Paper:
Ney, L., Wade, M., Reynolds, A., Zuj, D., Dymond, S., Matthews, A. & Felmingham, K. (2018). Critical evaluation of current data analysis strategies for psychophysiological measures of fear conditioning and extinction in humans. *International Journal of Psychophysiology, 134*, 95-107.
http://dx.doi.org/10.1016/j.ijpsycho.2018.10.010

Critical evaluation of current data analysis strategies for psychophysiological measures of fear conditioning and extinction in humans

Ney, L. J.[1*], Wade, M.[1], Reynolds, A.[1], Zuj, D. V.[2], Dymond, S.[2,3], Matthews, A.[1] & Felmingham, K. L.[4]

[1]School of Psychology, University of Tasmania

[2]Department of Psychology, Swansea University

[3]Department of Psychology, Reykjavik University

[4]School of Psychological Sciences, University of Melbourne

∗Corresponding author at: School of Medicine (Psychology), University of Tasmania, Private Bag 30, Sandy Bay, TAS 7005, Australia

Email address: luke.ney@utas.edu.au (Luke Ney)

Abstract

Fear conditioning and extinction is a construct integral to understanding trauma-, stress- and anxiety-related disorders. In the laboratory, associative learning paradigms that pair aversive with neutral stimuli are used as analogues to real-life fear learning. These studies use physiological indices, such as skin conductance, to sensitively measure rates and intensity of learning and extinction. In this review, we discuss some of the potential limitations in interpreting and analysing physiological data during the acquisition or extinction of conditioned fear. We argue that the utmost attention should be paid to the development of modelling approaches of physiological data in associative learning paradigms, by illustrating the lack of replicability and interpretability of results in current methods. We also show that statistical significance may be easily achieved in this paradigm without more stringent data and data analysis reporting requirements, leaving this particular field vulnerable to misleading conclusions. This review is written so that issues and potential solutions are accessible to researchers without mathematical training. We conclude the review with some suggestions that all laboratories should be able to implement, including visualising the full data set in publications and adopting modelling, or at least regression-based, approaches.

Keywords: Fear learning, extinction, methods, data analysis, return of fear, replicability crisis, conditioning, Skin conductance responses (SCR), physiological responding

1.1 General Introduction

Mechanisms underlying post-traumatic stress disorder (PTSD) and anxiety disorders are commonly tested in a laboratory paradigm based on Pavlovian fear learning and use psychophysiological responding as the primary outcome variable. In light of the replicability crisis (Button et al., 2013; Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011), this field of research has recently received significant attention for the renewal (Krypotos, Blanken, Arnaudova, Matzke, & Beckers, 2017; Krypotos, Klugkist, & Engelhard, 2017) and synchronisation (Lonsdorf et al., 2017; Lonsdorf & Merz, 2017) of methodological parameters underlying potential issues with the reliability of findings. In this review, we first describe the paradigm and replicability crisis, before demonstrating the many ways in which the currently accepted analysis strategies of psychophysiological data may leave research on fear conditioning vulnerable to systemic error. Next, we advise several strategies for optimising the interpretability of results, as well as how the field may be advanced to avoid many of the errors commonly attributed to psychological research. We argue that a concerted effort is required by researchers to devise and continuously improve data strategies that intuitively reflect and describe fear and extinction learning, rather than relying on null-hypothesis testing.

1.2 Physiological Testing of Fear Conditioning and Extinction in Humans

In studies of Pavlovian 'fear learning', participants learn to associate a neutral stimulus (e.g. a coloured circle on a computer screen) with an aversive 'unconditioned stimulus' (e.g. an electric shock, see Figure 1) after these stimuli are repeatedly paired together (LeDoux, 2014; Lonsdorf et al., 2017). The new conditional stimuli (CS+) are contrasted to the safety signal (CS-), which is never paired with an aversive outcome.

Repeated presentation of the CS+ with no aversive outcome in healthy populations results in reduced conditional responding to the CS+. The mechanism by which this occurs is termed 'fear extinction'; however, it is important to note that subjective fear itself is not necessarily manipulated during these paradigms (see LeDoux, 2014). For instance, one influential view postulates that associative memory of the CS+ is overridden by safety learning over successive presentations of a CS+ in absence of an aversive effect (Bouton, 2004). In this approach, the associative learning process underlies both real-life and simulated trauma, with the subjective experience of fear itself a by-product of this process at a different psychoneurological level. Alternatively, the propositional learning framework asserts that fear learning and extinction is affected by associative learning in combination with higher-order cognitive processes such as reasoning and existing knowledge (De Houwer, 2009). In this review, the terms 'associative learning' and 'fear learning or extinction' will be broadly employed as indexable physiologically, though readers should be aware that there is currently inconclusive evidence as to what processes best reflect the actual phenomenon. Since extinction of fear is essential to recovery from traumatic stress as well as anxiety disorders, understanding how and why fear extinction occurs has important clinical value (Milad & Quirk, 2012; Yehuda et al., 2015; Zuj, Palmer, Lommen, & Felmingham, 2016).

Fear extinction efficacy can be estimated in the laboratory. Physiological measures, such as skin conductance or fear-potentiated startle, are recorded to index participant arousal at different stages of the paradigm, with higher arousal reflective of heightened anticipation to the aversive stimulus. Measures such as skin conductance are conventionally scored by subtracting responding before stimulus onset from responding following stimulus onset, which is thought of as necessary due to the continuous stream of physiological data (Boucsein, 2012). In this way, researchers hone in on the differential response to the CS+ and CS- in terms of relative excitability (Lonsdorf et al., 2017). Physiological responding to

CS+ compared to CS- is essential to determining whether fear conditioning has occurred (indicated by increased responding to CS+ compared to CS-), as well as whether fear extinction has occurred (reduced differential responses between the CS+ and CS-). Reviews on the efficacy and diversity of associative learning paradigms (Lonsdorf et al., 2017), as well as the comparability of such research to real-world illnesses such as anxiety and PTSD are available (Graham, Callaghan, & Richardson, 2014; Milad & Quirk, 2012; Milad, Rosenbaum, & Simon, 2014; Zuj, Palmer, Lommen, et al., 2016). Clinical populations, such as those with PTSD, exhibit impaired reduction in CS+ responses during extinction and recall of learned extinction compared to healthy controls (Milad et al., 2009; Wessa & Flor, 2007).

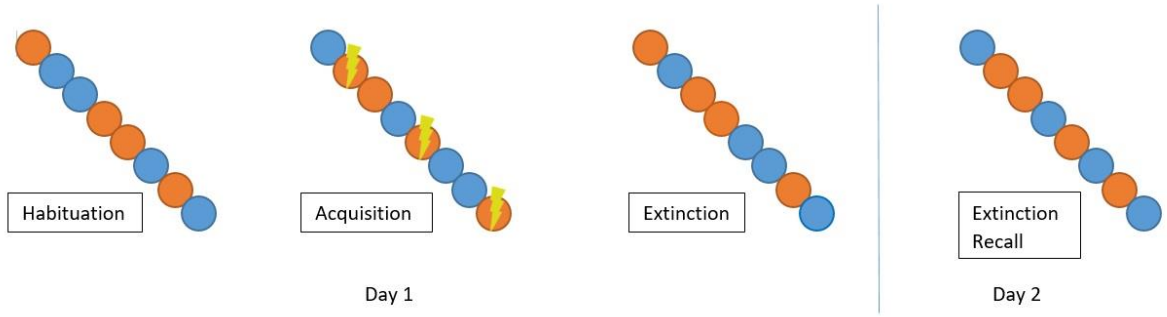

Figure 1. Simple two-day associative learning paradigm. Orange circle = CS+, blue circle = CS-, shock symbol represents unconditioned stimulus (eg. Electric shock)

Despite the translational basis of fear conditioning and extinction research, there are some existing issues with its operationalization in the laboratory (Bach et al., 2018; Beckers, Krypotos, Boddez, Effting, & Kindt, 2013; Lonsdorf et al., 2017; Lonsdorf & Merz, 2017; Sjouwerman, Niehaus, Kuhn, & Lonsdorf, 2016). Group differences between healthy and clinical populations, such as anxiety for which the primary aetiological mechanism is believed to be fear extinction, are often undetectable in individual studies and may only emerge in meta-analysis as negligible, small or small-moderate effect sizes, which have been

reported before correction for publication bias (Beckers et al., 2013; Duits et al., 2015; Lissek et al., 2005). Further, small paradigmatic differences, such as inclusion of contingency awareness instructions, difference in psychophysiological measurement tool or type of unconditioned stimulus can influence associative learning in terms of data output and presumably psychophysiological processes underlying data (Bach et al., 2018; Sjouwerman et al., 2016; Tzovara, Korn, & Bach, in press; Weidemann, Satkunarajah, & Lovibond, 2016). In a recent paper, fourteen European fear conditioning and extinction groups collectively noted an increasingly wide variation of paradigm design and methodologies, and hence produced a series of guidelines for enhancing replicability (Lonsdorf et al., 2017). The review covers variants of associative learning paradigms, physiological measurement strategies, data reduction strategy, and other methodological considerations. However, there were no recommended analytic strategies for associative learning research in this paper. Similarly, few recommendations for data analysis exist in the field. Recent efforts from a European laboratory have introduced the Bayesian approach to this paradigm (Krypotos, Blanken, et al., 2017; Krypotos & Engelhard, 2018; Krypotos, Klugkist, et al., 2017), which overcomes many of the problems associated with null hypothesis testing by using 'Bayes factors' to estimate the relative likelihood of the null and experimental hypotheses given an observed effect. This research group has developed a statistical package on R that can be easily and routinely applied to fear conditioning and extinction data (condir; see Krypotos, Klugkist, et al. (2017)).

However, on a deeper level, how accurately fear conditioning or extinction can truly be derived from the data generated when conducting this paradigm is uncertain. What is the relationship between psychophysiological responding and the process of fear extinction? In the rush to apply this paradigm to real-life problems, this issue has been almost completely overlooked, and not just in this field (Eisenberg et al., 2018). Some researchers have

developed modelling methods for skin conductance data by utilising the full data in time-series analyses sensitive to event-related responses (Bach et al., 2018; Bach, 2014; Bach, Daunizeau, Friston, & Dolan, 2010; Bach, Flandin, Friston, & Dolan, 2009, 2010; Bach, Friston, & Dolan, 2010, 2013; Staib, Castegnetti, & Bach, 2015). Whereas this approach has focussed largely on fear learning, we here primarily discuss fear extinction, as this is the process most relevant to therapy. Further, despite the availability of open-source code and program (PsPM, Bach, et al. 2009; Bach, et al. 2013), there has been little uptake of this or other modelling approaches in the field. Therefore, in the present paper we outline the critical statistical issues inherent to the conventional approach in a way that is conceptually accessible to the general psychological audience, with the aim of improving methodological outcomes in the near future.

2.1 Basis of Statistical Issues in Associative Learning Paradigms

Establishing whether there is a difference between responding over successive trials in this paradigm is difficult in practice, even when effect sizes should be large, such as between clinical and healthy populations (Beckers et al., 2013). Compared to real life trauma, effect sizes for the psychophysiological responding are rarely large (Duits et al., 2015; Lissek et al., 2005). Floor effects of responding are often observed after a few extinction trials; hence group differences must be observable between what is generally only two to three trials. This is because the aversive stimulus, compared to real life trauma, is only a mildly uncomfortable stressor that may not always produce a large difference in responding between groups, especially if there is only a small difference in responding expected between the groups. This is compounded by noisy data characterised by large individual variability, where differences between groups is likely masked by arbitrary differences between and within individuals (Bach, Flandin, et al., 2010; Beckers et al., 2013; Boucsein, 2012;

Lonsdorf & Merz, 2017). For this reason, if experimenters are interested in comparing more than two groups, it quickly becomes impractical to collect a large enough sample size to adequately power a psychophysiological associative learning experiment using even lenient multiple correction techniques. Further, there are a large number of independent variables to consider: with two stimuli (CS+ and CS-), a minimum of two experimental groups (often more), at least five trials in a phase (eg. extinction phase) and at least three phases (i.e. habituation, acquisition, extinction).

Most academics in psychology will be familiar with the false positive problem. Research projects that are inadequately powered, that have complex research designs with multiple or flexible outcomes, that have small effect sizes, and that are analysed in multiple methods are more likely to be false than true (Ioannidis, 2005). Aside from the conventional Type 1 and Type 2 errors, less known errors include 'Type M' and 'Type S' errors (Gelman & Carlin, 2014), which occur when the power is low and statistical significance is reached. More specifically, experiments that are underpowered and find a significant effect that is real necessarily overestimate the effect size by many orders of magnitude (Type M), and have a good probability of estimating the effect with the wrong sign (Type S). For instance, an effect found in a study with β=.06 has a 24% chance of having the wrong sign and is more likely than not to be at least nine times higher in magnitude than the true effect (Gelman, 2016). This is arguably worse than a non-significant finding, which at least reminds us that we need better measurement to accurately estimate the difference. Importantly, due to small effect sizes and noisy, imprecise psychophysiological measurements, studies using conventional data analysis strategies in the associative learning paradigm are typically hugely underpowered, with estimates for adequate power from ours and the University of Zurich laboratory (β=.80 and above) likely starting from N=200 for a single comparison of a small effect size (also see Bach et al. (2018); Bach, Tzovara, and Vunder (2017)). For instance, a

recent power analysis with an effect size based on existing data from a statistical paper (Khemka, Tzovara, Gerster, Quednow, & Bach, 2017) identified that 74 participants were required to detect a 50% reduction in responding during fear extinction of an experimental compared to a control group (Bach et al., 2017), which is a very large reduction. However, most studies include less than 74 subjects, anticipate small effect sizes and perform multiple comparisons.

Aside from the underwhelming power in most of these studies, a more subtle yet also critical issue with data analyses currently employed are the theoretical assumptions made by the most commonly chosen techniques. Most of this literature has analysed physiological scores between trials using unordered ANOVA, or has grouped means of trials together and compared the differences using ANOVA. It is often possible that patterns reflecting overall decreased arousal of one group compared to another will emerge by averaging trials. However, many of the patterns in the data that may better describe the hypothesized process of fear extinction (as indexed by reduction in physiological arousal) may not be well reflected by the averaged data. We will give an example of this later in this paper. Ideally, data should be modelled based on the patterns that would be expected to be seen given that an individual or a group is experiencing a certain psychological process (Farrell & Lewandowsky, 2015); however little attention has yet been given to how psychophysiological responding might acutely reflect psychological fear extinction or associative learning processes (Bach, Flandin, et al., 2010). Overall, it is concerning how little agreement there is between groups as to how these psychophysiological results should be analysed and reported. Not only does this increase false positive errors (Ioannidis, 2005), but it is not good scientific practice to have a field with many different acceptable methodologies and treat findings as equal when separate methodologies are actually testing different parameters and assumptions that are often not apparent in conventional data analysis strategies. As we will discuss later, the application of

statistical modelling to cognitive psychology (Forstmann & Wagenmakers, 2015) increases the relevance of data outcomes to the theoretical model being tested and is an ideal method of increasing power to detect important effects.

2.2 Brief Review of the Analysis Disparity between Studies

We will here illustrate what we consider to be the first major statistical issue in the analysis and reporting of psychophysiological data in associative learning paradigms: that of lack of power and consequent inconsistency of method design between studies. Before reading the following section, please note that this review is not intended to single out any particular paper or research group and it is not our opinion that any of the following analytic strategies are by themselves problematic. However, it is relevant to note the wide variation of statistical methodology across and within research groups, and this will later be discussed with reference to both the excessive flexibility in statistical decision-making as well as in making a case for more thoughtful analysis strategies in this field. We also acknowledge that similar findings across different studies using different methods and analytic approaches may suggest that the findings are robust and insensitive to minor paradigmatic changes. However, it is also a well-documented risk factor for poorer replicability (Ioannidis, 2005; Simmons et al., 2011) and we discuss later in this paper how the problems associated with it may be largely avoided by alternative approaches such as modelling.

There is a clear disparity between analysis strategies across studies in the associative learning paradigm. For example, Phelps and colleagues (2004) divided extinction phases from the first and second days of testing into early and late extinction trials, subtracted CS- from CS+ responses to yield a differential score for each participant, and compared the mean differential scores between early and late trials. This approach maximised power in a low-powered fMRI design where only eleven participants were included in the final analysis

(Button et al., 2013; Phelps, Delgado, Nearing, & LeDoux, 2004), and the authors consequently removed the first three trials of extinction recall on day two before their analysis due to suspected noise. In similar fMRI paradigms with thirty-one and fourteen participants respectively, Milad et al. (2009) and Milad et al. (2007), reported mean differences between CS+ and CS- in each phase, rather than differential scores and removed the last four trials during recall, only analysing the first four instead. We can immediately see how fMRI-fear extinction studies may be consistently underpowered. These two papers also reported a retention extinction index, which is a score that accounts for individual variability in physiological responding to the unconditioned stimulus during acquisition (Lonsdorf et al., 2017). However, in the first paper, only responses to the first two CS+ trials during extinction recall were considered (Milad et al., 2007), compared to the first four CS+ trials in the second paper (Milad et al., 2009), despite there being the same number of trials in each experiment. Differences in data strategy was not based on differences between goals and experimental design, since all three studies cited above examined extinction recall as the primary outcome in two day paradigms and it was admitted in these papers that post hoc removal of trials was used to account for floor effects on physiological responding. It is important to note that post hoc data inclusion/exclusion rules, as well as changes to analysis plan, form part of what is known as the 'garden of forking paths' or 'researcher degrees of freedom' and can both increase false positive error rates as well as greatly exaggerate effect sizes (Button et al., 2013; Gelman & Loken, 2013; Simmons et al., 2011). Undoubtedly this occurs in the majority of studies unintentionally and is, in our view, driven by the restraints of a paradigm that utilises an inefficient and underpowered design, which we will discuss later in the review. Therefore, we wish to again emphasise that these decisions are not the fault of any particular researcher but are generally the product of low power to detect an effect that in many cases will be a true effect. In cases where changes are made post-hoc to the data set –

no matter how justified they may be – findings using the full data set should be reported alongside the primary findings (Simmons et al., 2011).

Variation in analysis strategy becomes a more major issue when a field normalises excessive flexibility in trial reduction and analysis strategy, which results in higher 'researcher degrees of freedom' (Simmons et al., 2011; Wagenmakers et al., 2011). Changes to both data and methodology is a subtle, and usually unintentional (Gelman & Loken, 2013) method that increases the likelihood that a significant result will be found. Excessively flexible fields are problematic since variability in possible designs and data reduction post-hoc is analogous to multiple comparisons and is not reflective of an alpha level of α=.05 (Gelman & Loken, 2013; Simmons et al., 2011). As examples of the flexibility in this field, Phelps et al. (2004), Schiller et al. (2010) and Soliman et al. (2010) divided extinction trials into 'early' and 'late' phases, whereas Lonsdorf et al. (2015) and Raes and De Raedt (2012) did not compartmentalise any phases. Conversely, Pappens et al. (2014), Spoormaker et al. (2012) and Norrholm et al. (2011) divided extinction trials into three, five and six blocks of trials respectively for mean block comparisons using ANOVA. Other research groups have used the more broad strategy of conducting ANOVAs across ungrouped trials, rather than grouping them into smaller segments (Pace-Schott et al., 2009; Zuj, Palmer, Hsu, et al., 2016), or have tested the mean responses across trials of each experimental phase (Klumpers et al., 2012), though these papers struggled to find significant effects. Other papers have chosen and excluded different extinction or extinction recall trials to compare using ANOVA (Graham & Milad, 2013; Kindt, Soeter, & Vervliet, 2009; Rabinak et al., 2013). Interestingly, others have calculated an extinction retention index using again different methods to those described previously by selecting to use either all recall trials, the first CS+ trial, the first differential trial, the first two differential trials, or the first four CS+ trials (Graham & Milad, 2013; Milad et al., 2005; Milad et al., 2010; Raio, Brignoni-Perez,

Goldman, & Phelps, 2014; Zeidan et al., 2011). Critically, whilst we maintain that none of these methods are potentially any less justifiable than others, the large number of reporting styles provides a vast number of potential comparisons that statistically changes the likelihood that rejection of the null hypothesis is accurate (Gelman & Loken, 2013; Ioannidis, 2005; Rosenthal, Rosnow, & Rubin, 2000; Simmons et al., 2011). In summary, we recognise excessive variability in: (a) block combination and ordering; (b) trial inclusion and exclusion; (c) trial calculation; and, (d) data analysis strategy. It should also be noted that, in a recent addendum to their 2010 paper, Schiller et al. (2018) reported that changing complex SCR exclusion criteria changed the outcome and effect sizes of the study. Therefore, and as recommended in Lonsdorf, et al. (2017), consistent and justifiable exclusion criteria should also be a target for future studies.

There is not necessarily a problem with variability in methodology between projects and research groups. However, differences between study designs should ideally be openly reported, justified and evaluated against existing designs (Bach et al., 2018; Forstmann & Wagenmakers, 2015a; Gelman & Loken, 2013). In the rest of this paper, we therefore focus largely on how a priori (or post-hoc, in some cases) data analysis strategy and post-hoc corrections to data specifically contribute to lack of replicability (such choices made likely due to lack of power; see details on power analyses in the above sections). For these reasons, it is also our opinion that the field is essentially in an "exploratory phase" and all found results need to be weighed very carefully against both established theory and independent replicability. It is also likely that researchers will often find themselves frustrated when results are not intuitive or do not replicate, particularly when it comes to large effect findings of studies with small numbers of observations. Similarly, there may be many well-conducted studies that do not find significant effects, and this may be again due to a lack of power of the overall design and analysis strategy.

2.3 An Example using Unpublished Data

To make the implications of this issue completely clear, as well as to lead into the second major statistical problem detailed in this paper, we decided to use a few of the different analysis strategies to examine some unpublished data that we had. We recently conducted a study testing the stress hormone theory proposed by Joels and Baram (2009), where "first wave" hormones (catecholamines) may enhance extinction learning but "second wave" hormones (glucocorticoids) may impair extinction learning. Briefly, participants were exposed to a standard associative learning paradigm, where they underwent a single session consisting of habituation, acquisition, early extinction and late extinction phases. During fear acquisition, a 75% reinforcement schedule was used, with a mild electric shock serving as the US, and coloured circles serving as the CS+/-. In between acquisition and extinction, there was a twenty minute break where we administered an abbreviated version of the *Maastricht Acute Stress Test* (MAST; Smeets, et al. 2012) to some participants. Participants were separated into three groups: a first wave group (who underwent the stress procedure immediately before early extinction who thereby had increased adrenal hormones during extinction learning); a second wave group, who underwent the five-minute MAST at the beginning of the break (who thereby had increased glucocorticoid hormones during extinction learning); and a control group, who did not experience stress induction. Skin conductance response (SCR) was recorded as the dependent variable, with the mean SCR 2 seconds prior to stimulus onset subtracted from the mean SCR 12 post stimulus onset, as per standard 'peak and trough' procedure (Lonsdorf et al., 2017). Due to a short time schedule, limited resources and lower than expected effects of our laboratory stressor on salivary hormone and stress levels, this study was underpowered and we were left with clear null findings (unpublished data; Figure 2). Mixed models ANOVAs showed that the trial × group

and trial × group × stimulus interactions were not significant for early or late extinction (all $p>.05$), precluding any evidence that our experimental manipulation affected the fear extinction learning.

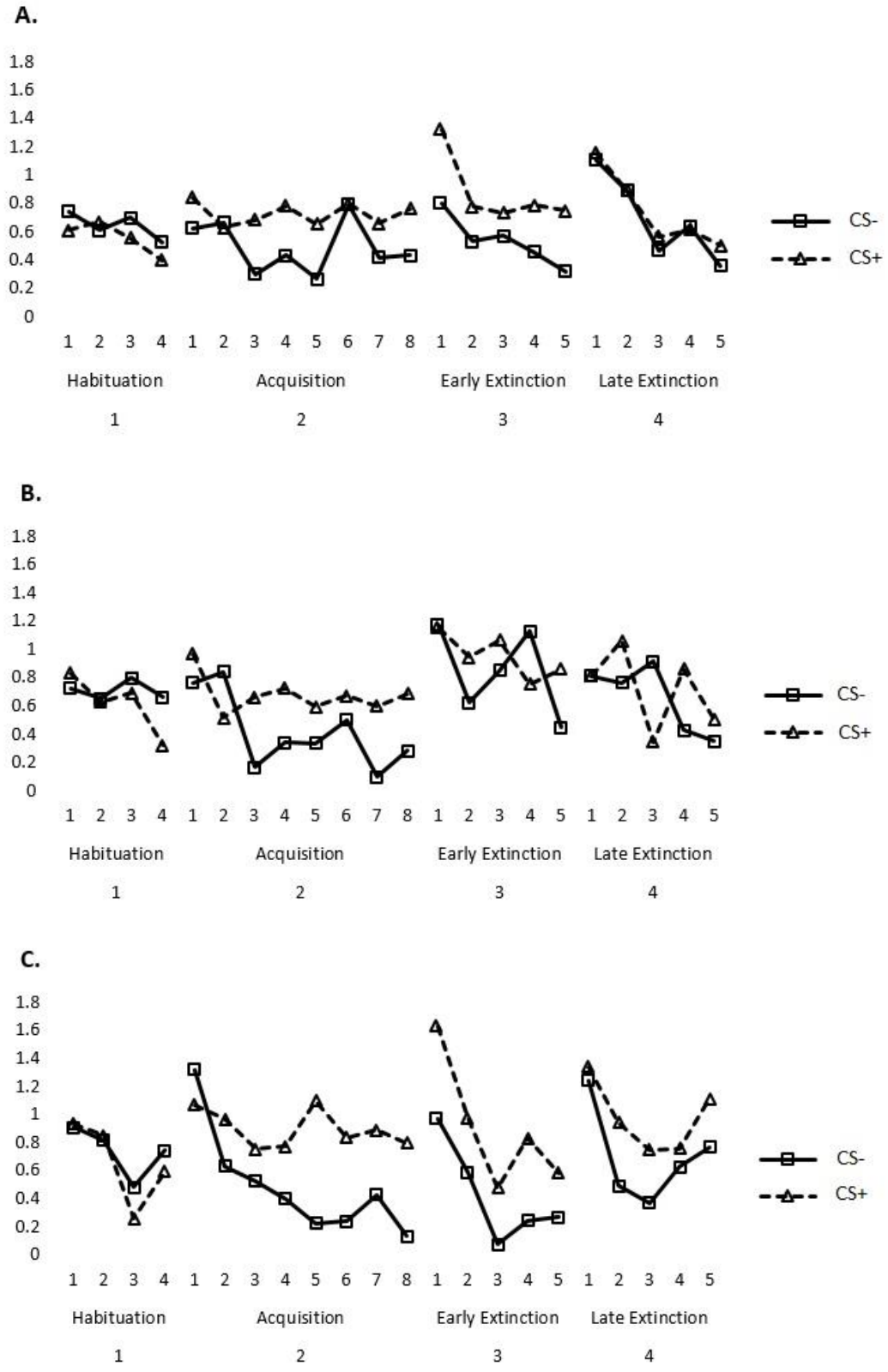

Figure 2. Skin Conductance Responses to CS+ and CS- by group and experimental phase. A panel = Control, B panel = First-wave group, C panel = Second wave group.

Our analysis strategy was to compare all trials in each phase between the stimulus conditions and between the groups. This means that, with a sample size of only 15 males per group, this analysis was certainly underpowered and we could have reasonably opted for a more simple method that would have enhanced the likelihood of finding a pattern given the small effect of our laboratory stressor. This choice may have seemed all the more reasonable when we realised post hoc that the effect of our stress manipulation was much smaller than anticipated (however this constitutes the garden of forking paths, see Gelman & Loken, 2013). Further, our new data analysis could have been conducted in any number of ways, if we drew our strategy from the existing literature.

For this example, we opted to compare our findings using several of the different techniques used in published fear extinction studies. We first examined the mean differential responses between acquisition and early extinction, and early extinction and late extinction, followed by the differential mean responses of the first, first two and first four trials before and after extinction. None of these analyses produced significant phase × group or phase × group × stimulus interactions (p>.1, data not shown). However, by grouping the CS+ and CS- responses together and analysing by phase instead of trial, we found significant differences between groups on overall response from the acquisition to early extinction phases (Figure 3; p=.024, η_p^2=.16), and from the early extinction to late extinction (Figure 3; p=.012, η_p^2=.19). Notably, these results were in opposite directions between the analyses with relatively large effect sizes, and it could be reasonably argued that, by grouping the trials in this way, we had found some tentative support for the hypothesis as late learning seemed to be impaired in the second wave group. None of the other approaches that we tried (mean CS+ and CS- responses examined during the first, first two and first four trials before and after the start of extinction, respectively) were successful in producing significant effects (p>.1, data not shown). Therefore, in our brief excursion to the practice of p-hacking, the

findings in Figure 3 may have been submitted for publication and justified, given that it is a intuitive effect given our expectations derived from our theoretical background.

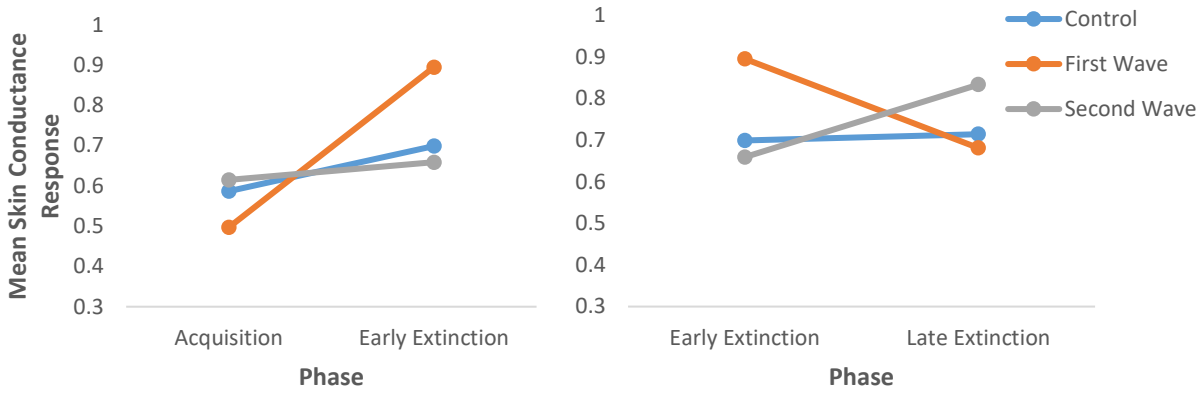

Figure 3. Mean Skin Conductance Responses during acquisition and early extinction phases, and early and late extinction phases, between experimental groups

Few researchers may test this number of potential comparisons, and instead may lean towards choosing a strategy post hoc after viewing the data (Gelman & Loken, 2013), if at all. However, it is likely that negligible effects are susceptible to optimisation when data is being analysed, as there is no doubt that the larger part of these studies suffers from lack of adequate power due to individual variability in psychophysiological responding, particularly for skin conductance paradigms.

Relatedly, the second critical issue in current methodological designs stems from optimisation of effects using inconsistent analyses. When data is analysed using any statistical technique, a model involving parameters and assumptions is applied to the data (Forstmann & Wagenmakers, 2015b). The way that data is analysed is meant to say something about the overall pattern observable in the data, such that reflects the psychological process being studied. The essence of this review is captured not so much by p-hacking but by how misleading our chosen analysis and conclusion is in context to the

patterns observable in our dataset, and subsequently how this choice of analysis limits the relevance of our conclusion to fear extinction itself. Firstly, the effect in Figure 3 is not particularly evident when looking at the skin conductance responses across trials in Figure 2. The hypothesised effect occurs only during late extinction in the final two trials of the second wave group, where they seem to display increasing responses compared to other groups, and in a heightened response among first wave participants in a single early extinction trial (Figure 2). This does not intuitively reflect what we might think the psychological process underlying fear extinction would look like over time. Rather, a real effect for fear extinction, as captured using physiological responding, might have shown a more pronounced and enduring decrease in the rate of linear extinction of the second wave compared to control groups, given that theoretically extinction learning is an inhibitory learning that occurs over time (Bouton, 2004). Our final chosen analysis was therefore able to detect an unintuitive effect that was not well described in context of the data and certainly does not provide certain distinctiveness from noise between trials. Therefore, despite finding a significant effect, our ability to interpret the effect in context to both the data and fear extinction itself is limited, due to the assumptions and parameters imposed by the analysis. Relatedly, a small sample size with the reported effect sizes is unconvincing when visualising the actual data, suggesting that a type M error in our analysis visualised in Figure 3 was likely and ultimately highly misleading (Gelman & Carlin, 2014).

Keeping in mind that accurate estimates of individual variability are the basis for most statistical tests, it is important to recognise that in this particular paradigm it is actually standard procedure to use only a small proportion of the information collected thought to be best relevant to the hypothesis (due to temporal proximity to stimulus presentation). Most commonly, researchers will subtract the mean SCR values immediately prior to the stimulus onset from the mean SCR values after the stimulus onset, to garner an index of relative

anticipation between trials. Usually these data points are only a few seconds long, and a small amount of the actual information collected, with inter-trial interval information discarded prior to analysis (Lonsdorf et al., 2017). Therefore, the larger part of information collected that may have otherwise improved error estimates is not used in analysis. Inter-trial interval information would otherwise allow us to recognise the rate and magnitude of SCR fluctuation of each individual participant, as well as to gauge the rate of change in SC level. Notably, SC level is known to slowly change over time and is not accounted for using conventional techniques (Benedek & Kaernbach, 2010a; Boucsein, 2012). This problem becomes even more pronounced when researchers are given license to remove trials that are reaching a floor effect during extinction or are otherwise unusual. We can therefore also expect that actual error estimates obtained in most analyses are poor, which in turn will have unfortunate effects when we come to running our analyses in either the case of type 1 or type 2 errors. We also did not correct for multiple comparisons – if we had done so, these effects would have been marginally, or not, significant, depending on which correction was used. Similarly, corrections for *implicit* multiple comparisons – that is, the multiple comparisons that could be conducted but are not (Rosenthal et al., 2000) – were not conducted and have not been attempted in any fear conditioning paper to our knowledge.

2.4 Differences in Psychophysiological Measures

One of the most profound current problems in this paradigm is that we are largely unaware of how psychophysiological outcomes reflect the actual processes underlying fear conditioning and extinction, despite adequate knowledge of how these outcomes are derived from psychological stimuli (Bach et al., 2018). Importantly, each psychophysiological measure is derived through different sympathetic nervous system pathways and are therefore likely to reflect different aspects of psychological responsivity (Bach et al., 2018). For

instance, whereas skin conductance responding is derived from sympathetic sudomotor nerve fibres to sweat glands (Boucsein, 2012), fear-potentiated startle is modulated by various stimuli through activation of the ventrocaudal pontine reticular formation which is linked to the amygdala and thus central nervous system (Davis, 1992; Yeomans, Li, Scott, & Frankland, 2002). Whereas skin conductance is an index primarily of arousal, fear-potentiated startle is sensitive to both valence and arousal, meaning that it is useful for a wider breadth of less invasive stimuli, such as images (Lang, 1995). Indeed, of the measures routinely employed, skin conductance and fear-potentiated startle appear to index fear extinction most sensitively (Jovanovic, Norrholm, Sakoman, Esterajher, & Kozarić-Kovačić, 2009); however, there is recent evidence for possible uses of measurements of pupil dilation (Korn, Staib, Tzovara, Castegnetti, & Bach, 2017), bradycardia (Castegnetti et al., 2016) and respiration (Castegnetti, Tzovara, Staib, Gerster, & Bach, 2017).

Despite a slow uptake it is increasingly recognised that different measures do not index fear conditioning and extinction identically; for instance, it has previously been reported that skin conductance responding is less sensitive to differences in fear acquisition than fear-potentiated startle (Glover et al., 2011). Further, some research suggests that skin conductance may reflect anticipation due to uncertainty during the paradigm, rather than solely reflecting anticipation of the aversive stimulus (Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; Tzovara et al., in press; Zhang, Mano, Ganesh, Robbins, & Seymour, 2016). Despite this, however, remarkably little is known about the relationship between psychophysiological responding and the psychological process of fear extinction itself, and studies using different measures are too often treated equally despite indexing different processes. This understanding might be greatly improved by indexing fear extinction using simultaneous measurement methods that assess different physiological aspects of psychological responding (Bach et al., 2018). Further, it is important to consider that there

will be pronounced differences in outcomes between studies using different measures (Bach et al., 2018; Glover et al., 2011), and these studies may not be reliably comparable.

3.1 Potential Solutions

Given all of these problems, what may be the best way to provide a reliably accurate description of this form of physiological data? We have several suggestions that might help, though acknowledge that without a significant amount of statistical development this field will be prone to a serious lack of replicability. Most obviously – though probably least practically – is that increasing sample size will always enhance the probability that found effects are true findings that reflect the wider populations being studied. We will discuss later in this article, however, how simply increasingly sample size is not the most efficient fix for this problem. Secondly, non-averaged responses should be reported and visualised in the publication, so that the reality of an effect, or no effect, is more apparent and readers can better understand the effect they are being told is there. In terms of data analysis strategy, we argue that, in most cases, analysing the full data set (applicable to both the untransformed as well as transformed data points) is preferable to the averaged, or reduced, trials. However, at this stage it is unlikely that more descriptive methods of full transformed data sets will be more successful in detecting group differences than analysing averaged trials using ANOVA. What we argue for here is a movement towards identifying common patterns observed across trials during fear extinction that may not reflect group differences otherwise identified by averaging trials, with the view of increasing power using modelling to detect these descriptive effects in the future. Moreover, different trials are usually compared across studies, leaving little room for collective interpretation of expected physiological patterns in conditioning and extinction. For instance, had we only had slightly more power, we may have found post-hoc that trial 3 was different from trial 4 in early extinction (Figure 2), but

this would have contradicted the overall pattern of the data. It is clear, therefore, that an unfocused ANOVA is a poor reflection of the full data set across trials. Regardless, unfocused ANOVA is the prevailing statistical option in this field currently.

Our foremost suggestion therefore is that modelling-based approaches should be adopted for psychophysiological responding in this paradigm. The best motivation for developing models is that they give us something that as a field agreeably represents the actual process of under investigation, rather than verbal and diffuse descriptions between research groups that result from different combinations of unfocused and tests with unspecified parameters and assumptions (Farrell & Lewandowsky, 2015). If researchers have competing hypotheses as to what parameters observed in the data constitute fear extinction, these can be played out with competing models that can easily be applied to lots of existing data (both sides can model each other's data). When a new model is proposed, it can be tested against old data, provided the data are made available. Constantly improving models can continuously better estimate patterns in data, and thus perform better, more intuitive and reliable comparisons that are meaningful in context of the mechanism being explored (Farrell & Lewandowsky, 2015). For instance, the increase in SCR towards the end of extinction in our second wave group (Figure 2) could have resulted from a form of gambler's fallacy, where an increasing number of unreinforced CS presentations triggers the expectancy that the chance of shock increases. A research group could adopt this as their working hypothesis, such that poor extinction learning exhibits this effect. A competing group might model a more linear, sustained pattern during extinction learning for an impaired group. The model fit for each of these working hypotheses could be compared easily across multiple datasets, with the better fitting model across multiple samples evidently most likely to be true.

In effect, this approach to modelling data also begins to reverse what statistics is conventionally used for: instead of being a tool to provide group differences in an existing theory, it can be used to feedback information to what is happening during fear extinction, associative learningor relevant process. Statistical analysis then becomes an additional and powerful tool that helps us to understand the process being studied. As a result of this, application of statistical modelling can also massively boost power to detect effects that are directly relevant to the research question (Forstmann & Wagenmakers, 2015b), however these parameters must first be identified, specified and rigorously tested (Heathcote, Brown, & Wagenmakers, 2015).

Statistical modelling has been successfully applied to other psychological fields. For instance, choice reaction time has a strong modelling history with great improvements in the ability of this field to predict decision-making (Brown & Heathcote, 2008; Logan, Cowan, & Davis, 1984; Ratcliff & McKoon, 2008; Ratcliff & Smith, 2004). Advances in this field have led to successful predictive models that can be applied to other fields, such as neuroscience and schizophrenia (Culbreth, Westbrook, Daw, Botvinick, & Barch, 2016; Forstmann & Wagenmakers, 2015b; Heathcote, Suraev, et al., 2015; Smith & Ratcliff, 2015). However, perhaps the most relevant advance in this field has been acute, strategic predictions of responding across research groups, with rigorously and meaningfully defined boundaries for what constitutes effects. Ideally, the associative learning paradigm will need to move in this direction before any true certainty about the processes that we are observing, and hence ultimate replicability, may be achieved. Indeed, work by statisticians at the Zurich University have already developed several models that utilise the full physiological data, improve error estimates and increase power to detect relevant effects (Bach, 2014; Bach, Daunizeau, et al., 2010; Bach et al., 2009; Bach, Flandin, et al., 2010; Bach, Friston, et al., 2010; Bach et al., 2013; Staib et al., 2015).

Finally, it is important to note that, despite modelling having major advantages over conventional statistical approaches, many arbitrary decisions that can affect the overall efficacy of the model still exist when models are constructed (Heathcote, Brown, et al., 2015). What sets modelling as an approach aside from conventional statistics is that decisions are based on rational criteria and are openly validated between different models (Heathcote, Brown, et al., 2015). One proposition for the fear conditioning and extinction literature is that this may be achieved using predictive validity, whereby known methods that inference fear extinction from physiological data are compared against a new method (Bach et al., 2018). Surprisingly, very few studies have compared different physiological measures or methods during fear extinction, and yet results are usually interpreted equally despite a lack of knowledge whether the process is being measured identically. Similarly, effect sizes resulting from models that specify different parameters can be compared with effect sizes from known methods to determine whether arbitrary changes between models are appropriate (Bach et al., 2018). In this regard, however, researchers will need to be sensitive to the validity and reliability of combining multiple measures in the same experiment given recent evidence that inclusion of the startle measure in addition to skin conductance delayed fear learning (Sjouwerman et al., 2016). This line of research, then, will require careful study design to be successful.

3.2 Simple Models for Psychophysiological Data that Intuitively Describe Fear Extinction

In this section, we describe several different basic models that may intuitively reflect what most research groups presume – but most likely do not directly test – is occurring at a physiological level during their associative learning paradigms. We would like to first note that we are not ourselves tied to any particular model given the lack of available empirical evidence comparing such models.

The most simple model that may intuitively describe fear conditioning and extinction data is the linear model. The linear model might be suitable for this paradigm as we are expecting change in responding to occur as a function of time, for successful fear extinction. This is contrast to general ANOVA, which is insensitive to the order of means and does not take into consideration the overall pattern of the data. Therefore, linear functions provide more power for detecting effects where the change in the data over trials is the interest of the researcher (Rosenthal et al., 2000). Linear regression, contrasts within the ANOVA function or mixed models may be used for this, and could indeed be used routinely and immediately without any additional model development. All three of the above-mentioned approaches are undoubtedly better suited to physiological responses, and are readily available in all statistical packages. Mixed models may be superior to contrasts in most study designs due to contrasts, as part of the ANOVA function, assuming measurements (in this case trials) are categorical and therefore equally spaced. This is often not the case for the associative learning paradigms, where timing between trials is often randomised. However, we might also expect that psychophysiological responding might not change in a linear fashion – conversely, we might expect that at the start of fear extinction responding might remain high and then later decrease more rapidly. Other simple models aside from linear patterns might therefore be similar to exponential decay functions, which intuitively might reflect floor effects in physiological responses at later stages of extinction and thus removing the necessity and unfortunate liberty of removing trials post hoc. Even at this simple stage, it is clear that these choices of test more intuitively describe the process (fear extinction) that we intend to study than does unfocused ANOVA.

Further, given that fear extinction is achieved when there is no response to a previously feared stimulus, it might be a useful approach to count and compare the number of trials that experimental groups take before reaching negligible responding. More efficient

fear extinction would be indexed by shorter time to negligible responding, and poorer fear extinction learning by longer. Although research would need to outline what constitutes negligible responding, one suggestion might be that three or more consecutive trials where the peak and trough SCR score is not significantly different to zero. This simple approach is theoretically intuitive and might also circumvent the need to remove floor effect trials during extinction learning. Alternatively, using untransformed time-series data, signal detection to establish negligible responding may be employed as routinely done in other fields (eg. electromyography, see Hinder, Schmidt, Garry, Carroll, & Summers, 2011 for example, and Raez, Hussain, & Mohd-Yasin, 2006 for review). Other fields using physiological measures detect signals using signal-noise ratios, such as if the signal is four times higher than background noise it is considered a true signal.

The small amount of data collected per subject, but access to large numbers of participants with a particular focus on group comparisons rather than individual comparisons also make this field an ideal candidate for hierarchical modelling, which is a Bayesian approach. Hierarchical modelling is appealing as it formally addresses the problem of group comparisons with many subjects per group, which is increasingly recognised as an important contributor to the noise in physiological responding (Lonsdorf & Merz, 2017). It also allows that some effects are going to be purely individually driven, and some are going to differ because of group, which is relevant to fields with high individual variability. If researchers are hesitant to use hierarchical modelling (a Bayesian approach), then multi-level modelling may also be useful and, similar to mixed models or simple linear models, can be conducted with a standard statistics package.

What would be most ideal would be full development of models that account for each specific parameter assumed to be important to both collection of psychophysiological responding (that is, the relationship between the psychophysiological measure used to

physiological arousal) and the hypothesised process underlying fear extinction. For instance, in Figure 4 we illustrate how psychophysiological responding during fear extinction might reduce as an exponential product of the reinforcement rate and the number of extinction CS+ trials, with individual responding rate as a factor too. This may be expressed by the equation:

$$n_{i+1} = n_i * (1\text{-}RR)^{ni+1}+b+\varepsilon$$

where i is the trial number, n is the arousal level, RR is the reinforcement rate and b is the baseline arousal. This model for fear extinction tells us that arousal will reduce as it becomes less likely that no shock was delivered due to chance, based on the original reinforcement rate. This is necessarily an exponential decrease in arousal, as it becomes exponentially less likely, at a 75% reinforcement rate, for example, that no shock would occur on multiple consecutive occasions. Obviously, this is too simplistic given the complexities of the relationship between the sympathetic nervous system and fear extinction learning, however it should illustrate how using model-based statistics may enhance our understanding of the physiological process during fear extinction learning. Further, it makes the important assumption that trials will be correlated with one another, which is not necessarily made using unfocussed ANOVA. Further to this, and due to the tendency for CS- responses to also increase and decrease during the paradigm, the baseline variable would need to account for the tonic drift in SC level and has been accordingly discussed elsewhere (Benedek & Kaernbach, 2010a). Beyond the expected tonic fluctuation in SC level that affect the SC-, we might also expect in situations where participants respond physiologically to the CS- even though it is the safety signal. This response might be modelled by indexing average deviance from the baseline SC level; if the average deviance is significantly higher than zero than this

may imply generalisation of associative learning to the CS-. However, as with the CS+, this idea serves just as an example and may be elaborated with further work.

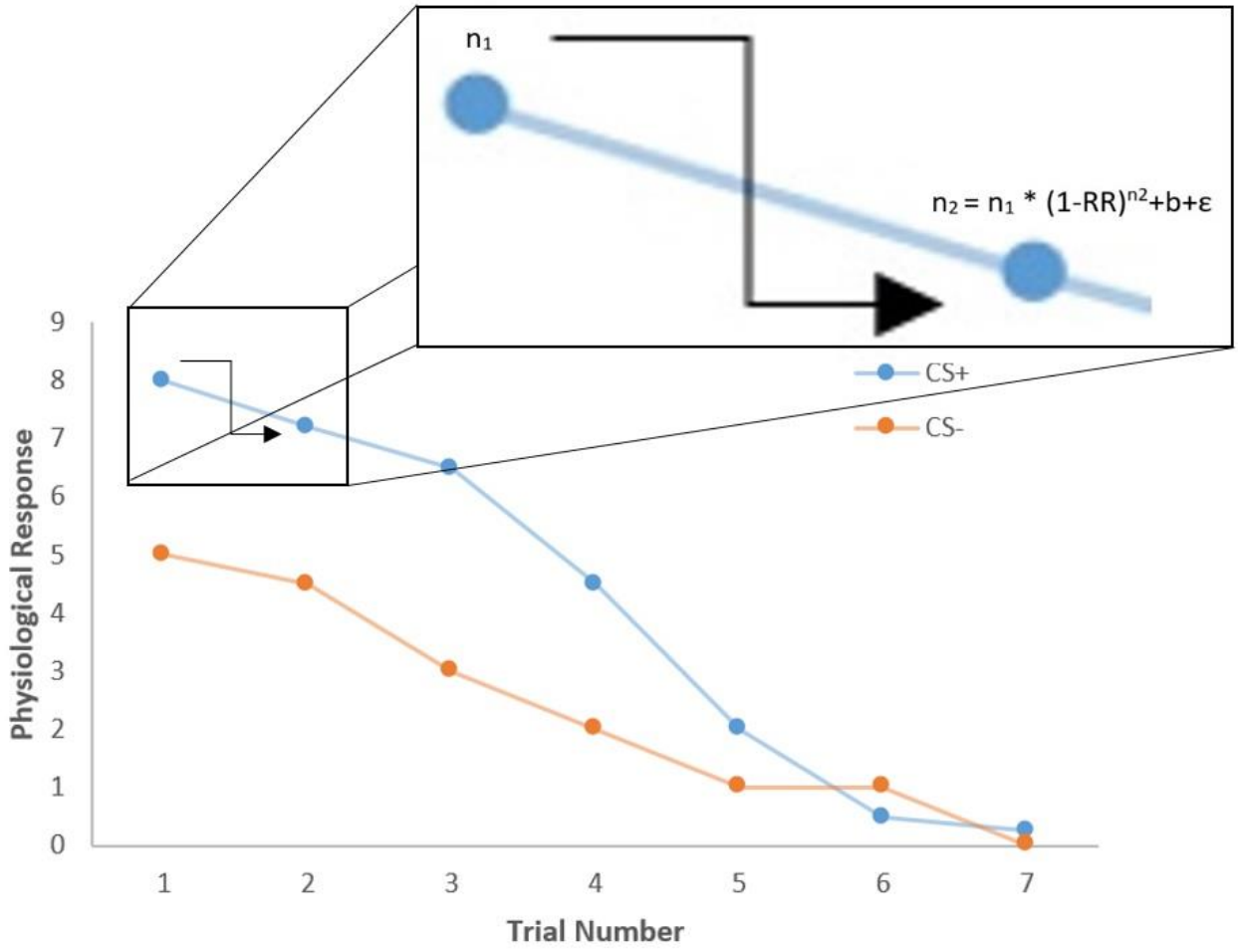

Figure 4. Physiological extinction as a product of previous trial arousal and likelihood of shock given reinforcement rate. Note: RR = reinforcement rate, b = baseline arousal

Work at the Zurich University, and others, have begun to optimise our understanding of the relationships between various physiological measures and psychological responding and have thus significantly expanded on the modelling approach to fear learning (Bach, 2014; Bach, Daunizeau, et al., 2010; Bach, Flandin, et al., 2010; Bach, Friston, et al., 2010; Benedek & Kaernbach, 2010a, 2010b). In these studies, the underlying relationship between

sympathetic nervous system responsivity and the psychological processes during associative learninghas been modelled using time-series data (ie. untransformed psychophysiological data) with promising future directions of precisely enhancing our understanding of what exactly is being measured during this paradigm (Bach et al., 2018). However, there is need for more research groups to attempt modelling strategies, and most of the more appealing approaches for fear extinction specifically are yet to be identified and tested. This argument is not a case for uniform analyses; rather, when models begin to be identified and tested, they may be routinely assessed by other researchers and continuously improved estimations of fear extinction in physiological responding may be obtained. In summary, modelling of physiological data during fear extinction can improve our understanding of the actual mechanisms involved as we are forced to specify and estimate processes that occur in the actual data (Forstmann & Wagenmakers, 2015a).

3.2.1 A Rerun of the Unpublished Data using a Simple Linear Method

Running the analyses for our unpublished data from earlier using linear trends, we observed several informative findings. During acquisition, there was a trial × stimulus linear contrast p=.018, η_p^2=.13, which suggested that SCRs for the CS- reduced across trials, whereas the CS+ did not (see Figure 2). This suggested that there was increased anticipation to the CS+ presentations, as would have been expected. The ANOVA was also significant (p=.032, η_p^2=.05), however as predicted this effect was smaller. During early extinction, there was a significant linear contrast of trial (p<.001, η_p^2=.42), suggesting that skin conductance responses reduced over trials overall. No other significant main effects or interactions were observed. Again, during late extinction there was a significant linear trial contrast, p<.001, η_p^2=.35, suggesting an overall decrease in skin conductance over trials. However, there were again no other significant effects. For both early and late extinction

phases the ANOVA effect for trial was significant, though with smaller effect sizes (η_p^2 = .21 and η_p^2=.14, respectively). For both of these effects, the only significant differences between trials were between trials one and the rest of the trials, which further highlights the limitations of using ANOVA for this form of data. Examining Figure 2, we decided to run exploratory quadratic trend analyses for the trial × group interactions for early and late extinction. These yielded a non-significant trend for early extinction (*p*=.085), and a significant trend for late extinction (*p*=.020; not significant after correction for multiple comparisons – see Figure 5). Our conclusion from this analysis would be that there was an interesting curve towards higher responding in the second wave group at the end of late extinction, which was captured by an exploratory quadratic trend that was not significant after correcting for multiple comparisons. Further investigation of why fear decreased and then increased for the second wave group could be the subject of replication so as to establish whether this effect was due to noise or to a mechanism involved in impaired fear extinction learning. Though, given our small sample size (*N*=45) and post-hoc analysis, this 'finding' would be admittedly highly exploratory and would require future research to establish that the effect was not driven by error variance.

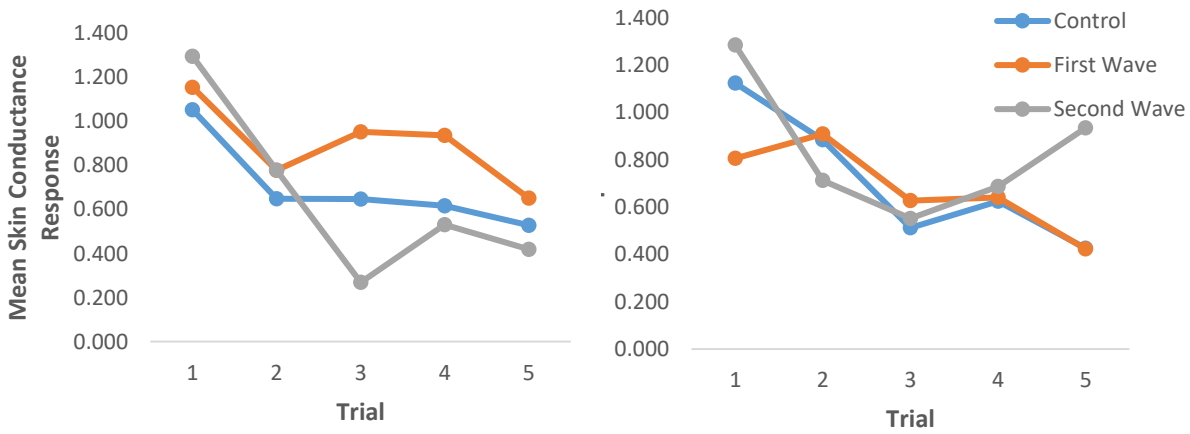

Figure 5. Quadratic trends for early and late extinction

In this analysis, we present and analyse the whole of the data that we collected. Due to using individual trial means and focused tests, we are able to accurately and transparently describe the patterns in our data, rather than leaving readers relying on distal and averaged numbers as evidence. If we had decided to use averaged data, readers would look at this data strategy, see the relationship to the data itself in Figure 2 and be able to interpret our findings in context. This is not to say that our averaged ANOVA method in Figure 2 was not useful for finding an effect. However, the use of focused tests rather than the more general ANOVA method allowed us to identify the physiological pattern occurring during fear extinction, as well as resulting in much higher effect sizes over the full dataset that reflected our improved ability to describe the actual patterns found in the data. The focused test also allowed us to narrow in on the exact nature of the difference between groups in the late extinction phase, whereas in our first analysis we were only able to recognise that there was a difference somewhere between groups on early and late extinction. This is clearly a more powerful, transparent and intuitive way of analysing the data and leaves the reader with no questions.

3.3 The Modelling Solution versus the Power Problem

At this stage, one important question that researchers might have would be: due to low power in most study designs, might not we sometimes expect the data to be linear, and sometimes not? How could we factor for this in advance? It is because of this very question that modelling the full data set is a critical issue. A practical matter in modelling physiological responding in this paradigm will be enhancing estimates of individual variability. As mentioned earlier, psychophysiological data is noisy data that is highly dependent on individual differences, a myriad of physiological processes and environmental or chance factors. One suggestion is that improving estimations of variability may potentially be achieved by increasing the number of trials in each experiment. If researchers and ethics

committees are agreeable, extra trials may then be added to all phases; otherwise, trials added to the end of studies (i.e. extinction phases) when no aversive stimuli are experienced would be acceptable. Whilst additional trials may only reflect floor effects in terms of extinction, additional trials at these times would give additional information as to the individual variability that is based on factors not associated with learning or extinction. However, this suggestion is secondary to utilising the data in between trials, such as using time series modelling. Including additional trials will also improve and strengthen estimations of trends across data, whether said trends be linear, quadratic, exponential or otherwise. However, if researchers wish to use this approach and leave their study open to the possibility that trends might present as different kinds of functions, then corrections for multiple comparisons will be essential. We recommend using more liberal techniques, such as the false discovery rate technique (Benjamini & Hochberg, 1995). Most efficiently, however, estimations of individual variability may also be improved by incorporating the data not included in the original data point calculations. Treating the data like a time-series, or as a continuous trace (or other model forms), reduction of the data may not be necessary (Bach et al., 2009). Intriguingly, there would also be potential to detect intuitive effects in the unused information, in addition to improving error estimates for the reliability and detection of conventional effects.

The ability to detect effects with relatively high power and precision using modelling of trends (or other models) should to some extent preclude the use of the retention index, which is used to determine relative fear during extinction recall. This is because effects found using statistical modelling during extinction recall will be based on comparisons between predicted slope and direction, rather than absolute physiological response levels, which the extinction index corrects for. In support of the extinction recall index, however, we found in combined fear acquisition data from the study described above and another

similar study (combined N=93) that the highest CS+ during acquisition was significantly correlated with the average CS+ during extinction (r=.70, p<.001), implying that indexing recall scores using this method might be valid. If researchers wish to use the extinction recall index, it is strongly recommended that a consistent strategy be adopted across studies.

4.1 Summary of Fundamental Statistical Issues and Recommendations

The major statistical issues identified in this paper are summarised below:

1. **Psychophysiological response data is "noisy".** High levels of individual variability in physiological data reduces power and makes true effects more difficult to establish due to random, arbitrary fluctuations in the data.
2. **Poor estimates of individual variability due to removal of information.** Although standard practice, removal of the larger part of the information (such as inter-trial interval information or removal of CS trials) during data preparation reduces estimates of individual variability, making the data even 'noisier'.
3. **Lack of modelling approaches.** Current data analysis strategies are not guided by theoretical parameters inherent to either fear extinction or psychophysiological responding and hence are poor explanations of the data set.
4. **Studies are underpowered.** Following from #3, studies are unable to detect effects as methods are not developed to detect appropriate effects from the available data. Current data analysis strategies used in the field are not able to cope with the lack of power and must be swapped and changed post hoc to detect effects.
5. **Lack of consistency of analyses between studies.** A wide variation in possible strategies that researchers can justifiably use greatly increases researcher degrees of freedom. Choosing an analysis post hoc is a common researcher degree of freedom

that seems innocent yet still has repercussions for error rates (Gelman & Loken, 2013).

6. **Lack of transparency in reported data.** Accurate interpretation of significant effects discovered with data averaged across trials is difficult. This is particularly troublesome given the prevalence of type 1, type M and type S errors in small sample sizes (Gelman & Carlin, 2014).
7. **Lack of consistency of blocks between studies.** Similar to the point above, there are dozens, even hundreds, of possible comparisons due to trial × block combinations. This problem is exacerbated by lack of universally agreed comparisons by which blocks reflect the processes of fear conditioning and extinction.
8. **Routine removal of trials, determined post hoc.** Although often reasonable and well justified, removal of data post hoc makes an implicit adjustment to false positive error rate, discussed in Gelman and Loken (2013).

We also have the following recommendations for improving on these issues. We are certain that most research groups will be able to implement at least a few of these strategies, even if they are not able or willing to work on a modelling approach. However, with a concerted and united effort we would expect that initially inaccessible statistical strategies and concepts would be able to be routinely implemented by a wide range of laboratories after some initial models begin to be developed.

1. **Modelling the error based on unused information.** Since most of the physiological data is unused when the data is prepared (ie. removal of CS trials and inter-trial responding), statisticians may consider optimising the amount of data removed

performed prior to analysis. Under a modelling approach, such as using time-series, removal of data would not be necessary.

2. **Full data made visible across all trials, or provided for independent analysis.** This removes the problems associated with apparently arbitrary, or otherwise justified, exclusion of experimental trials, and allows readers to interpret and visualise findings, reducing the rate of false positive results based on illogical or otherwise arbitrary patterns found in the full dataset.
3. **Rigorous development of predictive modelling, such as linear trends, exponential decay or Bayesian approaches.** This is likely to provide better, more intuitive descriptions of the data over time. More targeted strategies can boost power and reduce the number of comparisons, as well as make interpretation more intuitive and potentially consistent across studies.
4. **Testing increased number of trials.** Increasing the number of trials may be provide a better estimate of individual variability when learning/extinction has finished, potentially resulting more accurate and reliable findings if modelling as time-series data is not possible. However, this needs to be tested with issues such as study duration considered.
5. **Increased sample sizes to as many as practical.** An absolute minimum of 20 participants per group is likely to be necessary (Button et al., 2013; Simmons et al., 2011); however this is likely to remain problematic until error estimates are improved. Effect sizes in studies with sample sizes with less than 20 participants per group are very likely to be overestimated (Gelman & Carlin, 2014).
6. **Consistency of fear extinction index.** A consistent analysis strategy for the fear extinction index needs to be adopted across studies; however, adoption of one or more standard statistical models may make an index redundant.

7. **Multiple comparisons corrected using more liberal techniques, such as False Discovery Rate.** Due to relatively low power, conservative techniques may be unfavourable. FDR boosts power to detect real effects and should be adopted for researchers using exploratory modelling with simple techniques.
8. **Moving towards a hierarchical or multi-level modelling approach.** Any data collected tells us something both about that individual and the group that an individual comes from. Hierarchical modelling formally addresses this.

5.1 Conclusion

In this paper, we have outlined several existing problems in the statistical side of an important associative learning paradigm that is instrumental to our understanding of anxiety disorders and PTSD. This is an important issue, as the key point is that significant yet poorly interpretable results can be easily achieved in the current paradigm due to the lack adequate power and inconsistent analytical strategy between research teams. We have attempted to illustrate both how prevalent these issues are, as well as how easily, and perhaps subtly, results may be misconstrued and poor evidence may be convincingly presented as confirmatory. Not only does the absence of such a strategy limit certainty about findings, but it also detracts from an agreed definition of what fear extinction might actually look like physiologically, and a huge opportunity for furthering our understanding of fear extinction mechanisms based on feedback from our physiological data is being missed. In light of recent efforts in validating and standardising this area of research (Bach et al., 2018; Bach et al., 2013; Beckers et al., 2013; Lonsdorf et al., 2017; Lonsdorf & Merz, 2017; Scheveneels, Boddez, Vervliet, & Hermans, 2016; Staib et al., 2015), we feel that inconsistencies and lack of power in the research may be at least partly statistical- rather than design- or even sample-based. It is our duty to report our opinion that the field is consequently in an exploratory

phase, and researchers need to be aware that many findings will be either false, exaggerated or in the wrong direction. Critically, we do not intend to criticise any particular research team; the main point of this paper is to motivate interest in modelling physiological fear extinction data.

This conclusion, however, does not preclude the hope that good, informative science is the direction and ultimate outcome from our efforts. These issues are currently widespread in psychological research (Button et al., 2013; Gelman & Carlin, 2014; Gelman & Loken, 2013; Ioannidis, 2005; Open Science Collaboration, 2015; Wagenmakers et al., 2011) and most of the issues mentioned here are not entirely specific to this field. In many areas of psychology, teams are working towards ensuring that their paradigms are informative and reliable, and statistical modelling is a key step in achieving this (Forstmann & Wagenmakers, 2015a). Rigorous revision of the statistical properties of this paradigm is one of the essential developments that will greatly improve replicability. Many of the suggestions in this paper, namely the recommendation of simple modelling such as with linear, polynomial or exponential trends, are a preliminary attempt at making the essential concepts available to the wider field who are not experts in statistics and these techniques remain untested in larger databases. However, lessons can be learnt from existing modelling attempts by the Zurich group (Bach, 2014; Bach, Daunizeau, et al., 2010; Bach et al., 2009; Bach, Flandin, et al., 2010; Bach, Friston, et al., 2010; Bach et al., 2013), and modelling software for this data is now developed for MATLAB (PsPM, Bach, et al. 2009; Bach, et al. 2013). Efforts also need to be made so that modelling approaches are easily accessible to the broader research field who do not have advanced statistical training. It is ours, and others (Krypotos, Klugkist, et al., 2017; Lonsdorf et al., 2017; Staib et al., 2015) optimistic opinions, that a concerted and transparent effort from the researchers in this field is achievable; an effort that will greatly improve the reputation and validity of the field.

Acknowledgements

This work was supported by an National Health and Medical Research Council project grant to KLF (APP1050848).

References

Bach, D., Castegnetti, G., Korn, C. W., Gerster, S., Melinscak, F., & Moser, T. (2018). Psychophysiological modeling: Current state and future directions. *Psychophysiology, in press*. doi: 10.1111/psyp.13209

Bach, D. R. (2014). A head-to-head comparison of SCRalyze and Ledalab, two model-based methods for skin conductance analysis. *Biol Psychol, 103*, 63-68. doi:10.1016/j.biopsycho.2014.08.006

Bach, D. R., Daunizeau, J., Friston, K. J., & Dolan, R. J. (2010). Dynamic causal modelling of anticipatory skin conductance responses. *Biol Psychol, 85*(1), 163-170. doi:10.1016/j.biopsycho.2010.06.007

Bach, D. R., Flandin, G., Friston, K. J., & Dolan, R. J. (2009). Time-series analysis for rapid event-related skin conductance responses. *J Neurosci Methods, 184*(2), 224-234. doi:10.1016/j.jneumeth.2009.08.005

Bach, D. R., Flandin, G., Friston, K. J., & Dolan, R. J. (2010). Modelling event-related skin conductance responses. *Int J Psychophysiol, 75*(3), 349-356. doi:10.1016/j.ijpsycho.2010.01.005

Bach, D. R., Friston, K. J., & Dolan, R. J. (2010). Analytic measures for quantification of arousal from spontaneous skin conductance fluctuations. *Int J Psychophysiol, 76*(1), 52-55. doi:10.1016/j.ijpsycho.2010.01.011

Bach, D. R., Friston, K. J., & Dolan, R. J. (2013). An improved algorithm for model-based analysis of evoked skin conductance responses. *Biol Psychol, 94*(3), 490-497. doi:10.1016/j.biopsycho.2013.09.010

Bach, D. R., Tzovara, A., & Vunder, J. (2017). Blocking human fear memory with the matrix metalloproteinase inhibitor doxycycline. *Mol Psychiatry*. doi:10.1038/mp.2017.65

Beckers, T., Krypotos, A. M., Boddez, Y., Effting, M., & Kindt, M. (2013). What's wrong with fear conditioning? *Biol Psychol, 92*(1), 90-96. doi:10.1016/j.biopsycho.2011.12.015

Benedek, M., & Kaernbach, C. (2010a). A continuous measure of phasic electrodermal activity. *J Neurosci Methods, 190*(1), 80-91. doi:10.1016/j.jneumeth.2010.04.028

Benedek, M., & Kaernbach, C. (2010b). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology, 47*(4), 647-658. doi:10.1111/j.1469-8986.2009.00972.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300. doi:10.2307/2346101

Boucsein, W. (2012). *Electrodermal activity (2nd Edition)*. New York: Springer.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learn Mem, 11*(5), 485-494. doi:10.1101/lm.78804

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*(3), 153-178. doi:https://doi.org/10.1016/j.cogpsych.2007.12.002

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci, 14*(5), 365-376. doi:10.1038/nrn3475

Castegnetti, G., Tzovara, A., Staib, M., Gerster, S., & Bach, D. R. (2017). Assessing fear learning via conditioned respiratory amplitude responses. *Psychophysiology, 54*(2), 215-223. doi:10.1111/psyp.12778

Castegnetti, G., Tzovara, A., Staib, M., Paulus, P. C., Hofer, N., & Bach, D. R. (2016). Modeling fear-conditioned bradycardia in humans. *Psychophysiology, 53*(6), 930-939. doi:10.1111/psyp.12637

Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M., & Barch, D. M. (2016). Reduced model-based decision-making in schizophrenia. *J Abnorm Psychol, 125*(6), 777-787. doi:10.1037/abn0000164

Davis, M. (1992). The role of the amygdala in conditioned fear. In J. P. Aggleton (Ed.), *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 255-306). New York, NY, US: Wiley-Liss.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior, 37*(1), 1-20. doi:10.3758/LB.37.1.1

Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., . . . Baas, J. M. (2015). Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depress Anxiety, 32*(4), 239-253. doi:10.1002/da.22353

Eisenberg, I., Bissett, P., Zeynep Enkavi, A., Li, J., MacKinnon, D., Marsch, L., & Poldrack, R. (2018). *Uncovering mental structure through data-driven ontology discovery*.

Farrell, S., & Lewandowsky, S. (2015). An Introduction to Cognitive Modeling. In B. U. Forstmann & E. J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience*. New York: Springer.

Forstmann, B. U., & Wagenmakers, E. J. (2015a). *An Introduction to Model-Based Cognitive Neuroscience*.

Forstmann, B. U., & Wagenmakers, E. J. (2015b). Model-Based Cognitive Neuroscience: A Conceptual Introduction In B. U. Forstmann & E. J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience*. New York: Springer.

Gelman, A. (2016). More on my paper with John Carlin on Type M and Type S errors. Retrieved from http://andrewgelman.com/2016/11/13/more-on-my-paper-with-john-carlin-on-type-m-and-type-s-errors/

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspect Psychol Sci, 9*(6), 641-651. doi:10.1177/1745691614551642

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/

Glover, E. M., Phifer, J. E., Crain, D. F., Norrholm, S. D., Davis, M., Bradley, B., . . . Jovanovic, T. (2011). Tools for translational neuroscience: PTSD is associated with heightened fear responses using acoustic startle but not skin conductance measures. *Depress Anxiety, 28*(12), 1058-1066. doi:10.1002/da.20880

Graham, B. M., Callaghan, B. L., & Richardson, R. (2014). 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, 62*, 3-16. doi:10.1016/j.brat.2014.07.012

Graham, B. M., & Milad, M. R. (2013). Blockade of estrogen by hormonal contraceptives impairs fear extinction in female rats and women. *Biol Psychiatry, 73*(4), 371-378. doi:10.1016/j.biopsych.2012.09.018

Heathcote, A., Brown, S., & Wagenmakers, E. J. (2015). An Introduction to Good Practices in Cognitive Modeling. In B. U. Forstmann & E. J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience*. New York: NY: Springer.

Heathcote, A., Suraev, A., Curley, S., Gong, Q., Love, J., & Michie, P. T. (2015). Decision processes and the slowing of simple choices in schizophrenia. *J Abnorm Psychol, 124*(4), 961-974. doi:10.1037/abn0000117

Hinder, M. R., Schmidt, M. W., Garry, M. I., Carroll, T. J., & Summers, J. J. (2011). Absence of cross-limb transfer of performance gains following ballistic motor practice in older adults. *J Appl Physiol (1985), 110*(1), 166-175. doi:10.1152/japplphysiol.00958.2010

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124. doi:10.1371/journal.pmed.0020124

Joels, M., & Baram, T. Z. (2009). The neuro-symphony of stress. *Nat Rev Neurosci, 10*(6), 459-466. doi:10.1038/nrn2632

Jovanovic, T., Norrholm, S. D., Sakoman, A. J., Esterajher, S., & Kozarić-Kovačić, D. (2009). Altered Resting Psychophysiology and Startle Response in Croatian Combat Veterans with PTSD. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology, 71*(3), 264-268. doi:10.1016/j.ijpsycho.2008.10.007

Khemka, S., Tzovara, A., Gerster, S., Quednow, B. B., & Bach, D. R. (2017). Modeling startle eyeblink electromyogram to assess fear learning. *Psychophysiology, 54*(2), 204-214. doi:10.1111/psyp.12775

Kindt, M., Soeter, M., & Vervliet, B. (2009). Beyond extinction: erasing human fear responses and preventing the return of fear. *Nat Neurosci, 12*(3), 256-258. doi:10.1038/nn.2271

Klumpers, F., Denys, D., Kenemans, J. L., Grillon, C., van der Aart, J., & Baas, J. M. (2012). Testing the effects of Delta9-THC and D-cycloserine on extinction of conditioned fear in humans. *J Psychopharmacol, 26*(4), 471-478. doi:10.1177/0269881111431624

Korn, C. W., Staib, M., Tzovara, A., Castegnetti, G., & Bach, D. R. (2017). A pupil size response model to assess fear learning. *Psychophysiology, 54*(3), 330-343. doi:10.1111/psyp.12801

Krypotos, A. M., Blanken, T. F., Arnaudova, I., Matzke, D., & Beckers, T. (2017). A Primer on Bayesian Analysis for Experimental Psychopathologists. *J Exp Psychopathol, 8*(2), 140-157. doi:10.5127/jep.057316

Krypotos, A. M., & Engelhard, I. M. (2018). Testing a novelty-based extinction procedure for the reduction of conditioned avoidance. *J Behav Ther Exp Psychiatry, 60*, 22-28. doi:10.1016/j.jbtep.2018.02.006

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

Lang, P. J. (1995). The emotion probe. Studies of motivation and attention. *Am Psychol, 50*(5), 372-385.

LeDoux, J. E. (2014). Coming to terms with fear. *Proc Natl Acad Sci U S A, 111*(8), 2871-2878. doi:10.1073/pnas.1400335111

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nat Neurosci, 14*(10), 1250-1252. doi:10.1038/nn.2904

Lissek, S., Powers, A. S., McClure, E. B., Phelps, E. A., Woldehawariat, G., Grillon, C., & Pine, D. S. (2005). Classical fear conditioning in the anxiety disorders: a meta-analysis. *Behav Res Ther, 43*(11), 1391-1424. doi:10.1016/j.brat.2004.10.007

Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the Ability to Inhibit Simple and Choice Reaction Time Responses: A Model and a Method. *Journal of Experimental Psychology: Human Perception and Performance, 10*(2), 276-291.

Lonsdorf, T. B., Haaker, J., Schumann, D., Sommer, T., Bayer, J., Brassen, S., . . . Kalisch, R. (2015). Sex differences in conditioned stimulus discrimination during context-dependent fear learning and its retrieval in humans: the role of biological sex, contraceptives and menstrual cycle phases. *Journal of Psychiatry and Neuroscience, 40*(6), 368-375. doi:10.1503/140336

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., . . . Merz, C. J. (2017). 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, 77*, 247-285. doi:10.1016/j.neubiorev.2017.02.026

Lonsdorf, T. B., & Merz, C. J. (2017). More than just noise: Inter-individual differences in fear acquisition, extinction and return of fear in humans - Biological, experiential, temperamental factors, and methodological pitfalls. *Neurosci Biobehav Rev, 80*, 703-728. doi:10.1016/j.neubiorev.2017.07.007

Milad, M. R., Pitman, R. K., Ellis, C. B., Gold, A. L., Shin, L. M., Lasko, N. B., . . . Rauch, S. L. (2009). Neurobiological basis of failure to recall extinction memory in posttraumatic stress disorder. *Biol Psychiatry, 66*(12), 1075-1082. doi:10.1016/j.biopsych.2009.06.026

Milad, M. R., Quinn, B. T., Pitman, R. K., Orr, S. P., Fischl, B., & Rauch, S. L. (2005). Thickness of ventromedial prefrontal cortex in humans is correlated with extinction memory. *PNAS, 102*(30), 10706-10711. doi:10.1073/pnas.0502441102

Milad, M. R., & Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: ten years of progress. *Annu Rev Psychol, 63*, 129-151. doi:10.1146/annurev.psych.121208.131631

Milad, M. R., Rosenbaum, B. L., & Simon, N. M. (2014). Neuroscience of fear extinction: implications for assessment and treatment of fear-based and anxiety related disorders. *Behav Res Ther, 62*, 17-23. doi:10.1016/j.brat.2014.08.006

Milad, M. R., Wright, C. I., Orr, S. P., Pitman, R. K., Quirk, G. J., & Rauch, S. L. (2007). Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biol Psychiatry, 62*(5), 446-454. doi:10.1016/j.biopsych.2006.10.011

Milad, M. R., Zeidan, M. A., Contero, A., Pitman, R. K., Klibanski, A., Rauch, S. L., & Goldstein, J. M. (2010). The influence of gonadal hormones on conditioned fear extinction in healthy humans. *Neuroscience, 168*(3), 652-658. doi:10.1016/j.neuroscience.2010.04.030

Norrholm, S. D., Jovanovic, T., Olin, I. W., Sands, L. A., Karapanou, I., Bradley, B., & Ressler, K. J. (2011). Fear extinction in traumatized civilians with posttraumatic stress disorder: relation to symptom severity. *Biol Psychiatry, 69*(6), 556-563. doi:10.1016/j.biopsych.2010.09.013

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251).

Pace-Schott, E. F., Milad, M. R., Orr, S. P., Rauch, S. L., Stickgold, R., & Pitman, R. K. (2009). Sleep Promotes Generalization of Extinction of Conditioned Fear. *Sleep, 32*(1), 19-26.

Pappens, M., Schroijen, M., Sütterlin, S., Smets, E., Van den Bergh, O., Thayer, J. F., & Van Diest, I. (2014). Resting Heart Rate Variability Predicts Safety Learning and Fear Extinction in an Interoceptive Fear Conditioning Paradigm. *PLoS One, 9*(9), e105054. doi:10.1371/journal.pone.0105054

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction learning in humans: role of the amygdala and vmPFC. *Neuron, 43*(6), 897-905. doi:10.1016/j.neuron.2004.08.042

Rabinak, C. A., Angstadt, M., Sripada, C. S., Abelson, J. L., Liberzon, I., Milad, M. R., & Phan, K. L. (2013). Cannabinoid facilitation of fear extinction memory recall in humans. *Neuropharmacology, 64*, 396-402. doi:10.1016/j.neuropharm.2012.06.063

Raes, A. K., & De Raedt, R. (2012). The effect of counterconditioning on evaluative responses and harm expectancy in a fear conditioning paradigm. *Behav Ther, 43*(4), 757-767. doi:10.1016/j.beth.2012.03.012

Raez, M. B. I., Hussain, M. S., & Mohd-Yasin, F. (2006). Techniques of EMG signal analysis: detection, processing, classification and applications. *Biological Procedures Online, 8*, 11-35. doi:10.1251/bpo115

Raio, C. M., Brignoni-Perez, E., Goldman, D., & Phelps, E. A. (2014). Acute stress impairs the retrieval of extinction memory in humans. *Neurobiol Learn Mem, 112*, 212-221. doi:http://dx.doi.org/10.1016/j.nlm.2014.01.015

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation, 20*(4), 873-922.

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review, 111*(2), 333-367.

Rosenthal, R., Rosnow, R. L., & Rubin, B. (2000). *Contrasts and Effect Sizes in Behavioural Research: A Correlational Approach*. Cambridge: Cambridge University Press.

Scheveneels, S., Boddez, Y., Vervliet, B., & Hermans, D. (2016). The validity of laboratory-based treatment research: Bridging the gap between fear extinction and exposure treatment. *Behav Res Ther, 86*, 87-94. doi:10.1016/j.brat.2016.08.015

Schiller, D., Monfils, M.-H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2018). Addendum: Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature, 562*(7727), E21-E21. doi:10.1038/s41586-018-0405-7

Schiller, D., Monfils, M. H., Raio, C. M., Johnson, D. C., Ledoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature, 463*(7277), 49-53. doi:10.1038/nature08637

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci, 22*(11), 1359-1366. doi:10.1177/0956797611417632

Sjouwerman, R., Niehaus, J., Kuhn, M., & Lonsdorf, T. B. (2016). Don't startle me-Interference of startle probe presentations and intermittent ratings with fear acquisition. *Psychophysiology, 53*(12), 1889-1899. doi:10.1111/psyp.12761

Smeets, T., Cornelisse, S., Quaedflieg, C. W., Meyer, T., Jelicic, M., & Merckelbach, H. (2012). Introducing the Maastricht Acute Stress Test (MAST): a quick and non-invasive approach to elicit robust autonomic and glucocorticoid stress responses. *Psychoneuroendocrinology, 37*(12), 1998-2008. doi:10.1016/j.psyneuen.2012.04.012

Smith, P. L., & Ratcliff, R. (2015). An Introduction to the Diffusion Model of Decision Making. In B. U. Forstmann & E. J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience*. New York: Springer.

Soliman, F., Glatt, C. E., Bath, K. G., Levita, L., Jones, R. M., Pattwell, S. S., . . . Casey, B. J. (2010). A Genetic Variant BDNF Polymorphism Alters Extinction Learning in Both Mouse and Human. *Science (New York, N.Y.), 327*(5967), 863-866. doi:10.1126/science.1181886

Spoormaker, V. I., Schroter, M. S., Andrade, K. C., Dresler, M., Kiem, S. A., Goya-Maldonado, R., . . . Czisch, M. (2012). Effects of rapid eye movement sleep deprivation on fear extinction recall and prediction error signaling. *Hum Brain Mapp, 33*(10), 2362-2376. doi:10.1002/hbm.21369

Staib, M., Castegnetti, G., & Bach, D. R. (2015). Optimising a model-based approach to inferring fear learning from skin conductance responses. *J Neurosci Methods, 255*, 131-138. doi:10.1016/j.jneumeth.2015.08.009

Tzovara, A., Korn, C. W., & Bach, D. (in press). Human Pavlovian fear conditioning conforms to probabilistic learning. *PLOS Computational Biology*.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *J Pers Soc Psychol, 100*(3), 426-432. doi:10.1037/a0022790

Weidemann, G., Satkunarajah, M., & Lovibond, P. F. (2016). I Think, Therefore Eyeblink: The Importance of Contingency Awareness in Conditioning. *Psychol Sci, 27*(4), 467-475. doi:10.1177/0956797615625973

Wessa, W., & Flor, H. (2007). Failure of Extinction of Fear Responses in Posttraumatic Stress Disorder: Evidence From Second-Order Conditioning. *American Journal of Psychiatry, 164*(11), 1684-1692. doi:10.1176/appi.ajp.2007.07030525

Yehuda, R., Hoge, C. W., McFarlane, A. C., Vermetten, E., Lanius, R. A., Nievergelt, C. M., . . . Hyman, S. E. (2015). Post-traumatic stress disorder. *Nature Reviews Disease Primers, 1*, 15057. doi:10.1038/nrdp.2015.57

Yeomans, J. S., Li, L., Scott, B. W., & Frankland, P. W. (2002). Tactile, acoustic and vestibular systems sum to elicit the startle reflex. *Neurosci Biobehav Rev, 26*(1), 1-11.

Zeidan, M. A., Igoe, S. A., Linnman, C., Vitalo, A., Levine, J. B., Klibanski, A., . . . Milad, M. R. (2011). Estradiol modulates medial prefrontal cortex and amygdala activity during fear extinction in women and female rats. *Biol Psychiatry, 70*(10), 920-927. doi:10.1016/j.biopsych.2011.05.016

Zhang, S., Mano, H., Ganesh, G., Robbins, T., & Seymour, B. (2016). Dissociable Learning Processes Underlie Human Pain Conditioning. *Curr Biol, 26*(1), 52-58. doi:10.1016/j.cub.2015.10.066

Zuj, D. V., Palmer, M. A., Hsu, C. M., Nicholson, E. L., Cushing, P. J., Gray, K. E., & Felmingham, K. L. (2016). Impaired Fear Extinction Associated with Ptsd Increases with Hours-since-Waking. *Depress Anxiety, 33*(3), 203-210. doi:10.1002/da.22463

Zuj, D. V., Palmer, M. A., Lommen, M. J., & Felmingham, K. L. (2016). The centrality of fear extinction in linking risk factors to PTSD: A narrative review. *Neurosci Biobehav Rev, 69*, 15-35. doi:10.1016/j.neubiorev.2016.07.014