001	< .001	< .001	< .001	.976	.002															
EXT	*r*	-.077	.275	-.288	.964**	-.427	-.082	.057	-.543*	-.931**	.293	.458	-.375	.968**	-.177	-.217	-.986**	-.180	-.132	-.026	.047	.339
	p	.786	.321	.298	< .001	.112	.771	.839	.037	< .001	.289	.086	.169	< .001	.527	.437	< .001	.521	.640	.928	.869	.216
EXT-EXT	*r*	.407	.496	.482	.540*	.562*	.492															
	p	.132	.060	.069	.038	.029	.062															
Simulated effect b																						
ACQ-EXT	*r*	-.892**	.884**	-.697**	-.893**	.405	-.622*															
	p	< .001	< .001	.003	< .001	.120	.010															
EXT	*r*	-.078	.402	-.326	.958**	-.747**	-.777**	.203	-.433	-.969**	.794**	.833**	-.429	.958**	-.416	-.071	-.989**	.417	.057	-.289	-.167	.777**
	p	.773	.122	.218	< .001	.001	< .001	.451	.094	< .001	< .001	< .001	.098	< .001	.109	.795	< .001	.108	.833	.277	.537	< .001
EXT-EXT	*r*	-.204	.412	-.282	.587*	.619*	.531*															
	p	.449	.113	.290	.017	.011	.034															

Note: EXT = Static Extinction, ACQ-EXT = acquisition to extinction, EXT-EXT = early to late extinction.

*p < .05,

**p < .001

https://doi.org/10.1371/journal.pone.0268814.t002

When a group-level effect was simulated, however, these results changed. Only Strategies 1 and 3 showed positive but increasingly weak correlative improvements in the acquisition with increasing sample size (Fig 2 and S1–S7 Tables), whereas combinations of other strategies were increasingly significantly and negatively correlated with increased sample sizes, meaning that they estimated fear responding in opposite directions to one another and that this pattern got worse with a larger sample (Fig 2 and Table 2).

EXT

Some of the correlations between the static extinction strategies failed to be supported compared to our previous study in the data without simulated group level effects (S29–S35 Tables). These were mainly between strategies 1, 4 and 5, which were not supported in data derived from the new data sets but had been correlated in our previous report. Correlations between Strategies 1 and 3; 2, 6 and 7; and 5, 3, and 4 continued to be supported of the static extinction strategies (S29–S35 Tables and Fig 2).

There were very few supported correlations in static extinction Strategies when group effects were simulated (i.e., high ${}_{T}b$ values, primarily correlations between 1 and 3 were supported), but some of these improved with increased sample size (Fig 2, Table 2, and S8–S14 Tables). Correlations between Strategies 1, 3, and 5 for static extinction improved significantly with increased sample size, whereas correlations between Strategies 2, 4, and 5 were significantly negatively correlated with increasing sample size. Other combinations of strategies showed no change with increasing sample size (Fig 2 and Table 2).

EXT-EXT

At higher sample sizes, some of the significant correlations from our earlier study [47] within the early-late extinction strategies were no longer significant, though this did not follow a particularly consistent pattern (S40–S42 Tables). In all cases, Strategies 3 and 4 of the early-late extinction category continued to be correlated (Fig 2 and S36–S42 Tables), but this did not improve with higher sample size (Table 2).

Strategies 2 and 4, 3 and 4, as well as 2 and 3 of early-late extinction showed some moderate-high evidence of correlation that improved logarithmically when group level effects were simulated (Fig 2 and S15–S21 Tables). Unexpectedly, these results were not substantially affected by the type of simulated effect (Fig 2). However, only Strategies correlating with Strategy 1 from early-late extinction changed by improving with increased sample size, after correcting for multiple comparisons using False Discovery Rate Q = .1.

Sample size and average effect sizes

During data compilation, we noticed large decreases in effect sizes with increased sample size. As an exploratory analysis, we correlated the average effect size (ηp^2) from each category of analytical strategies with the sample size. The average effect size for each set of analyses decreased significantly as a function of sample size (all $p < .001$), as shown by Fig 3A. Effect sizes of all three types of analyses reduced at a similar rate. This effect was replicated when the data was resampled from full participant rows (i.e., in real data, Fig 3B).

Discussion

In this study we investigated whether the decreased robustness that arises from inconsistent analytic strategy [54] could be amended by increased sample sizes. To do so, we tested whether greater sample sizes affected the robustness of outcomes via lower divergence of results

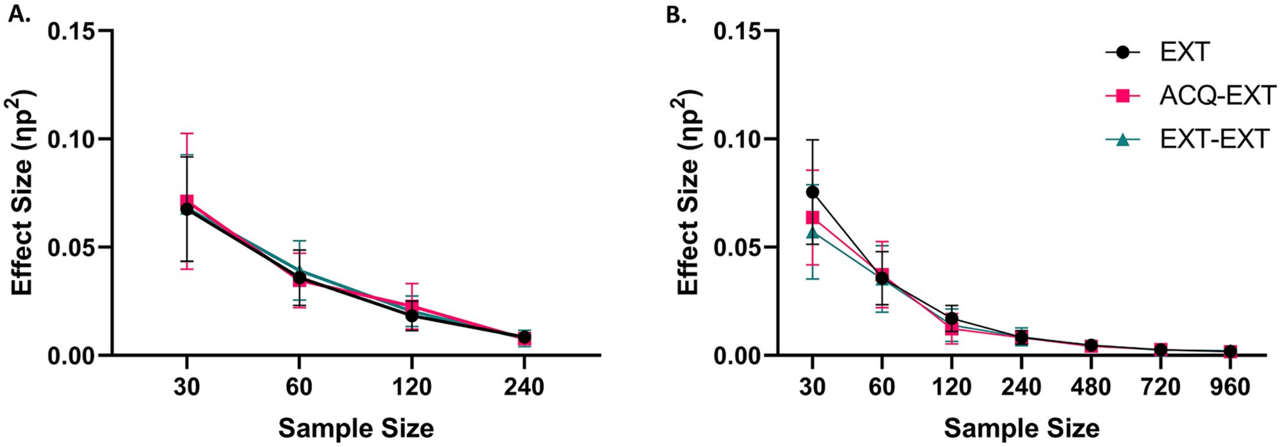

Fig 3. Average effect size decreased as sample size increased for all types of analyses ($p < .001$). Panel A is the correlation using resampled data of responses. Panel B is the correlation using resampled data of responses. Error bars are 95% Confidence Intervals.

https://doi.org/10.1371/journal.pone.0268814.g003

obtained across varied analytical strategies. Robustness did not improve when sample size was increased for any of the strategies included in our analysis that did not include a simulated effect. However, in contrast to our hypothesis, a simulated effect resulted in several changes in robustness, particularly within strategies that examined extinction as a single index. The kind of effect that was simulated (in terms of the gradient of fear responding across trials) did not substantially affect these results. These findings have several implications for study design and statistical analysis of fear extinction via SCRs.

Our previous study provided evidence that heterogeneity in analytical strategy in the assessment of fear extinction can reduce robustness of effects when tested across different data sets [47]. This problem has been reported in other fields, such as human neuroimaging [76, 77], and high flexibility in data analysis is an established cause of increased false positives [49]. It is possible, however, that some types of strategies produce more robust results than others. The current findings support several assertions that we made in our previous paper in this regard. First, studies that examine change in SCRs from acquisition to extinction will show varying robustness depending on what sections of acquisition and extinction are used, but robustness does not seem to be affected if small variations in number of trials or use of averaged compared to maximal values are used. Similarly, analysing extinction on a trial-by-trial basis is inconsistent with strategies that averaged across trials, but both strategies are internally consistent regardless of the number of trials included, or whether differential responses (CS+ > CS-) were calculated. Finally, we found mixed evidence that number of trials and use of differential responding affects robustness of strategies examining change during extinction, which were associations that we had previously identified as having moderate support [47].

The main aim of the current study was to understand whether improving power, by increasing sample size, would improve robustness that is affected by heterogeneity of analytical strategies. As we had anticipated, the limitations imposed by varied analytic strategies holds even when applied to samples of greater size, but this appeared to apply only when no effect was present in the data. While this supports the validity of our prior study [54], it also challenges our previous findings in several ways. Firstly, in the data containing simulated group-level effects, some strategies improved markedly in terms of robustness as sample size increased. However, these cases were contradicted by several other strategies that showed weaker robustness with increasing sample size. Most importantly, not all strategies showed these patterns, and marked improvement in robustness in the static-extinction strategies was primarily observed at the higher sample sizes, which research groups would not have the capacity to collect. Improving robustness of strategies examining changes in fear responding from early to late extinction might be achievable by increasing sample size to an amount that is viable with respect to research resources. Critically, these results were not substantially affected by alteration of the gradient of the group-level effect. This implies that it is possible that an improved set of data analysis strategies for fear extinction data could be applied robustly across fear extinction phenotypes, which were recently identified in [50]. These findings provide critical boundary conditions and caveats to our previous findings, and we strongly emphasise that not all analytical approaches that we highlighted as problematic (or robust) in our previous report will be applicable to all real-world samples. Critically, by extension our current data also demonstrates that a statistical approach that seems unrobust in this study could be robust if the underlying effect is different.

This research is important because sample considerations are frequently the first criticism addressed in experimental psychology research, and supports the notion that further methodological innovations are required to enhance fear extinction research, beyond simply increasing study power [43]. Several research groups have begun moving towards Bayesian inference in fear extinction [78–81] and computational modelling has also been explored in assessing physiological responses to fear conditioning [45]. It is possible that these contemporary statistical frameworks may offer solutions to the deficits imposed by heterogeneous analytic strategies in extinction research. However, further research is needed to explore this as a viable possibility to conventional data analysis, particularly in terms of accessibility to non-statisticians.

While compiling the data in the current study, we observed strong effects of lower sample size resulting in higher effect size. It is likely that the reduced effect size we observed with increasing sample size reflected increasing precision of the effect size, which is reflected in the increasingly narrow confidence intervals. The relationship between smaller samples reaching significance with higher effect sizes is intrinsically wedded to the parameters of power analysis in null-hypothesis significance testing (NHST) [82]. In a simple case, when performing *a-priori* power analysis (to determine an appropriate sample size), specifying higher r (e.g., effect size) and a significance criterion (α) of $p < 0.05$ will result in a generally lower N, all things being equal [83]. The propensity for studies with small sample sizes to inflate effect sizes is well documented [84–86]. This is sometimes attributed to publication bias [85], but in the context of the current study, higher variability in our smaller samples is a likely cause, as indicated by 95% confidence intervals. Our findings suggest that these issues are likely to be prevalent until a minimum of $n = 40$ participants per group for a 3-group design (which may vary depending on the number of groups). However, it has been reported that this estimate is improved by advanced SCR scoring methods [39]. Interestingly, the inflection point of logarithmic improvement in some of the strategies in terms of robustness was at this same sample size, raising the possibility that there may be some relationship between adequately powered data and the propensity for certain strategies (mainly the early-late extinction strategies) to perform robustly. Relatedly, power analyses using single point estimates from previous fear conditioning studies is likely problematic given that heterogeneity of experimental parameters and effect sizes that are chosen by researchers affect power calculations. Instead, it might be more useful to estimate the expected variability and build a power analysis based on the precision of the anticipated effect size (i.e., the effect size's confidence interval) [87].

Considering this finding, as well as the overall results of the current report, we suggest two implications for the enhancement of robustness in fear extinction research via SCR. First, in line with our prior report (Ney et al., 2020), it is critical that a specific analytic strategy is implemented only when the experimenter seeks to measure a specific aspect of fear extinction, one that corresponds clearly to the strategy in question. For instance, some of the analytic strategies identified in this and the prior study [54] can credibly be used to measure distinct aspects of extinction learning. For instance, subtracting early and late extinction responses might represent a principled measure of extinction learning per se, while subtracting mean extinction responses from mean acquisition responses could represent something quite different, albeit equally worthy of investigation. Critically, if these different strategies are used, it is incumbent on the experimenter to interpret the results consistently. Labelling all different strategies under a homogenised term (i.e., 'extinction learning') could otherwise incur costs to robustness, and ultimately, failures to replicate. Similarly, it is important that standardised methods for comparing extinction between group relative to acquisition learning are developed, because there is significant heterogeneity in current methods that do this [46], yet some relative estimation is essential given that the effects observed during extinction are often contingent on responses during acquisition.

Second, this study illustrates that the pervasive issue of measuring one construct by a diverse array of analyses remains an issue even in the face of some methodological changes, in this case, sample size. An implication of this is that other methodological changes may also be unable to ameliorate this effect, but more critically, that future research should strive to find ways to analysis extinction learning that circumvent the effect altogether. In other words, analysing data in different ways will almost always lead to different outcomes, and reduced robustness or replicability. Therefore, rather than finding ways to homogenise between different analytic strategies as a path forward, ongoing work could seek to characterise extinction via more principled quantitative approaches. It is critical to consider that fear acquisition and extinction are multifaceted processes that cannot be captured by a single parameter. In many cases, researchers will make different statistical decisions based on the type of learning process that they are interested in—for example, analysing data trial-by-trial may assess the rate of learning, whereas comparing mean responses during extinction to acquisition might assess someone's relative performance between phases. One way of addressing the propensity for different studies to use different types of analyses is to use multiverse approaches. Multiverse analysis is an approach that assesses a statistical problem with multiple analytical methods [88]. In fear conditioning, multiverse packages have been written for R [89], and can potentially directly address the issues highlighted within this paper by increasing transparency of statistical decision making as well as the relative importance of a reported result [53]. In this way, not only does multiverse analysis reduce the potential of p-hacking, but it also facilitates comparison between studies that may have otherwise analysed their results in incomparable ways. Similar to this, it is almost certain based on this and recent data that different experimental designs (e.g., number of trials, induction of uncertainty via instructions, etc) are likely to produce different outcomes that may not be readily comparable between studies. We are aware of current work aiming to produce 'typical' fear conditioning experiments that may help to standardise the field, but in the meantime it is also possible that further investigation of the relationship between specific statistical analyses and experimental designs may help to improve the comparability of findings between fear conditioning studies.

The current study is primarily limited by the possibility of our findings not generalising to other fear extinction designs. For instance, we have a relatively low number of trials and long-duration stimuli (12 s), which are not the case for many studies. Further, these results may not be transferable to different data pre-processing methods and will need to be checked independently by groups that use these methods. One issue that we did not explicitly examine was the effect of number of trials on statistical outcomes—however, it is probable that the number of trials included in a study presents another significant heterogeneity factor that, when analysed using similar methods, may reduce robustness. Our experimental phases were all separated by brief instruction screens, including between early and late extinction learning, and this detail may have impacted on the patterns observed in our results. Third, our sample included a small proportion of PTSD participants, though this was done to replicate our previous study [47]. While we do not anticipate that this would affect our primary outcome, some variability in the bootstrapped samples may have been due to participant characteristics such as this. Next, we only simulated one type of potential group-level effect in our data and this may have resulted in some strategies showing greater or lesser robustness, depending on the aim of the strategy. Therefore, we cannot be prescriptive concerning which strategy may perform best with group-level effects; however, it is relevant to note that a model that best describes extinction has not been formalised and thus it is unknown what group-level extinction data should look like. Finally, there may be many more analytical strategies in the literature that were not included in the present paper. These strategies could alter the robustness between strategies reported here. The strategies reported here were identical to those identified in the previous paper—based on highly cited examples; hence, it is possible that there are different analytical strategies reported in less cited studies.

In conclusion, we found that larger sample size does not improve the robustness of fear extinction results when assessed across heterogeneous analytical strategies when no effect is simulated but does alter robustness under some circumstances when an effect is simulated. We also report that smaller sample sizes (less than $N = 120$, or $n = 40$ per group) result in inflated effect sizes, both in simulated and original data. Although this issue is not unique to fear extinction, formal identification of it may encourage better powered studies and more progressive methods in the future. Future studies should examine how robustness of fear extinction analyses can be improved and ensure that studies are adequately powered such that effect sizes are not artificially inflated.

Supporting information

S1 Table. Conditioning—Extinction, N = 30. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S2 Table. Conditioning—Extinction, N = 60. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S3 Table. Conditioning—Extinction, N = 120. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S4 Table. Conditioning—Extinction, N = 240. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S5 Table. Conditioning—Extinction, N = 480. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S6 Table. Conditioning—Extinction, N = 720. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S7 Table. Conditioning—Extinction, N = 960. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S8 Table. Static Extinction, N = 30. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with a static extinction learning efficacy estimated.
(DOCX)

S9 Table. Static Extinction, N = 60. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with a static extinction learning efficacy estimated.
(DOCX)

S10 Table. Static Extinction, N = 120. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with a static extinction learning efficacy estimated.
(DOCX)

S11 Table. Static Extinction, N = 240. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with a static extinction learning efficacy estimated.
(DOCX)

S12 Table. Static Extinction, N = 480. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with a static extinction learning efficacy estimated.
(DOCX)

S13 Table. Static Extinction, N = 720. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with a static extinction learning efficacy estimated.
(DOCX)

S14 Table. Static Extinction, N = 960. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with a static extinction learning efficacy estimated.
(DOCX)

S15 Table. Early—Late Extinction, N = 30. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes during extinction learning estimated.
(DOCX)

S16 Table. Early—Late Extinction, N = 60. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes during extinction learning estimated.
(DOCX)

S17 Table. Early—Late Extinction, N = 120. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes during extinction learning estimated.
(DOCX)

S18 Table. Early—Late Extinction, N = 240. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes during extinction learning estimated.
(DOCX)

S19 Table. Early—Late Extinction, N = 480. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes during extinction learning estimated.
(DOCX)

S20 Table. Early—Late Extinction, N = 720. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes during extinction learning estimated.
(DOCX)

S21 Table. Early—Late Extinction, N = 960. Strategy comparisons using Kendall rank correlation coefficient between effect-simulated datasets with changes during extinction learning estimated.
(DOCX)

S22 Table. Conditioning—Extinction, N = 30. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S23 Table. Conditioning—Extinction, N = 60. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S24 Table. Conditioning—Extinction, N = 120. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S25 Table. Conditioning—Extinction, N = 240. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S26 Table. Conditioning—Extinction, N = 480. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S27 Table. Conditioning—Extinction, N = 720. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S28 Table. Conditioning—Extinction, N = 960. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes from Conditioning to extinction learning phases estimated.
(DOCX)

S29 Table. Static Extinction, N = 30. Strategy comparisons using Kendall rank correlation coefficient between datasets with a static extinction learning efficacy estimated.
(DOCX)

S30 Table. Static Extinction, N = 60. Strategy comparisons using Kendall rank correlation coefficient between datasets with a static extinction learning efficacy estimated.
(DOCX)

S31 Table. Static Extinction, N = 120. Strategy comparisons using Kendall rank correlation coefficient between datasets with a static extinction learning efficacy estimated.
(DOCX)

S32 Table. Static Extinction, N = 240. Strategy comparisons using Kendall rank correlation coefficient between datasets with a static extinction learning efficacy estimated.
(DOCX)

S33 Table. Static Extinction, N = 480. Strategy comparisons using Kendall rank correlation coefficient between datasets with a static extinction learning efficacy estimated.
(DOCX)

S34 Table. Static Extinction, N = 720. Strategy comparisons using Kendall rank correlation coefficient between datasets with a static extinction learning efficacy estimated.
(DOCX)

S35 Table. Static Extinction, N = 960. Strategy comparisons using Kendall rank correlation coefficient between datasets with a static extinction learning efficacy estimated.
(DOCX)

S36 Table. Early—Late Extinction, N = 30. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes during extinction learning estimated.
(DOCX)

S37 Table. Early—Late Extinction, N = 60. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes during extinction learning estimated.
(DOCX)

S38 Table. Early—Late Extinction, N = 120. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes during extinction learning estimated.
(DOCX)

S39 Table. Early—Late Extinction, N = 240. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes during extinction learning estimated.
(DOCX)

S40 Table. Early—Late Extinction, N = 480. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes during extinction learning estimated.
(DOCX)

S41 Table. Early—Late Extinction, N = 720. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes during extinction learning estimated.
(DOCX)

S42 Table. Early—Late Extinction, N = 960. Strategy comparisons using Kendall rank correlation coefficient between datasets with changes during extinction learning estimated.
(DOCX)

S1 Fig. Overall responding in the real data set.
(TIF)

Author Contributions

Conceptualization: Luke J. Ney, Patrick A. F. Laing, Trevor Steward, Ben Harrison, Kim L. Felmingham.

Data curation: Luke J. Ney.

Formal analysis: Luke J. Ney.

Methodology: Luke J. Ney.

Supervision: Kim L. Felmingham.

Writing – original draft: Luke J. Ney, Patrick A. F. Laing, Trevor Steward, Daniel V. Zuj, Simon Dymond, Ben Harrison, Bronwyn Graham, Kim L. Felmingham.

Writing – review & editing: Luke J. Ney, Patrick A. F. Laing, Trevor Steward, Daniel V. Zuj, Simon Dymond, Ben Harrison, Bronwyn Graham, Kim L. Felmingham.

References

1. Craske M.G., et al., Treatment for anxiety disorders: Efficacy to effectiveness to implementation. Behaviour research and therapy, 2009. 47(11): p. 931–937. https://doi.org/10.1016/j.brat.2009.07.012 PMID: 19632667
2. Lebois L.A.M., et al., Augmentation of extinction and inhibitory learning in anxiety and trauma-related disorders. Annual review of clinical psychology, 2019. 15: p. 257–284. https://doi.org/10.1146/annurev-clinpsy-050718-095634 PMID: 30698994
3. Bouton M.E., Context and behavioral processes in extinction. Learn Mem, 2004. 11(5): p. 485–94. https://doi.org/10.1101/lm.78804 PMID: 15466298
4. Bouton M.E., Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. Biol Psychiatry, 2002. 52(10): p. 976–86. https://doi.org/10.1016/s0006-3223(02)01546-9 PMID: 12437938
5. Kalisch R., et al., Context-dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. The Journal of neuroscience: the official journal of the Society for Neuroscience, 2006. 26(37): p. 9503–9511. https://doi.org/10.1523/JNEUROSCI.2021-06.2006 PMID: 16971534
6. Zuj D.V., et al., The centrality of fear extinction in linking risk factors to PTSD: A narrative review. Neurosci Biobehav Rev, 2016. 69: p. 15–35. https://doi.org/10.1016/j.neubiorev.2016.07.014 PMID: 27461912
7. Duits P., et al., Updated meta-analysis of classical fear conditioning in the anxiety disorders. Depress Anxiety, 2015. 32(4): p. 239–53. https://doi.org/10.1002/da.22353 PMID: 25703487
8. Lonsdorf T.B., et al., Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. Neurosci Biobehav Rev, 2017. 77: p. 247–285. https://doi.org/10.1016/j.neubiorev.2017.02.026 PMID: 28263758
9. Ney L.J., et al., Dopamine, endocannabinoids and their interaction in fear extinction and negative affect in PTSD. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 2021. 105: p. 110118. https://doi.org/10.1016/j.pnpbp.2020.110118 PMID: 32991952
10. Abraham A.D., Neve K.A., and Lattal K.M., Dopamine and extinction: a convergence of theory with fear and reward circuitry. Neurobiol Learn Mem, 2014. 108: p. 65–77. https://doi.org/10.1016/j.nlm.2013.11.007 PMID: 24269353
11. Hill M.N., et al., Integrating Endocannabinoid Signaling and Cannabinoids into the Biology and Treatment of Posttraumatic Stress Disorder. Neuropsychopharmacology, 2018. 43(1): p. 80–102. https://doi.org/10.1038/npp.2017.162 PMID: 28745306

12. Ney L.J., et al., Modulation of the endocannabinoid system by sex hormones: Implications for Posttraumatic Stress Disorder. Neurosci Biobehav Rev, 2018. 94: p. 302–320. https://doi.org/10.1016/j.neubiorev.2018.07.006 PMID: 30017748

13. Lebron-Milad K., Graham B.M., and Milad M.R., Low Estradiol Levels: A Vulnerability Factor for the Development of Posttraumatic Stress Disorder Biological Psychiatry, 2012. 72: p. 6–7. https://doi.org/10.1016/j.biopsych.2012.04.029 PMID: 22682395

14. Gogos A., et al., Sex differences in schizophrenia, bipolar disorder and PTSD: Are gonadal hormones the link? British Journal of Pharmacology, 2019. 176(21): p. 4119–4135. https://doi.org/10.1111/bph.14584 PMID: 30658014

15. Merz C.J., et al., Neural Underpinnings of Cortisol Effects on Fear Extinction. Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology, 2018. 43(2): p. 384–392. https://doi.org/10.1038/npp.2017.227 PMID: 28948980

16. Stockhorst U. and Antov M.I., Modulation of Fear Extinction by Stress, Stress Hormones and Estradiol: A Review. Front Behav Neurosci, 2015. 9: p. 359. https://doi.org/10.3389/fnbeh.2015.00359 PMID: 26858616

17. Zuj D.V., et al., Endogenous cortisol reactivity moderates the relationship between fear inhibition to safety signals and posttraumatic stress disorder symptoms. Psychoneuroendocrinology, 2017. 78: p. 14–21. https://doi.org/10.1016/j.psyneuen.2017.01.012 PMID: 28135580

18. Ney L.J., et al., BDNF genotype Val66Met interacts with acute plasma BDNF levels to predict fear extinction and recall. Behaviour Research and Therapy, 2021. 145: p. 103942. https://doi.org/10.1016/j.brat.2021.103942 PMID: 34340176

19. Ney L.J., et al., Translation of animal endocannabinoid models of PTSD mechanisms to humans: Where to next? Neuroscience & Biobehavioral Reviews, 2022. 132: p. 76–91.

20. Graham B.M., Callaghan B.L., and Richardson R., Bridging the gap: Lessons we have learnt from the merging of psychology and psychiatry for the optimisation of treatments for emotional disorders. Behav Res Ther, 2014. 62: p. 3–16. https://doi.org/10.1016/j.brat.2014.07.012 PMID: 25115195

21. Zuj D.V. and Norrholm S.D., The clinical applications and practical relevance of human conditioning paradigms for posttraumatic stress disorder. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 2019. 88: p. 339–351. https://doi.org/10.1016/j.pnpbp.2018.08.014 PMID: 30134147

22. Milad M.R. and Quirk G.J., Fear extinction as a model for translational neuroscience: ten years of progress. Annu Rev Psychol, 2012. 63: p. 129–51. https://doi.org/10.1146/annurev.psych.121208.131631 PMID: 22129456

23. Lange I., et al., Neural responses during extinction learning predict exposure therapy outcome in phobia: results from a randomized-controlled trial. Neuropsychopharmacology, 2020. 45(3): p. 534–541. https://doi.org/10.1038/s41386-019-0467-8 PMID: 31352467

24. Fullana M.A., et al., Human fear conditioning: From neuroscience to the clinic. Behaviour Research and Therapy, 2020. 124: p. 103528. https://doi.org/10.1016/j.brat.2019.103528 PMID: 31835072

25. Picó-Pérez M., et al., Common and distinct neural correlates of fear extinction and cognitive reappraisal: A meta-analysis of fMRI studies. Neuroscience & Biobehavioral Reviews, 2019. 104: p. 102–115. https://doi.org/10.1016/j.neubiorev.2019.06.029 PMID: 31278951

26. Scheveneels S., et al., The validity of laboratory-based treatment research: Bridging the gap between fear extinction and exposure treatment. Behav Res Ther, 2016. 86: p. 87–94. https://doi.org/10.1016/j.brat.2016.08.015 PMID: 27590839

27. Vervliet B., Craske M.G., and Hermans D., Fear extinction and relapse: state of the art. Annu Rev Clin Psychol, 2013. 9: p. 215–48. https://doi.org/10.1146/annurev-clinpsy-050212-185542 PMID: 23537484

28. Simmons J.P., Nelson L.D., and Simonsohn U., False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci, 2011. 22(11): p. 1359–66. https://doi.org/10.1177/0956797611417632 PMID: 22006061

29. Wagenmakers E.J., et al., Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). J Pers Soc Psychol, 2011. 100(3): p. 426–32. https://doi.org/10.1037/a0022790 PMID: 21280965

30. Koul A., Becchio C., and Cavallo A., Cross-Validation Approaches for Replicability in Psychology. Frontiers in Psychology, 2018. 9(1117).

31. Wingen T., Berkessel J.B., and Englich B., No Replication, No Trust? How Low Replicability Influences Trust in Psychology. Social Psychological and Personality Science, 2019. 11(4): p. 454–463.

32. Rabeyron T., Why Most Research Findings About Psi Are False: The Replicability Crisis, the Psi Paradox and the Myth of Sisyphus. Frontiers in Psychology, 2020. 11(2468). https://doi.org/10.3389/fpsyg.2020.562992 PMID: 33041926

33. Beckers T., et al., What's wrong with fear conditioning? Biol Psychol, 2013. 92(1): p. 90–6. https://doi.org/10.1016/j.biopsycho.2011.12.015 PMID: 22223096

34. Pöhlchen D., et al., No robust differences in fear conditioning between patients with fear-related disorders and healthy controls. Behaviour Research and Therapy, 2020. 129: p. 103610. https://doi.org/10.1016/j.brat.2020.103610 PMID: 32302820

35. Abend R., et al., Anticipatory Threat Responding: Associations With Anxiety, Development, and Brain Structure. Biological psychiatry, 2020. 87(10): p. 916–925. https://doi.org/10.1016/j.biopsych.2019.11.006 PMID: 31955915

36. Vervliet B. and Boddez Y., Memories of 100 years of human fear conditioning research and expectations for its future. Behav Res Ther, 2020. 135: p. 103732. https://doi.org/10.1016/j.brat.2020.103732 PMID: 33007544

37. Ryan K.M., et al., The need for standards in the design of differential fear conditioning and extinction experiments in youth: A systematic review and recommendations for research on anxiety. Behaviour Research and Therapy, 2019. 112: p. 42–62. https://doi.org/10.1016/j.brat.2018.11.009 PMID: 30502721

38. Melinscak F. and Bach D.R., Computational optimization of associative learning experiments. PLOS Computational Biology, 2020. 16(1): p. e1007593. https://doi.org/10.1371/journal.pcbi.1007593 PMID: 31905214

39. Bach D. and Melinscak F., Psychophysiological modelling and the measurement of fear conditioning. Behaviour Research and Therapy, 2020. 127: p. 103576. https://doi.org/10.1016/j.brat.2020.103576 PMID: 32087391

40. Benedek M. and Kaernbach C., Decomposition of skin conductance data by means of nonnegative deconvolution. Psychophysiology, 2010. 47(4): p. 647–58. https://doi.org/10.1111/j.1469-8986.2009.00972.x PMID: 20230512

41. Green S.R., et al., Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. Int J Psychophysiol, 2014. 91(3): p. 186–93. https://doi.org/10.1016/j.ijpsycho.2013.10.015 PMID: 24184342

42. Jentsch V.L., Wolf O.T., and Merz C.J., Temporal dynamics of conditioned skin conductance and pupillary responses during fear acquisition and extinction. International Journal of Psychophysiology, 2020. 147: p. 93–99. https://doi.org/10.1016/j.ijpsycho.2019.11.006 PMID: 31760105

43. Ney L.J., et al., Critical evaluation of current data analysis strategies for psychophysiological measures of fear conditioning and extinction in humans. International Journal of Psychophysiology, 2018. 134: p. 95–107. https://doi.org/10.1016/j.ijpsycho.2018.10.010 PMID: 30393110

44. Krypotos A.M. and Engelhard I.M., Testing a novelty-based extinction procedure for the reduction of conditioned avoidance. J Behav Ther Exp Psychiatry, 2018. 60: p. 22–28. https://doi.org/10.1016/j.jbtep.2018.02.006 PMID: 29486371

45. Tzovara A., Korn C.W., and Bach D., Human Pavlovian fear conditioning conforms to probabilistic learning. PLOS Computational Biology, 2018. 14(8): p. e1006243. https://doi.org/10.1371/journal.pcbi.1006243 PMID: 30169519

46. Lonsdorf T.B., Merz C.J., and Fullana M.A., Fear extinction retention: Is it what we think it is? Biological Psychiatry, 2019. 85(12): p. 1074–1082. https://doi.org/10.1016/j.biopsych.2019.02.011 PMID: 31005240

47. Ney L.J., et al., Inconsistent analytic strategies reduce robustness in fear extinction via skin conductance response. Psychophysiology, 2020. 57(11): p. e13650. https://doi.org/10.1111/psyp.13650 PMID: 32748977

48. Bach D.R., et al., Dynamic causal modeling of spontaneous fluctuations in skin conductance. Psychophysiology, 2011. 48(2): p. 252–7. https://doi.org/10.1111/j.1469-8986.2010.01052.x PMID: 20557485

49. Ioannidis J.P., Why most published research findings are false. PLoS Med, 2005. 2(8): p. e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722

50. Galatzer-Levy I.R., et al., Utilization of machine learning for prediction of post-traumatic stress: a re-examination of cortisol in the prediction and pathways to non-remitting PTSD. Transl Psychiatry, 2017. 7(3): p. e0. https://doi.org/10.1038/tp.2017.38 PMID: 28323285

51. Weathers F., et al., The PTSD Checklist (PCL): Reliability, Validity, and Diagnostic Utility, in Annual Convention of the International Society for Traumatic Stress Studies. 1993: San Antonio, TX.

52. Weathers, F., et al. The PTSD Checklist for DSM-5 (PCL-5)—Standard [Measurement instrument]. 2013.

53. Sjouwerman, R., et al., A data multiverse analysis investigating non-model based SCR quantification approaches. 2021.

54. Johnson R.W., An Introduction to the Bootstrap. Teaching Statistics, 2001. 23(2).

55. Mundfrom D.J., Shaw D.G., and Ke T.L., Minimum Sample Size Recommendations for Conducting Factor Analyses. International Journal of Testing, 2005. 5(2): p. 159–168.

56. Collins G.S., Ogundimu E.O., and Altman D.G., Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in Medicine, 2016. 35(2): p. 214–226. https://doi.org/10.1002/sim.6787 PMID: 26553135

57. Kocovsky P.M., Adams J.V., and Bronte C.R., The Effect of Sample Size on the Stability of Principal Components Analysis of Truss-Based Fish Morphometrics. Transactions of the American Fisheries Society, 2009. 138(3): p. 487–496.

58. Schönbrodt F.D. and Perugini M., At what sample size do correlations stabilize? Journal of Research in Personality, 2013. 47(5): p. 609–612.

59. Ney L.J., et al., Cannabinoid polymorphisms interact with plasma endocannabinoid levels to predict fear extinction learning. Depress Anxiety, 2021. https://doi.org/10.1002/da.23170 PMID: 34151472

60. Graham B.M. and Milad M.R., Blockade of estrogen by hormonal contraceptives impairs fear extinction in female rats and women. Biol Psychiatry, 2013. 73(4): p. 371–8. https://doi.org/10.1016/j.biopsych.2012.09.018 PMID: 23158459

61. Milad M.R., et al., The influence of gonadal hormones on conditioned fear extinction in healthy humans. Neuroscience, 2010. 168(3): p. 652–8. https://doi.org/10.1016/j.neuroscience.2010.04.030 PMID: 20412837

62. White E.C. and Graham B.M., Estradiol levels in women predict skin conductance response but not valence and expectancy ratings in conditioned fear extinction. Neurobiol Learn Mem, 2016. 134 Pt B: p. 339–48.

63. Grady A.K., et al., Effect of continuous and partial reinforcement on the acquisition and extinction of human conditioned fear. Behav Neurosci, 2016. 130(1): p. 36–43. https://doi.org/10.1037/bne0000121 PMID: 26692449

64. Milad M.R., et al., Neurobiological basis of failure to recall extinction memory in posttraumatic stress disorder. Biol Psychiatry, 2009. 66(12): p. 1075–82. https://doi.org/10.1016/j.biopsych.2009.06.026 PMID: 19748076

65. Zuj D.V., et al., Impaired fear extinction associated with PTSD increases with hours-since-waking. Depress Anxiety, 2016. 33(3): p. 203–10. https://doi.org/10.1002/da.22463 PMID: 26744059

66. Garfinkel S.N., et al., Impaired contextual modulation of memories in PTSD: an fMRI and psychophysiological study of extinction retention and fear renewal. The Journal of neuroscience: the official journal of the Society for Neuroscience, 2014. 34(40): p. 13435–13443. https://doi.org/10.1523/JNEUROSCI.4287-13.2014 PMID: 25274821

67. Schiller D., et al., Preventing the return of fear in humans using reconsolidation update mechanisms. Nature, 2010. 463(7277): p. 49–53. https://doi.org/10.1038/nature08637 PMID: 20010606

68. Milad M.R., et al., Presence and acquired origin of reduced recall for fear extinction in PTSD: results of a twin study. Journal of psychiatric research, 2008. 42(7): p. 515–520. https://doi.org/10.1016/j.jpsychires.2008.01.017 PMID: 18313695

69. Milad M.R., et al., Fear conditioning and extinction: influence of sex and menstrual cycle in healthy humans. Behav Neurosci, 2006. 120(6): p. 1196–203. https://doi.org/10.1037/0735-7044.120.5.1196 PMID: 17201462

70. Blechert J., et al., Fear conditioning in posttraumatic stress disorder: evidence for delayed extinction of autonomic, experiential, and behavioural responses. Behav Res Ther, 2007. 45(9): p. 2019–33. https://doi.org/10.1016/j.brat.2007.02.012 PMID: 17442266

71. Michael T., et al., Fear conditioning in panic disorder: Enhanced resistance to extinction. J Abnorm Psychol, 2007. 116(3): p. 612–7. https://doi.org/10.1037/0021-843X.116.3.612 PMID: 17696717

72. Phelps E.A., et al., Extinction learning in humans: role of the amygdala and vmPFC. Neuron, 2004. 43(6): p. 897–905. https://doi.org/10.1016/j.neuron.2004.08.042 PMID: 15363399

73. Milad M.R., et al., Deficits in Conditioned Fear Extinction in Obsessive-Compulsive Disorder and Neurobiological Changes in the Fear Circuit. JAMA Psychiatry, 2013. 70(6): p. 608–618. https://doi.org/10.1001/jamapsychiatry.2013.914 PMID: 23740049

74. Soliman F., et al., A genetic variant BDNF polymorphism alters extinction learning in both mouse and human. Science (New York, N.Y.), 2010. 327(5967): p. 863–866. https://doi.org/10.1126/science.1181886 PMID: 20075215

75. Zeidan M.A., et al., Estradiol modulates medial prefrontal cortex and amygdala activity during fear extinction in women and female rats. Biol Psychiatry, 2011. 70(10): p. 920–7. https://doi.org/10.1016/j.biopsych.2011.05.016 PMID: 21762880

76. Carp J., The secret lives of experiments: Methods reporting in the fMRI literature. NeuroImage, 2012. 63(1): p. 289–300. https://doi.org/10.1016/j.neuroimage.2012.07.004 PMID: 22796459

77. Botvinik-Nezer R., et al., Variability in the analysis of a single neuroimaging dataset by many teams. Nature, 2020. 582(7810): p. 84–88. https://doi.org/10.1038/s41586-020-2314-9 PMID: 32483374

78. Sjouwerman R. and Lonsdorf T.B., Experimental boundary conditions of reinstatement-induced return of fear in humans: Is reinstatement in humans what we think it is? Psychophysiology, 2020. 57(5): p. e13549. https://doi.org/10.1111/psyp.13549 PMID: 32072648

79. Krypotos A.M., et al., A Primer on Bayesian Analysis for Experimental Psychopathologists. J Exp Psychopathol, 2017. 8(2): p. 140–157. https://doi.org/10.5127/jep.057316 PMID: 28748068

80. Krypotos A.M., Klugkist I., and Engelhard I.M., Bayesian hypothesis testing for human threat conditioning research: an introduction and the condir R package. Eur J Psychotraumatol, 2017. 8(sup1): p. 1314782. https://doi.org/10.1080/20008198.2017.1314782 PMID: 29038683

81. Cameron G., Schlund M.W., and Dymond S., Generalization of socially transmitted and instructed avoidance. Frontiers in Behavioral Neuroscience, 2015. 9(159). https://doi.org/10.3389/fnbeh.2015.00159 PMID: 26150773

82. Hoenig J.M. and Heisey D.M., The Abuse of Power. The American Statistician, 2001. 55(1): p. 19–24.

83. Cohen J., Statistical Power Analysis. Current Directions in Psychological Science, 1992. 1(3): p. 98–101.

84. Button K.S., et al., Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 2013. 14(5): p. 365–376. https://doi.org/10.1038/nrn3475 PMID: 23571845

85. Kühberger A., Fritz A., and Scherndl T., Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. PLOS ONE, 2014. 9(9): p. e105825. https://doi.org/10.1371/journal.pone.0105825 PMID: 25192357

86. Hackshaw A., Small studies: strengths and limitations. European Respiratory Journal, 2008. 32(5): p. 1141. https://doi.org/10.1183/09031936.00136408 PMID: 18978131

87. Lakens, D., Sample Size Justification. 2021.

88. Steegen S., et al., Increasing transparency through a multiverse analysis. Perspect Psychol Sci, 2016. 11(5): p. 702–712. https://doi.org/10.1177/1745691616658637 PMID: 27694465

89. Lonsdorf, T.B., et al., Multiverse analyses in fear conditioning research. 2021.