

UNIVERSITÄTSKLINIKUM HAMBURG-EPPENDORF

Institut für Systemische Neurowissenschaften

Prof. Dr. med. Christian Büchel

**THE ROBUSTNESS OF FEAR:
Challenges of methodological heterogeneity in fear conditioning research**

Dissertation

zur Erlangung des Doktorgrades Dr. rer. biol. hum.
an der Medizinischen Fakultät der Universität Hamburg.

vorgelegt von:

Maren Klingelhöfer-Jens
aus Dillenburg

Hamburg 2023

(wird von der Medizinischen Fakultät ausgefüllt)

**Angenommen von der
Medizinischen Fakultät der Universität Hamburg am:**

15.09.2023

**Veröffentlicht mit Genehmigung der
Medizinischen Fakultät der Universität Hamburg.**

Prüfungsausschuss, der/die Vorsitzende:

Prof. Dr. Tina B. Lonsdorf

Prüfungsausschuss, zweite/r Gutachter/in:

Prof. Dr. Anja Riesel

Für Matti, Tommi & Torben

List of publications

The present thesis draws on **Study I – IV**, which are cited in the text using Roman numerals:

- I.** Ehlers, M. R., Nold, J., Kuhn, M., **Klingelhöfer-Jens, M.**, & Lonsdorf, T. B. (2020). Revisiting potential associations between brain morphology, fear acquisition, and extinction through new data and a literature review. *Scientific Reports*, *10*(1), 19894. <https://doi.org/10.1038/s41598-020-76683-1>

- II.** **Klingelhöfer-Jens, M.**, Ehlers, M. R., Kuhn, M., Keyaniyan, V., & Lonsdorf, T. B. (2022). Robust group- but limited individual-level (longitudinal) reliability and insights into cross-phases response prediction of conditioned fear. *ELife*, *11*, e78717. <https://doi.org/10.7554/eLife.78717>

- III.** Lonsdorf, T. B., **Klingelhöfer-Jens, M.**, Andreatta, M., Beckers, T., Chalkia, A., Gerlicher, A., Jentsch, V. L., Meir Drexler, S., Mertens, G., Richter, J., Sjouwerman, R., Wendt, J., & Merz, C. J. (2019). Navigating the garden of forking paths for data exclusions in fear conditioning research. *ELife*, *8*, e52465. <https://doi.org/10.7554/eLife.52465>

- IV.** Lonsdorf, T. B., Gerlicher, A., **Klingelhöfer-Jens, M.**, & Krypotos, A.-M. (2022). Multiverse analyses in fear conditioning research. *Behaviour Research and Therapy*, *153*, 104072. <https://doi.org/10.1016/j.brat.2022.104072>

List of additional publications

Additional manuscripts that were written or published during the duration of the PhD program but were not incorporated into the thesis content:

- V. **Klingelhöfer-Jens, M.***, Morriss, J.*, & Lonsdorf, T. B. (2022). Effects of intolerance of uncertainty on subjective and psychophysiological measures during fear acquisition and delayed extinction. *International Journal of Psychophysiology*, *177*, 249–259.
<https://doi.org/10.1016/j.ijpsycho.2022.05.006>
- VI. Werner, F.*, **Klingelhöfer-Jens, M.***, Schumann, D., Gamer, M., Kalisch, R., Sommer, T., & Lonsdorf, T. B. (Preprint). Limited temporal stability of the Spielberger State-Trait Inventory over 3.5 years but excellent test-retest reliability for shorter time intervals.
<https://doi.org/10.31234/osf.io/mubgv>
- VII. Ruge, J. Ehlers, M. R., Kastrinogiannis, A., **Klingelhöfer-Jens., M.**, Koppold, A., & Lonsdorf, T. B. (Preprint). How adverse childhood experiences get under the skin: A systematic review, integration and methodological discussion on threat and reward learning mechanisms.
- VIII. **Klingelhöfer-Jens, M.**, Hutterer, K., Schiele, M., Leehr, E. J., Schumann, D., Rosenkrantz, K., Böhnlein, J., Repple, R., Deckert, J., Domschke, K., Dannlowski, U., Lueken, U., Reif, A., Romanos, M., Zwanzger, P., Pauli, P., Gamer, M., & Lonsdorf, T. B. (in preparation). Reduced discrimination between signals of danger and safety but not overgeneralization is linked to exposure to childhood adversity in a large sample of healthy adults.

* These authors contributed equally

Contents

1	Introduction.....	1
1.1	Meta-science and the three ‘R-terms’	2
1.2	The garden of forking paths	4
1.3	Overarching aims	5
1.4	Fear, anxiety, and pathology	5
1.5	The fear conditioning paradigm.....	7
1.5.1	Fear acquisition training.....	7
1.5.2	Extinction training.....	8
1.5.3	ROF manipulation and test	9
1.6	Quantifying the conditioned response.....	11
1.6.1	Skin conductance response (SCR)	11
1.6.2	Verbal reports.....	12
1.6.3	Neuroimaging: fMRI.....	13
1.6.4	Further readout measures and multivariate testing	14
1.7	The shift toward individual differences	15
1.8	The difficulty of conceptual replication attempts in an example (Study I)	17
1.9	Reliability as a prerequisite for robustness and replicability (Study II)	18
1.10	The garden of forking participant exclusions (Study III)	20
1.11	Introducing the multiverse idea: A compass for the garden of forking paths (Study IV)	22
2	Materials and methods	24
2.1	Participants.....	24
2.2	Experimental design.....	25
2.3	Experimental fear conditioning protocol and stimuli	25
2.4	Fear ratings and contingency awareness	26
2.5	Skin conductance response	27
2.6	Brain imaging.....	28
2.6.1	Study I: MRI data acquisition and analysis	28
2.6.2	Study II: fMRI data acquisition and analysis.....	28
2.7	Further outcome measures of no interest	29
2.7.1	Questionnaires.....	29

2.7.2	Other physiological outcomes.....	30
2.8	Additional complementary data sets.....	31
2.8.1	Additional data set used in Study III.....	31
2.8.2	Additional data set used in Study IV.....	32
2.9	Statistical analyses.....	32
3	Results and brief discussions.....	35
3.1	Study I: Revisiting potential associations between brain morphology, fear acquisition, and extinction through new data and a literature review.....	35
3.2	Study II: Robust group- but limited individual-level (longitudinal) reliability and insights into cross-phases response prediction of conditioned fear.....	39
3.3	Study III: Navigating the garden of forking paths for data exclusions in fear conditioning research.....	44
3.4	Study IV: Multiverse analyses in fear conditioning research.....	48
4	General discussion.....	52
5	List of Abbreviations.....	60
6	References.....	62
7	Study I.....	85
8	Study II.....	86
9	Study III.....	87
10	Study IV.....	88
11	Abstract.....	89
12	Zusammenfassung.....	90
13	Statement of author contribution.....	92
14	Danksagung.....	93
15	Curriculum vitae.....	94
16	Eidesstattliche Versicherung.....	95

“Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.”

– Marie Curie

1 Introduction

There are certainly things in our world that are rightly feared, such as the acute climate crisis or the current war in Ukraine – however, I believe that the central point of Marie Curie’s quote is surely true: the more we understand, the less we have to fear. In the context of anxiety- and fear-related disorders, we could also frame it this way: The more we understand underlying mechanisms through fear-, anxiety- and stress-related research, the better equipped we are to prevent or combat psychopathologies, and the fewer individuals have to experience and suffer from magnified anxiety and fears. The number of individuals affected by these conditions is tremendous: According to a nationwide survey conducted by Jacobi et al. (2014), around one-third of the participants fulfilled the criteria for at least one psychological disorder in the previous 12 months, with anxiety disorders being the most prevalent at 15.3%. This high prevalence is not only associated with considerable suffering for an enormous number of affected individuals, but also with high costs that burden the healthcare system arising directly from anxiety disorders themselves, but also indirectly from several physical diseases that are associated with anxiety disorders (Bandelow & Michaelis, 2015; Härter et al., 2003; Sareen et al., 2005).

To study fear- and anxiety-related processes in the laboratory, the fear conditioning paradigm, which is considered to be one of the most promising paradigms for translating basic research findings into clinical practice (Anderson & Insel, 2006; Beckers et al., 2023), is commonly used in humans and animals. Even though fear conditioning research has a long tradition (Milad & Quirk, 2012), open questions regarding its optimal experimental parameters, operationalization, and measurements remain. These questions are addressed in *meta-science* or *meta-research*, which aims at the pursuit of the scientific process by ensuring the protection of and adherence to research methods and standards of analysis (Ioannidis, 2018).

1.1 Meta-science and the three ‘R-terms’

In the field of meta-science, science investigates its own methods and practices (Ioannidis, 2018; Ioannidis et al., 2015). In essence, meta-science can be described as research on research. More precisely, the aim of meta-science is “the study of research itself: its methods, reporting, reproducibility, evaluation, and incentives.” (cf. Ioannidis, 2018, page 1).

There are three important and related ‘R-terms’ (McIntosh & Chambers, 2020) linked to meta-science. Even though they are sometimes used interchangeably in the literature, they differ in important aspects and should therefore be differentiated: *reproducibility*, *robustness*, and, *replication* (Nosek et al., 2022). *Reproducing* a prior finding refers to applying the same analysis to the same data whereas testing the *robustness* of a prior finding refers to the application of different analysis specifications to the same data to answer the same research question. In order to *replicate* a prior finding, the same analysis is applied to a different data set acquired using the same data recording and study design specifications (i.e., collection of new data in a new study; National Academies of Sciences, Engineering, and Medicine, 2019). With respect to replication, it is important to differentiate between direct and conceptual replications (Earp & Trafimow, 2015): Direct replication involves repeating the methods and procedures of an experiment as similar as possible to the original study, while conceptual replication involves using somewhat different methods and/or procedures (Schmidt, 2009). Both replication types serve different purposes: While a direct replication is more concerned with confirmation, a conceptual replication goes further by investigating generalizability and contributing to the theoretical understanding (Schmidt, 2009). However, replications also pose challenges in terms of their definition - or more precisely, it is still an open question when exactly a finding is considered to be replicated (e.g., LeBel et al., 2018).

All these aspects have received growing attention in recent years as they are fundamentally important for the credibility of scientific findings (Nosek et al., 2022; Simmons et al., 2011) and have been examined in detail in several studies: As an example, in a work on (computational) reproducibility of Artner et al., (2021), 30% of 232 identified findings in the field of psychology could not be reproduced. This is an alarming amount, considering that the identical analysis of identical data should invariably generate identical results. Other studies on this topic have reported similar rates of unsatisfying (computational) reproducibility (Hardwicke et al., 2021; Maassen et al., 2020). These studies also showed that the lack of reproducibility was either related to the fact that the original results were incorrect or that the analysis was not performed as described – both of which are a threat to the credibility of the findings.

Similarly, testing robustness in psychological research by supplying the same data and the same research question to different researchers yielded significantly different results, illustrating the strong dependence of results on the type of analysis selected and performed (Botvinik-Nezer et al., 2020; Silberzahn et al., 2018). The variability in both the significance of results and effect sizes was evident. Strikingly, none of the 70 teams included in the study of Botvinik-Nezer et al. (2020) used the same analytical approach. Neither level of expertise nor peer assessment of analysis quality explained the variability in approaches (Silberzahn et al., 2018). This concerning observation of limited robustness might signify that some findings are rather fragile, which might pose a threat to replicability and generalizability (Nosek et al., 2022) as results may be confined within specific experimental boundary conditions or in specific samples. In the studies by Botvinik-Nezer et al. (2020) and Silberzahn et al. (2018), the robustness of results was investigated across multiple individuals or laboratories, but it can also be assessed at other levels, such as using different types of data analysis for the same data.

The results of the direct replication attempts of psychological studies by the Open Science Collaboration (2015) are probably best known (but also criticized for underestimating replication rates, see Anderson et al., 2016; Gilbert et al., 2016): More than one-half of the studies examined could not be replicated. Other psychological or social science projects have undertaken similar replication attempts (Camerer et al., 2018; Protzko et al., 2020; Soto, 2019) – with some studies showing even relatively high replication rates (Protzko et al., 2020). Taken together, the results of these replication projects conducted by Camerer et al. (2018), Open Science Collaboration (2015), Protzko et al. (2020), and Soto (2019) and various multisite replications (i.e., Many Labs projects) revealed that only 64% of the studies showed statistically significant results in the same direction (Nosek et al., 2022) – even though these projects have been methodologically criticized for the definition and operationalization of a successful replication (e.g., the significance of p-values; LeBel et al., 2018).

In sum, these findings emphasize the importance of meta-science and the continuing challenges of producing reproducible, robust, and replicable results in psychological research. An important aspect that accentuates this challenge is the high degree of methodological flexibility that researchers encounter throughout the research process: With numerous decisions to be made during this process, a plethora of options arise at each decision point, resulting in substantial methodological heterogeneity in the literature. This considerable number of decisions involved in the research process is also known as ‘researcher degrees of freedom’ (Simmons et al., 2011) or the ‘garden of forking paths’ (Gelman & Loken, 2013).

1.2 The garden of forking paths

The term ‘garden of forking paths’ (Gelman & Loken, 2013) derives from the idea that with each decision, the researcher enters a path from which, in turn, further paths branch off (for an illustration, see **Figure 1**). Researchers frequently opt for a specific path concerning for instance the selection of a specific study design as well as the sample and outcome measures, data collection, and analysis types. This is done in good faith and with a rationale behind it, such as choosing the most apparent path, unfamiliarity with alternative paths, or relying on paths taken by the majority of the field. In addition to the difficulty of keeping track of all the options, making these choices is further complicated by the fact that there may not be a clear “best path” to answer a specific research question, and that none of the options may necessarily be wrong, but even equally justifiable (Silberzahn et al., 2018; Simmons et al., 2011). Yet, the consequences of choosing different paths may be substantial.

However, the impact of this heterogeneity can be systematically investigated in so-called manyverse- or multiverse approaches, in which some (i.e., manyverse, Kuhn et al., 2022) or most to all (i.e., multiverse, Del Giudice & Gangestad, 2021; Steegen et al., 2016) different plausible and equally justifiable options are investigated simultaneously (see **Figure 1B and C**). In contrast, opting solely for one specific path is also referred to as ‘universe’ (see **Figure 1A**). The many- and multiverse analyses might serve as a compass for the garden of forking paths. Their application is described in more detail in sections 1.9 – 1.11.

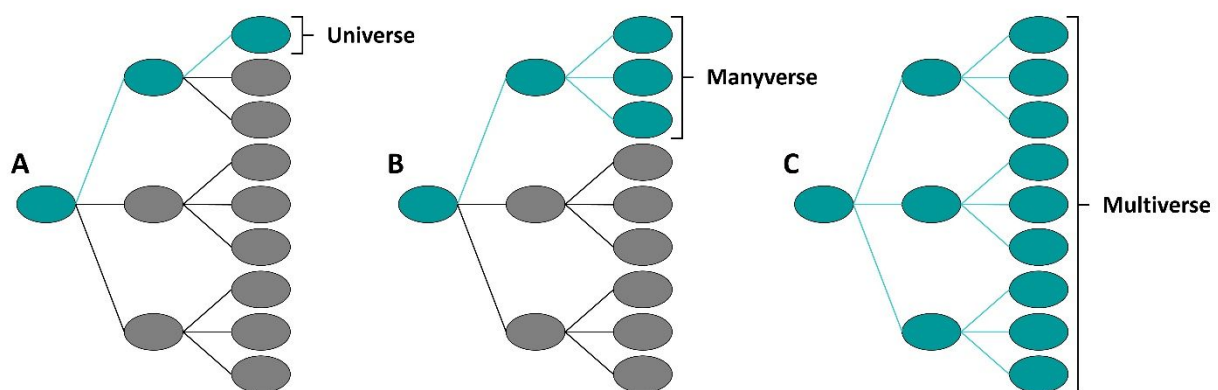


Figure 1 Illustration of the ‘garden of forking paths’ (Gelman & Loken, 2013) visualized through ellipses representing various decisions made (in green) throughout the research process. Typically, researchers follow one path resulting in a single universe as depicted in A. The manyverse includes several, but not all possible options (B) while the multiverse approach explores most or all potential and equally justifiable choices simultaneously (C).

1.3 Overarching aims

In this thesis, the meta-scientific topics of methodological heterogeneity, reproducibility, replicability, and in particular robustness are systematically addressed in four different fear conditioning studies, which will be discussed in more detail below. Even though this thesis does not address (computational) reproducibility empirically, the work included here takes (computational) reproducibility into account as two studies (i.e., **Study III** and **Study IV**) have been written as reproducible manuscripts using the open-source software R Markdown (<https://rmarkdown.rstudio.com/>). This allows others to generate the full manuscript file from the data and code shared publicly. Three of the here included studies focus on different aspects of robustness (**Studies II – IV**), and one study concentrates on conceptual replicability (**Study I**). More specifically, **Study I** attempts to conceptually replicate specific previously reported brain-behavior associations, while **Studies II – IV** focus on the robustness of results by examining the reliability of commonly used outcome measures in fear conditioning (**Study II**), and by addressing the impact of methodological heterogeneity related to exclusion criteria (**Study II**) and analytical approaches (**Study IV**) in fear conditioning research. In sum, the overall objective of this thesis was to elucidate the challenges produced by the high amount of methodological heterogeneity and work out potential remedies in fear conditioning research.

In the field of fear conditioning, meta-science has gained momentum with a growing body of meta-scientific studies contributing to methodological and theoretical knowledge. However, meta-science in the fear conditioning field is still in its infancy with only a few studies focusing on methodological heterogeneity or robustness (Haaker et al., 2014; Kuhn et al., 2022; Lonsdorf et al., 2019; Ney et al., 2020, 2022; Sjouwerman et al., 2022; Sjouwerman & Lonsdorf, 2020), replication (Bauer et al., 2020; Chalkia et al., 2020; Luyten & Beckers, 2017) and – to my knowledge – no study on (computational) reproducibility. Prior to introducing the studies included in this thesis, I will provide a comprehensive overview and explanation of fear and anxiety as well as the fear conditioning paradigm.

1.4 Fear, anxiety, and pathology

Fear is – despite being intensely unpleasant – a crucial emotion from an evolutionary perspective, as it is a fundamental aspect of survival (Ekman & Cordaro, 2011) and plays a central role in adapting to an ever-changing environment. When confronted with a predator, fear triggers a physical response that prepares the body for an adaptive defensive reaction: pulse and breathing

rates increase to supply optimal oxygen to the muscles, and adrenaline is released to prime the body for either fight or flight (Cannon, 1929; Lang et al., 2000). After defeating or escaping the predator, fear unfolds its second critical learning function by teaching us to recognize and respond to similar threats in the future. Consequently, fear has two essential functions: to react appropriately to current dangerous situations and to anticipate or avoid future hazardous situations by applying prior experiences to comparable stimuli or situations.

Whereas the fear described above is thought to be elicited by a specific threatening, time-locked stimulus, i.e. a ‘phasic response’, *anxiety* is considered to be more long-lasting and evoked by less specific stimuli or by anticipated threatening experiences in the future, i.e. a ‘sustained response’ (Davis et al., 2010; Lang et al., 2000). Most importantly, this distinction – although not necessarily mutually exclusive (Bublitzky et al., 2013; Stegmann et al., 2022) – provides a basis for the theoretical mechanisms behind different anxiety- and stress-related disorders: Phobias and post-traumatic stress disorder (PTSD), for instance, appear to be more associated with phasic fear processes, while generalized anxiety and panic disorder may be more related to sustained anxiety processes (Grillon, 2008).

These disorders relate to the dark side of fear, where magnified anxiety and fear lead to pathological conditions (for review, see Rosen & Schulkin, 1998). Fear then no longer fulfills its adaptive functions, but becomes dysfunctional by occurring in innocuous situations. The spectrum of innocuous situations that can potentially trigger fear is wide, and the resulting anxiety- and stress-related disorders range from specific phobias over post-traumatic stress disorder to generalized anxiety disorder. If left untreated, pathological fear is not only likely to persist, but also tends to generalize to similar stimuli (Cooper, van Dis, et al., 2022; Lissek, 2012), intensifying the personal burden by occurring in an increasing number of situations.

Fortunately, several treatment options exist that are highly effective – including, most notably, cognitive behavioral therapy (CBT; Barlow et al., 2007; Hofmann & Smits, 2008; Olatunji et al., 2013) the key component of which is believed to be exposure (Dunsmoor et al., 2015; Milad & Quirk, 2012; Rachman, 1989; Sánchez-Meca et al., 2010; Wolitzky-Taylor et al., 2008). Most critically, however, these therapeutic gains are often not long-lasting and patients frequently experience an increase of fear after a period of time, which can result in a full-blown relapse (Vervliet et al., 2013). Investigating the conditions under which therapeutic effects are preserved is therefore of major interest (Vervliet et al., 2013), but also answering other related clinical questions such as “Why do some people develop an anxiety disorder and others do not? What factors contribute to the perpetuation of the disorder and generalization of fear? Which individuals benefit

from which treatment?” The majority of fundamental anxiety and fear-related mechanisms – particularly at the individual level – are still unclear and urgently need to be studied in more detail (Lonsdorf & Merz, 2017). An important experimental tool for investigating these mechanisms constitutes the classical fear conditioning paradigm.

1.5 The fear conditioning paradigm

The acquisition, treatment, and relapse of (pathological) fear can be experimentally modeled in the laboratory by implementing the classical fear conditioning paradigm which can comprise the experimental phases *fear acquisition training*, *extinction training*, *return of fear (ROF) manipulation*, and *ROF-test*. These are described in more detail in the following sections.

1.5.1 Fear acquisition training

One assumed mechanism for the acquisition of (pathological) fear is Pavlovian conditioning, which is modeled in the first experimental phase - the acquisition training phase (Mineka & Oehlberg, 2008; Mineka & Zinbarg, 2006; Öhman & Mineka, 2001): In cue conditioning paradigms, an initially neutral stimulus (NS), e.g. a geometric shape (see **Figure 2**), is repeatedly paired with an unpleasant stimulus, the unconditioned stimulus (US), such as an electro-tactile stimulation, which represents a threat to the individual and elicits an unconditioned (fear) reaction (UR). Through repeated pairing, the NS gains the power to predict the US and becomes a conditioned stimulus (CS+). The CS+ thus serves as a danger cue that elicits the anticipatory, now conditioned, reaction (CR). CRs encompass orienting, fear, or defensive responses to threat (Lonsdorf et al., 2017), and to be elicited, a new memory trace has to be formed in an excitatory learning process that consists of three elements: mental representations of the CS+, of the US and their relation (Vervliet et al., 2013). This excitatory learning process results in a CS-US association which is stored in the new memory trace – the conditioned memory.

Another stimulus, the CS-, is never paired with the US, but instead of being a neutral control stimulus, it is assumed to gain the power to predict the absence of danger and to become a signal of safety in an inhibitory learning process (Lissek, Baas, et al., 2005; Lissek, Powers, et al., 2005). The implementation of two different stimuli refers to the so-called differential fear conditioning paradigm, which is the most commonly used paradigm in humans, whereas in rodents, single cue paradigms are frequently employed, in which the CS+ is presented without the CS- (Lonsdorf et al., 2017). An advantage of the differential fear conditioning paradigm is that the CRs can be quantified

as amplitude/strength differences between CS+ and CS- responses while controlling for differences between subjects in general response levels. Another advantage over the single-cue protocol is that there is no necessity for a control group, and within-subject designs are associated with increased statistical power and time efficiency (Lonsdorf et al., 2017).

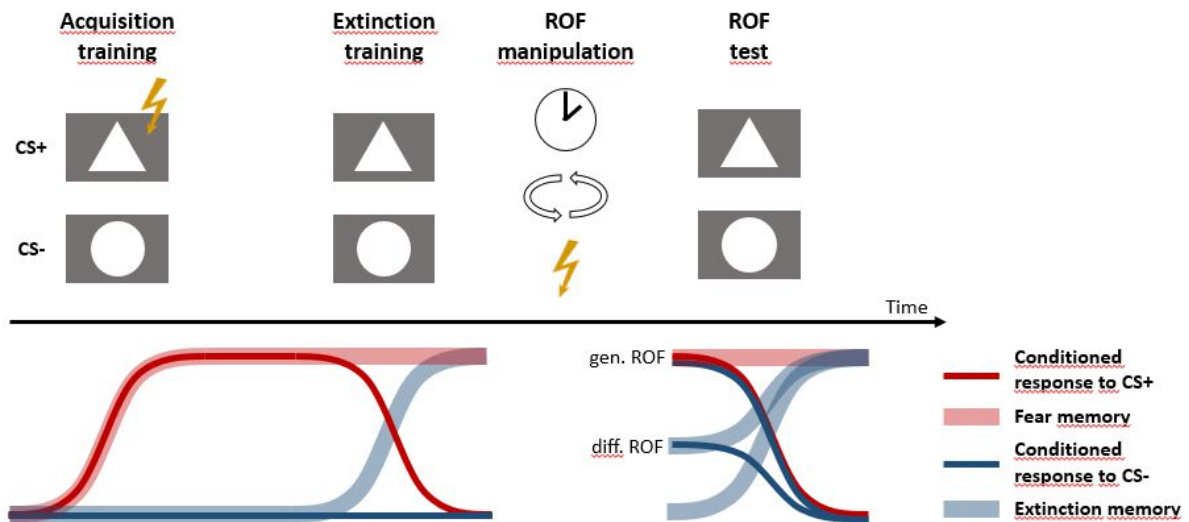


Figure 2 Illustration of the presented stimuli and ROF manipulation types (upper panel) as well as conditioned responses and memory traces across experimental phases (lower panel). Upper panel: The ROF manipulation types presented here include spontaneous recovery (depicted as a clock), contextual change (depicted as two circular arrows) and reinstatement (depicted as a flash). Lower panel: Darker, non-transparent lines represent conditioned responses to the CS+ (red) and CS- (blue), while lighter, transparent lines represent fear (red) and extinction (blue) memory traces. Note that after successful acquisition and extinction training, responding during ROF test can be either generalized with increased responses to the CS+ and the CS- or differential with increased responses solely to the CS+. CS = conditioned stimulus, ROF = return of fear, gen. = generalized, diff. = differential.

1.5.2 Extinction training

The extinction training phase serves as an experimental model for a secondary learning process (i.e., extinction learning) in which the contingencies change such that the CS+ no longer predicts the US and, as a result, the CR gradually wanes. Extinction learning appears to be the crucial ingredient in exposure interventions (Graham & Milad, 2011; Milad & Quirk, 2012; Rachman, 1989; Vervliet et al., 2013), proposing extinction training as a translational model for the treatment of anxiety- and stress-related disorders in the clinical context (Craske et al., 2018). The validity of extinction training as a translational model may be, however, less established than previously assumed (Scheveneels et al., 2016) implying the need for more extinction research in depth.

During extinction training, both CSs are presented again, however, in an absence of the US, typically resulting in a stepwise reduction of the CR over time or trials. Throughout this phase, a CS-noUS association is thought to be formed through an inhibitory learning process, producing a new memory trace: the extinction memory. Thus, it is assumed that the initial conditioning memory trace is not erased and remains intact, but is in conflict with the extinction memory (Bouton, 1993, 2004, 2014; Kim & Richardson, 2007; Myers & Davis, 2002, 2007; Rescorla, 1993, 2001). Thus, after successful extinction training, the CS+ includes both excitatory and inhibitory information. According to this *retrieval model* (Bouton, 2002; Craske et al., 2018), the dominance of one memory trace over the other determines which association (i.e., CS-US or CS-noUS) is retrieved resulting in the expression or suppression of the CR. In line with this, manipulations that undermine the CS-noUS association such as a change of context (Bouton & Bolles, 1979b; Bouton & Swartzentruber, 1989, see section 1.5.3) appear to weaken the inhibitory memory retrieval and facilitate ROF.

Extinction training can either follow immediately after acquisition training (i.e., *immediate extinction*) or with a certain time delay (i.e., *delayed extinction*). Most human work employs immediate extinction, which is more time and cost-efficient (for an overview, see Lonsdorf et al., 2017). Yet, there is also some evidence that immediate extinction may promote fragmented and slower extinction learning and reduce ROF effects (Golkar & Öhman, 2012; Norrholm et al., 2008) – a phenomenon termed *immediate extinction deficit* (Chang & Maren, 2009; Maren, 2014) – which is, however, not consistently observed across studies (Huff et al., 2009; Merz et al., 2016). On the contrary, allowing the fear memory to consolidate prior to extinction training is thought to be a more naturalistic model for the treatment of anxiety- and stress-related disorders, as there is usually a time delay between the acquisition and treatment of pathological fear (Haaker et al., 2014; Lonsdorf et al., 2017), and may thus enhance ecological validity (for discussion, see Maren, 2014). In addition, delayed extinction provides the opportunity to investigate another phenomenon called *fear recall*, which refers to the increased responses to the CS+ (i.e., *differential fear recall*) or to both CSs (i.e., *generalized fear recall*) during the first extinction training trials (Scharfenort et al., 2016). Delayed extinction training thus has a number of advantages – particularly for clinical translation (Lonsdorf et al., 2017).

1.5.3 ROF manipulation and test

Experimentally induced ROF, which can be implemented to probe retention of fear in the laboratory, has been proposed to model clinical relapse and is therefore employed systematically to investigate mechanisms that promote or prevent ROF (Scharfenort et al., 2016; Vervliet et al.,

2013). Following successful extinction training, ROF can be induced in the laboratory by i) the mere passage of time (*spontaneous recovery*), ii) contextual change (*renewal*), or iii) an unannounced re-exposure to the US (*reinstatement*). The subsequent paragraph provides brief explanations of spontaneous recovery and renewal, but a more detailed explanation of reinstatement, as this manipulation of the ROF was included in the study this thesis is based on. It is assumed that these ROF manipulations promote the dominance of the conditioned memory over the extinction memory, resulting in the re-occurrence of the CR. The success of ROF induction can be tested in the following phase, the ROF test phase, in which both CSs are presented again without US delivery. Similar to fear recall, ROF phenomena can be CS+ specific (Dirikx et al., 2007; Hermans et al., 2005; LaBar & Phelps, 2005) or non-specific, with the latter manifesting in an increased response to both CSs (Dirikx et al., 2009; Kull et al., 2012). These different response patterns are crucial as the tendency to generalize across safety and danger cues has been shown to be associated with pathological fear whereas intact discrimination between these cues is linked to resilience (Cooper, van Dis, et al., 2022; Craske et al., 2012; Duits et al., 2015; Lissek, Powers, et al., 2005).

Among these ROF manipulations, *spontaneous recovery* refers to the observable phenomenon of CR re-occurrence elicited by the mere passage of time – without any experimental manipulation – following successful extinction training and indexing retrieval of the fear memory which is also referred to as *ROF expression*. In contrast, *extinction retention*, whose examination is procedurally identical to that of spontaneous recovery, describes the absence of CR re-occurrence and indicates successful retrieval of the extinction memory (Lonsdorf et al., 2017). *Renewal*, as a further ROF phenomenon, describes the return of the CR following a change of context throughout the different experimental phases and after successful extinction training (Bouton, 2002).

Pavlov (1927) was the first to report the third ROF phenomenon described here, the *reinstatement*, which refers to the CR re-occurrence following an unsignalled representation of the US. While reinstatement has been examined in rodents since the 1970s, its studies in humans have only recently begun (Haaker et al., 2014). The retrieval model (described above, see section 1.5.2) provides a possible explanation for the emergence of the reinstatement phenomenon: The reinstated US is assumed to elicit contextual fear by retrieving the CS-US association that was formed during acquisition training. The context, which serves as an “occasion setter” (Holland, 1992; Schmajuk & Holland, 1998), is crucial as it conveys the re-validity of the CS-US association and promotes its dominance over the CS-noUS association acquired during extinction training (Bouton, 2004). Consistent with the retrieval model, experiments on the impact of contextual information support its pivotal role: the ROF appears to be stronger when the reinstatement context is identical to the

context of fear acquisition (Bouton, 2002) and extinction (Bouton & Bolles, 1979a; LaBar & Phelps, 2005). However, some findings challenge the retrieval model, such as the emergence of ROF in the case where the test context differs from the identical context of extinction and reinstatement (i.e., the absence of an occasion setter during the test phase; Westbrook et al., 2002). Nevertheless, the retrieval model can explain the majority of reinstatement findings in rodents and is the prevailing theory behind reinstatement effects (Haaker et al., 2014).

The fear conditioning paradigm is currently viewed as the most promising approach for transferring empirical findings on fear- and anxiety-related processes into clinical practice (Anderson & Insel, 2006; Beckers et al., 2023) and is therefore of great importance for anxiety- and stress-related research. In the following, *fear conditioning* will be used as an umbrella term encapsulating the different experimental stages (Lonsdorf et al., 2017).

1.6 Quantifying the conditioned response

As described above, differential fear conditioning protocols include both the CS+, which serves as a danger cue, and the CS-, which signals safety. The CR can be quantified as the difference between responses to the CS+ and CS-, also referred to as *CS discrimination* which ultimately reflects the associative learning process of interest (Lonsdorf et al., 2017).

As a multimodal and multidimensional construct, the defensive (conditioned) response to threat is experienced and expressed at different experiential and behavioral levels (Lonsdorf et al., 2017). Hence, it can be quantified at multiple response levels including i) subjective reports about experiences, ii) behavioral expression, and as changes in iii) physiological, and iv) neurobiological processes (Bradley & Lang, 2000). Typical outcome measures employed in human fear conditioning studies include the skin conductance response (SCR), fear ratings, and BOLD fMRI (blood oxygen level-dependent functional magnetic resonance imaging; Lonsdorf et al., 2017) which will be explained in more detail below, as they were obtained in the study upon which this thesis is primarily based.

1.6.1 Skin conductance response (SCR)

In the presence of unexpected, novel, presumably important, discrete stimuli – such as a threat in the form of the CS+ predicting the US – human sweat glands in the skin are filled with sweat, resulting in a lower resistance of the skin and hence better electrical conductance. These changes in electrodermal activity (EDA) can be measured as skin conductance responses (SCRs).

SCRs are directly related to sympathetic activity and thus reflect arousal processes (Hamm & Weike, 2005; Lipp, 2006), but also other processes or stimuli/task characteristics such as orientation, activation, attention, task significance, and affective stimulus intensity (Dawson et al., 2007).

More precisely, SCRs represent a phasic increase in EDA that typically emerges 0.5–5 s after stimulus onset (Boucsein et al., 2012) and can be quantified as the amplitude difference between the trough and the peak of this increase (Lykken & Venables, 1971). Typically, the CS+ as a danger cue elicits a larger SCR amplitude, whereas SCRs to the CS- – the safety cue – are lower (Lonsdorf et al., 2017). As explained above, the difference between responses to the CS+ and CS- yields CS discrimination. Taken together, SCRs are optimal for capturing defensive responding to discrete stimuli (Dawson et al., 2007) and are therefore predominantly employed in cue conditioning paradigms. Physiological responses such as SCRs have the additional advantage of being more objective and less prone to bias than verbal self-reports (Lonsdorf et al., 2017).

1.6.2 Verbal reports

Another commonly used measure to capture conditioned responding is to collect verbal self-reports that address subjective experiences of more affective aspects such as fear and distress (i.e., fear ratings), valence or arousal (i.e., valence or arousal ratings), or more cognitive aspects such as the expectancy/risk/probability of receiving the US (i.e., expectancy/risk/probability ratings) in the presence of CS+ and CS-. The latter also indicates the extent to which the contingency of CS and US co-occurrence has been learned. Alternatively, the awareness of this CS-US contingency can be explicitly checked retrospectively – e.g., in a post-experimental interview (for discussion, see Lonsdorf et al., 2017). These ratings are typically expressed on a visual analog scale (VAS), as forced choices (e.g., expected/not expected), or on another device, depending on the study design and apparatus. They can be provided either after each trial, after a block of trials, or after specific phases (for a discussion of the advantages and disadvantages of each procedure, see Lonsdorf et al., 2017), illustrating the learning progress in acquisition, extinction, and return of fear test at different resolutions.

Ratings are a common measure in fear conditioning studies as they can be easily obtained. They might, however, have the disadvantage of being susceptible to subjective biases that are generally present in self-reports (Choi & Pak, 2005), e.g. due to demand effects related to participants' assumption about the purpose of the experiment (Lipp, 2006). Furthermore, ratings may attenuate the learning progress during acquisition training (Atlas et al., 2022; Sjouwerman et

al., 2016). However, ratings contribute to the understanding of fear- and stress-related mechanisms as a multidimensional concept and are thus a valuable contribution to the overall picture.

1.6.3 Neuroimaging: fMRI

Functional magnetic resonance imaging (fMRI) is a complex procedure that – through further computational steps – results in functional images of the brain. As a detailed description would be beyond the scope of this thesis, only the general principles of this procedure will be outlined below.

fMRI is an imaging technique that takes advantage of the fact that neuronal activation is associated with increased blood flow (Fox et al., 1988; Fox & Raichle, 1986). When a specific brain region is activated, oxygen in the blood is expended resulting in an imbalance between deoxygenated and oxygenated blood. Deoxygenated and oxygenated blood have different magnetic properties which can be translated into an image contrast (i.e., blood oxygenation level dependent = BOLD contrast) which serves as an indirect measure of brain activation. Because of the need for a contrast between two experimental conditions, the CR as quantified by functional brain activation can only be determined for CS discrimination, but not for responses to CS+ or CS- alone.

In the following, I present some basic fMRI findings related to acquisition and extinction training as well as reinstatement, as these experimental phases were included in the design of the experiment which is the primary basis of this thesis. In a meta-analysis of Fullana et al. (2016), differential (i.e., CS+ vs. CS-) functional changes in neuronal activity were reported during fear acquisition in the insula, putamen, pallidum, caudate nucleus, nucleus accumbens, thalamus, precuneus, dorsal anterior cingulate cortex (dACC), dorsolateral prefrontal cortex (dlPFC), ventromedial prefrontal cortex (vmPFC), orbitofrontal cortex (OFC), hippocampus, somatosensory and motor areas as well as in the cerebellum. Surprisingly, the amygdala was not observed to be robustly involved as unanimously assumed (LeDoux, 2003; Öhman, 2009; Phelps et al., 2004), although there are potential explanations for these findings (Fullana et al., 2019). In another meta-analysis of Fullana et al. (2018) on differential neuronal involvement during extinction learning, the robustly identified regions encompassed dACC, mPFC, insula, dlPFC, putamen, caudate nucleus, pallidum, thalamus, motor cortical regions, and the pons. During differential and generalized reinstatement, distinct, but partially overlapping regions were activated. More precisely, the vmPFC, hippocampus, rectal gyrus, parietal operculum, and dorsal inferior temporal lobe were observed to be activated during differential reinstatement, and the thalamus, insula, occipital lobe, parietal operculum, inferior parietal lobe, supplementary motor area, cuneus,

cerebellum and bed nucleus of the stria terminalis (BNST) during generalized reinstatement suggesting distinct, but also intertwined functional processes (Scharfenort & Lonsdorf, 2016).

1.6.4 Further readout measures and multivariate testing

Other important measures, which will only be briefly explained here as they were not part of the present work, include fear-potentiated startle (FPS), heart rate (HR) and the pupillary response.

The startle reflex represents a defensive response that occurs during intense and abrupt auditory, visual or tactile events (Ramirez-Moreno & Sejnowski, 2012). Eyeblink responses are the most frequently measured electromyographic (EMG) changes among the startle reflex (Blumenthal et al., 2005). FPS describes the phenomenon of increased eyeblink responses under threat as compared to non-threatening stimuli (Brown et al., 1951; Davis & Astrachan, 1978; Hamm et al., 1993). It is considered to be an indicator of the valence of a stimulus (Lang et al., 1990; Lipp, 2006) and to have a strong translational value as the startle reflex is often used in rodent work (Kong et al., 2014).

In earlier and also more recent studies, HR has been demonstrated to be a useful, but also challenging tool in fear conditioning research (Castegnetti et al., 2016; Liu et al., 2013; Pappens et al., 2014; Tzovara et al., 2018) with HR showing both conditioned decelerations and accelerations (Castegnetti et al., 2016). While the (earlier) HR decelerations are assumed to reflect rather orienting responses (Hamm et al., 1993), HR accelerations appear to be related to defensive responding to the US predicted by the CS+ mirroring fear learning (Dimberg, 1987; Hamm et al., 1993; Moratti & Keil, 2005). Moreover, HR decelerations may be more likely to occur in the presence of neutral (Lipp & Vaitl, 1990) or safety stimuli (i.e., CS-; Ahrens et al., 2016), and some individuals habitually accelerate while others show decelerations. Thus, there is a multitude of individual, stimulus, and temporal factors to consider when designing or interpreting (the results of) a study employing HR (Lonsdorf et al., 2017).

Similar to SCRs, the pupillary response has been suggested to be linked to psychological arousal, but emerges comparably faster (Granholt & Steinhauer, 2004). It can be obtained in the behavioral laboratory by pupillometry or eye-tracking and represents a suitable CR indicator (Bitsios et al., 2004; Reinhard et al., 2006).

The majority of fear conditioning studies employ multiple read-out measures to capture defensive responding to threat at different response levels and as a multidimensional construct,

since different outcome measures are thought to tap into different anxiety- and stress-related processes in terms of content and timing: SCRs are presumed to reflect arousal and unfold rather slowly over seconds after CS onset, whereas FPS responses are thought to mirror valence and occur rather rapidly subsequent to the startle probe in the middle of CS presentations, close to the US, for a very short duration (Hamm & Weike, 2005; Lang et al., 1990; Lipp, 2006; Vrana et al., 1988). Thus, results obtained in the same study but at different response levels do not necessarily converge (for discussion, see Lonsdorf et al., 2017). Overall, the application of different outcome measures is recommended to capture the multidimensional nature of defensive responding (Haaker et al., 2014; Lonsdorf & Merz, 2017). However, as there is also evidence of mutual interference (Sjouwerman et al., 2016), a myriad of potential readout measures warrants a thoughtful selection and/or combination depending on the study design and goal (Lonsdorf et al., 2017).

1.7 The shift toward individual differences

Fear conditioning research to date has predominantly focused on general, basic mechanisms that reflect group-level effects such as the manipulation of experimental conditions (Lonsdorf & Merz, 2017). Tackling important – and still largely unanswered – clinical questions, such as why some individuals do not respond to treatment or experience relapse after successful treatment, requires a shift toward research that addresses individual differences, or more precisely, predictions at the individual level (Craske & Mystkowski, 2006; Fava et al., 2001; Yonkers et al., 2003).

The research of individual differences is part of a sub-discipline of psychology called *differential psychology*, which focuses, among other topics, on aspects in which individuals differ from each other or from themselves over time. Hence, individual differences can refer to differences between individuals (i.e., inter-individual differences), within an individual over time (i.e., intra-individual differences), or between individuals over time (inter-individual differences of intra-individual differences). This work will focus on inter-individual differences as these relate to the most pressing clinical questions outlined above. Unless otherwise indicated, the term *individual differences* will be used throughout this thesis to refer to inter-individual differences.

In fear conditioning research, individual differences have been largely neglected for decades, or considered “noise” or “unexplained variance” that needs to be eliminated in experiments to obtain robust group effects (Lonsdorf & Merz, 2017). These deviations from group average responding, which may contain important individual information, have only recently come into focus.

These individual differences might affect individual fear conditioning processes, which in turn may translate to and shed light on individual differences in clinical settings, such as the (non-) responsiveness to specific intervention programs, and thus can be assigned a key role in attempts to translate empirical findings into clinical applications. Recent work on individual differences has focused on temperamental factors such as trait anxiety (e.g., see Sjouwerman et al., 2020; Wroblewski et al., 2022; for review, see Lonsdorf & Merz, 2017;) or intolerance of uncertainty (e.g., Klingelhöfer-Jens et al., 2022; Mertens et al., 2022; for review, see Morriss et al., 2021), biological factors such as brain morphology (e.g., Abend et al., 2020; Cacciaglia et al., 2015) or genetic polymorphisms (e.g., Kastrati et al., 2022; Klumpers et al., 2015; Lonsdorf & Baas, 2017), experiential factors such as life adversity (e.g., Machlin et al., 2019; Scharfenort et al., 2016) and situational factors such as state anxiety (e.g., Glotzbach-Schoon et al., 2015; Kuhn et al., 2016) in order to unravel their contribution to the etiology and relapse of anxiety, but also to resilience mechanisms (Lonsdorf & Merz, 2017). These factors have been observed to be linked to altered fear processing mechanisms such as aberrant processing of the threat signal, decreased processing of the safety signal, reduced discrimination between threat and safety signals, impaired awareness of the CS-US coupling, or the tendency to generalize fear, all of which might be associated with increased anxiety to the point of psychopathology (e.g., Baas et al., 2008, Lissek et al., 2005; Lueken et al., 2014; for meta-analyses, see Duits et al., 2015; Lissek et al., 2009; for reviews, see Cooper, van Dis, et al., 2022; Lonsdorf & Merz, 2017; Nees et al., 2015).

Ultimately, a better understanding of individual difference factors might help us to identify individuals at risk or predict resilience, and may aid to develop individually tailored prevention and intervention programs, as has been done, for example, in medical disciplines (Insel, 2014). For instance, breast cancer patients could be assigned to different therapies based on genomic profiling to increase treatment success (i.e., precision medicine; Jiang et al., 2021).

However, while human fear conditioning research on individual difference factors is a growing field, there is still a need for improved study design, methodological procedures, and data analysis strategies that are tailored to this field. Apart from its meta-scientific objectives, this thesis will also examine several significant factors that might provide ideas and guidance for filling this room for improvement: While **Study I** investigates previously reported associations between individual differences in brain structure and defensive responding, **Study II** examines the extent of the feasibility of individual-level predictions using measures typically employed in the field. **Study III** explores the impact of excluding participants with specific individual differences on the outcomes, and finally, **Study IV** proposes a comprehensive and efficient method for testing various

analytic approaches which can be employed in individual difference research. In the next four sections 1.8 – 1.11, **Studies I – IV** and their specific aims are introduced in more detail.

1.8 The difficulty of conceptual replication attempts in an example (Study I)

The fundamental processes of fear conditioning and extinction, as well as the significance of individual differences in defensive reactions, are widely acknowledged, and the field of psychology and neuroscience has a rich tradition of exploring associations between brain morphology and behavior or physiology – known as structural-brain-behavior associations. It is therefore surprising that there has been so little research into how such differences in defensive responses might correspond to variations in brain structure, whose individual differences are commonly extracted from anatomical scans using magnetic resonance imaging (MRI) in human studies. However, structural-brain-behavior associations have been challenged by recent findings from a large cohort of healthy adults: It was demonstrated that significant associations are scarce and that the replication rates of such associations across various psychological measures appear to be low (Boekel et al., 2015; Genon et al., 2017; Kharabian Masouleh et al., 2019).

Previous studies on fear conditioning have shown that variations in brain morphology are linked to differences in conditioned responding during fear acquisition and extinction learning, as well as during retention test. While SCRs have been the most commonly used measure in these studies, a smaller number have employed FPS, ratings of valence, arousal, or awareness of CS-US contingency (Abend et al., 2020; Cacciaglia et al., 2015; Pohlack et al., 2012; Winkelmann et al., 2016). However, these studies investigating the link between brain morphology and individual differences in conditioned responding have produced inconsistent results: For instance, amygdala volume was associated with differential SCRs, but not arousal, valence, or CS-US contingency ratings during acquisition training in studies using cue conditioning paradigms (Cacciaglia et al., 2015; Winkelmann et al., 2016). More precisely, Winkelmann et al. (2016) showed effects for the right amygdala, and Cacciaglia et al. (2015) for the left amygdala. In these two studies, the association between amygdala volume and differential conditioned responding during acquisition training was positive, whereas, in another study, researchers reported an (although not significant) negative relationship (Hartley et al., 2011). Moreover, Winkelmann et al. (2016) observed the association between amygdala volume and differential conditioned responding in one sample for the early acquisition phase and in another sample for the late acquisition phase. Other brain structures associated with conditioned responding during acquisition training include the insula (Hartley et al., 2011), dACC (Milad, Quirk, et al., 2007), and dm/dIPFC (Abend et al., 2020). A

study using a contextual fear conditioning paradigm reported also an involvement of the hippocampus (Pohlack et al., 2012). During extinction training and extinction retention, differential responding was related to the thickness of the vmPFC (Hartley et al., 2011; Rauch et al., 2005; Winkelmann et al., 2016).

Of note, in most of these studies, sample sizes were rather small with a range of 14 (Milad et al., 2007; Rauch et al., 2005) to 52 participants (Cacciaglia et al., 2015). An exception to this constitutes the study of Abend et al. (2020) which included a larger sample of 351 participants. However, they focused on general reactivity rather than associative learning processes such as CS discrimination, as SCRs were averaged across CS+ and CS-. Despite the small sample sizes in most previous studies, which would be considered underpowered today, surprisingly strong correlations are reported.

Given that the robustness of structural brain-behavior associations has been recently called into question (Kharabian Masouleh et al., 2019; Masouleh et al., 2020), the aim of **Study I** was to conceptually replicate previous findings linking cortical thickness/subcortical volume to conditioned responding in SCRs and fear ratings, both in a large sample and within a single study.

In general, several factors appear to contribute to the success or failure of replication attempts. These include aspects such as operationalization of the constructs under study, but also methodological aspects, such as study design or type of data analysis. Whereas these aspects depend on decisions made across the research process, one important but often neglected aspect impacting replicability arises from the outcome measure itself: measurement reliability.

1.9 Reliability as a prerequisite for robustness and replicability (Study II)

Measurement reliability has gained momentum in the recent past, resulting in a growing demand for increased attention for and research in this domain (Fröhner et al., 2019; Hedge et al., 2018; Zuo et al., 2019). One reason might be that measurement reliability is central to meeting the challenges of robustness and replicability of results – what some researchers refer to as ‘replication crisis’ (Open Science Collaboration, 2015; Stroebe & Strack, 2014): Reliable measures increase the likelihood of replicating previous findings (LeBel & Paunonen, 2011), as reliable measures provide consistent data and produce similar results each time they are employed under the same conditions (Heale & Twycross, 2015). Yet another reason may be that the field of fear conditioning has shifted from investigating rather basic general principles derived from group averages to questions and processes at the individual level (Lonsdorf & Merz, 2017), for which measurement reliability is an

important prerequisite. Individual-level predictions are essential because studies using fear conditioning paradigms have significant potential to translate neuroscientific discoveries into clinical applications (Anderson & Insel, 2006; Cooper, van Dis, et al., 2022; Fullana et al., 2020; Milad & Quirk, 2012), where important potential for improvement remains, such as individual treatment progression and therapeutic efficacy. Although reliability is also important at the group level to safeguard group findings such as the effects of experimental manipulations (Lonsdorf & Merz, 2017). Furthermore, the reliability of a given measure is also crucial for investigating its correlations with other (individual difference) variables. This is because the reliability of a given measure sets an upper limit to the maximum observable correlation between that measure and another measure or (individual difference) variables (Spearman, 1910).

Study II aims to explicitly tackle reliability in fear conditioning research, as there has been surprisingly little effort in scrutinizing reliability in the field. To date, there are only five studies that addressed longitudinal reliability (Cooper, Dunsmoor, et al., 2022; Fredrikson et al., 1993; Ridderbusch et al., 2021; Torrents-Rodas et al., 2014; Zeidan et al., 2012), which in the literature is also referred to as test-retest reliability, and only one that has examined internal consistency (Fredrikson et al., 1993). While longitudinal reliability provides information about the extent to which responses of an individual (individual-level longitudinal reliability) or a group (group-level longitudinal reliability) are stable over time, internal consistency mirrors the extent to which items – or trials – measure the same construct. However, it is difficult to draw a consistent picture from these studies because they differ widely in types of outcome measures and reliability estimates included, as well as lengths of retest intervals and sample sizes.

The lack of empirical investigations is also true for the relationship of conditioned responding across very short intervals – or more precisely, between individual experimental phases. Interestingly, such relationships are often implicitly assumed – for instance, some researchers “control” responding in later experimental phases for responding in earlier experimental phases (e.g., see Milad et al., 2009, critically discussed in Lonsdorf et al., 2019). Both the presence (Foa et al., 1983; Gershman & Hartley, 2015; Rauch et al., 2004) and the absence of associations between responding across different experimental phases (Bouton et al., 2006; Plendl & Wotjak, 2010; Prenoveau et al., 2013; Shumake et al., 2014) or therapeutic sessions (Kozak et al., 1988; Pitman et al., 1996; Riley et al., 1995) have been supported by direct findings from animal studies or indirect findings from patient research. In human fear conditioning paradigms, however, these associations of responding across different experimental phases have rarely been directly investigated.

To address these gaps in the literature, the aims of **Study II** were to investigate i) longitudinal reliability of conditioned responding at the individual and the group level, ii) internal consistency as well as iii) the association of conditioned responding across experimental phases. This was done for different data specifications such as different phase operationalizations (e.g., the operationalization of acquisition training as averaged across all acquisition training trials or across the last two acquisition training trials), data transformations (e.g., log-transformation and range correction of SCR data), and reliability measures to account for the high level of methodological heterogeneity in the literature in a small manyverse approach.

In addition to reliability, another important aspect that is crucial for the robustness and the generalizability of results is the composition of the sample, which is – among other factors – influenced by the exclusion of specific individuals. The criteria for participant exclusion can be very heterogeneous. Possible procedures and rationales underlying these exclusions will be presented in the next section.

1.10 The garden of forking participant exclusions (Study III)

In the human fear conditioning field, research questions frequently focus on objectives such as modification of conditioned responses, generalization, consolidation, or retrieval. It has been commonly assumed that studying these processes necessitates acquiring a strong conditioned response in the first place. As a result, one of the researcher's decisions concerns the (often routinely conducted) exclusion of participants due to 'non-learning' or 'non-responding' in SCRs – the most frequently used outcome measure in fear conditioning (Lonsdorf et al., 2017): While 'non-learning' refers to the absence of physiological CS discrimination (i.e., in SCRs) during acquisition training, 'non-responding' refers to the absence of a stimulus-driven physiological response.

However, the exclusion of 'non-learners' and 'non-responders' poses a number of challenges: First, the definition of 'non-learners' and 'non-responders' vary considerably across studies: e.g., the exclusion of 'non-learners' is also referred to as 'performance-based exclusion' or 'exclusion of outliers'. Second, there is no consistent procedure for handling 'non-learners' and 'non-responders' as some researchers exclude 'non-learners' (e.g., Ahmed & Lovibond, 2019), while some exclude 'non-responders' (e.g., Morriss et al., 2018), and some exclude both (e.g., Hartley et al., 2014). Third, 'non-learners' are typically excluded based on a lack of CS discrimination in a single outcome measure (i.e., SCR), but it is common practice to exclude them from *all* analyses. However, as SCR serves only as one proxy to measure fear learning, this

procedure disregards the fact that CS discrimination might be evident in other simultaneously acquired outcome measures such as FPS or ratings. Fourth, the extent of CS discrimination was observed to vary as a function of individual difference factors such as intolerance of uncertainty (e.g., Johnson et al., 2022; Klingelhöfer-Jens et al., 2022; Morriss et al., 2021; Wroblewski et al., 2022) or trait anxiety (e.g., Gazendam et al., 2013; Indovina et al., 2011; Sjouwerman et al., 2020; Staples-Bradley et al., 2018) – even though results are mixed (e.g., Mertens et al., 2022; Torrents-Rodas et al., 2013; for an overview, see Lonsdorf & Merz, 2017;). Thus, excluding participants based on their ability to physiologically discriminate between conditioned stimuli (CSs) could lead to the exclusion of relevant subpopulations, including participants with subclinical symptoms. Taken together, excluding ‘non-learners’ and/or ‘non-responders’ might lead to a substantial sample bias. This would not only compromise the generalizability of results, as specific subpopulations could be overlooked, but also impede research on individual differences and the transfer into clinical applications.

As the concerns raised above are primarily of theoretical nature, we wanted to empirically approach the relevant topics of ‘non-learners’ and ‘non-responders’ in two different data sets (i.e., one main data set, which was obtained in the framework of this thesis and a complementary data set, see section 2.8) in **Study III** by i) systematically identifying the criteria of ‘non-learners’ and ‘non-responders’, ii) investigating the consequences of the exclusion of ‘non-learners’ and ‘non-responders’ according to different criteria on results and their interpretation and iii) elaborating our specific recommendations for future definition and handling ‘non-responders’ and ‘non-learners’ based on empirical evidence.

Navigating the complex garden of forking paths requires careful consideration of many decision points, including how to handle ‘non-responders’ and ‘non-learners’. The decision at this specific forking path predominantly concerns the pre-processing of the data. Selecting an appropriate statistical analysis model to follow the pre-processing step is equally crucial, but represents a similar significant challenge due to the multitude of options available and the heterogeneity in the literature. To effectively explore (most to) all possible paths simultaneously, a multiverse analysis (Del Giudice & Gangestad, 2021; Steegen et al., 2016), as demonstrated in **Study IV** and described in the following section, is a valuable tool.

1.11 Introducing the multiverse idea: A compass for the garden of forking paths (Study IV)

In the fear conditioning field, the garden of forking paths has been mainly scrutinized by focusing on heterogeneity in the operationalization of constructs or data pre-processing such as the impact of different definitions of ‘non-learners’/‘non-responders’ (see **Study III**), ‘extinction retention’ (Lonsdorf et al., 2019) or SCR quantification approaches (Kuhn et al., 2022; Sjouwerman et al., 2022). However, a systematic assessment of the impact of heterogeneity in the choice of statistical models is lacking so far. This choice concerns, for example, the use of a specific statistical procedure (e.g., analysis of variance [ANOVA] or mixed model), the inclusion of covariates, or how many trials are included in the analyses. Even though these decisions may be equally valid options, they might produce different outcomes resulting in different or even contrasting conclusions (Botvinik-Nezer et al., 2020; Dutilh et al., 2019; Kuhn et al., 2022; Lonsdorf et al., 2019; Silberzahn et al., 2018). This not only impedes the comparison of studies, but also complicates the integration of different study results, for instance within the framework of meta-analytical approaches.

The degrees of freedom in the choice of statistical models are enhanced by the frequent lack of formalization of psychological theories: These are predominantly verbally formulated in the form of a verbal description of latent constructs and their associations (Farrell & Lewandowsky, 2018; Lewandowsky & Farrell, 2010). As an example, a theory in the fear conditioning field might predict that the repeated pairing of CS+ and US will result in elevated SCRs to the CS+ compared to the CS-. This theory, however, does not define the magnitude of this response difference or the number of trials needed to observe this difference. The exact assumptions about this reside in the researcher degrees of freedom (Simmons et al., 2011) and thus, this theory may be translated into different statistical models (Muthukrishna & Henrich, 2019).

The missing ingredient to formalize verbal theories in the fear conditioning field constitutes, however, a deep understanding of how individual methodological specifications affect the results. A promising path to improve this understanding could be the implementation of a multiverse approach (Del Giudice & Gangestad, 2021; Steegen et al., 2016). The multiverse approach can include either all different plausible and equally justifiable i) preprocessing pipelines of the data (i.e., ‘data multiverse’), ii) applications of different statistical models (i.e., ‘model multiverse’), or iii) a combination of both. These multiverses consist of numerous universes which are generated based on the specific decisions for certain preprocessing or analysis pipelines (see **Figure 1C**). The multiverse approach aims to obtain better estimators of potential effects by including the full range

of possible decisions because these decisions are frequently made arbitrarily. Convergence of results based on different pipelines would suggest a rather robust effect. In contrast, heterogeneous results might indicate a systematic dependence of the effect on the precise pipeline specification putatively providing information on which boundary conditions may scale the effect size.

Thus, the aims of **Study IV** were to i) import the multiverse idea into the fear conditioning field, ii) demonstrate how the implementation of different statistical models (as identified via a systematic literature search) can affect the results in two independent data sets, and finally, iii) to introduce the open software R package *multifear* (<https://github.com/AngelosPsy/multifear>) with which such multiverse analyses can be conducted by typing a single line of code. Of note, the statistical models incorporate data that have undergone distinct data reduction approaches such as including single trial data or averages across all trials or specific blocks of an experimental phase. These data reduction approaches were also determined through the systematic literature review.

2 Materials and methods

2.1 Participants

Participants of the study on which this thesis is based were pre-selected from a larger cohort of the Collaborative Research Center CRC 58 on the basis of the absence of childhood maltreatment according to the Childhood Trauma Questionnaire and applying critical cut-offs (Bernstein & Fink, 1998; Häuser et al., 2011). Participants were contacted via phone and screened for exclusion criteria which comprised left-handedness, claustrophobia, cardiac pacemaker, non-MR-compatible metal implants, brain surgery, participation in pharmacological studies within the past 2 weeks, medication except for oral contraceptives, internal medical disorders, chronic pain, neurological disorders, psychiatric disorders, metabolic disorders, acute infections, complications with anesthesia in the past, and pregnancy. All participants were right-handers and had corrected to normal or normal vision. The Ethics Committee of the General Medical Council Hamburg (PV 5157) had approved the experimental protocol, to which all participants gave written informed consent prior to the study. The study was performed in conformity with the Declaration of Helsinki. All participants were unfamiliar with the experimental setup. In the case of participation at all measurement time points (T0 – T5, see next section 2.2), participants received 300 € as financial compensation of which 40 € were handed out in the form of 10 euro vouchers.

For **Studies I, III and IV**, data from the first measurement time point (T0) were included, and a different, albeit strongly overlapping, sample of participants was excluded because i) different phases and ii) outcome measures were involved in the analyses and iii) different research questions were addressed. For instance, non-responders (for a definition in this study, see section 1.6.1) who were excluded in **Study I**, were retained in the sample of **Study III** as they were subjects of interest. In **Study I**, participants were excluded due to a deviating protocol ($n = 1$), missing data ($n = 1$), technical issues ($n = 3$), and SCR non-responding ($n = 8$) resulting in $N = 107$ participants from whom MRI and SCR data were analyzed (female_N = 71, male_N = 36, age_M = 24.4, age_{SD} = 3.7). Another $n = 4$ and $n = 12$ participants were excluded from analyses due to missing fear rating data on day 1 and day 2 respectively. For analyses in **Study III**, one participant was excluded due to a deviating protocol, yielding $N = 119$ participants (female_N = 79, male_N = 40, age_M = 25, age_{SD} = 4), and in **Study IV**, four participants were excluded due to deviating protocol and technical issues leaving $N = 116$ participants for analyses (female_N = 77, male_N = 39, age_M = 24.38, age_{SD} = 0.34).

For cross-sectional analyses in **Study II**, the identical sample as in **Study I** was used. For longitudinal analyses, 16 participants were excluded due to technical issues, three participants due to a deviating protocol, and eleven participants due to SCR non-responding. In **Study II**, the same non-responder criteria were applied as in **Study I**. In addition, 21 participants dropped out during the time interval between T0 and T1 leaving 71 participants for longitudinal analyses (female_N = 41, male_N = 30, age_M = 24.6, age_{SD} = 3.8).

2.2 Experimental design

This work includes data from the first two measurement time points of a larger longitudinal study which comprised a total of six time points (T0 – T5) that were all approximately six months apart. At the first two measurement time points, a differential fear conditioning paradigm was conducted on two consecutive days in MR environment. Stimuli and the experimental procedure were identical at these two time points. A battery of questionnaires was completed at all time points. At the last time point T5, participants returned to the laboratory and a structural MR scan was obtained. As T5 data are not included in the studies of this thesis, no further details are provided here.

2.3 Experimental fear conditioning protocol and stimuli

The fear conditioning experiment started on day 1 with a habituation phase followed by an acquisition training phase in which CS+ and CS- were each presented 7 and 14 times (with a total of 14 and 28 trials) respectively. During extinction training and reinstatement-test phase on day 2, CS+ and CS- were each presented 14 and 7 times (with a total of 28 and 14 trials) respectively. Following a delay conditioning protocol during acquisition training, the CS+ co-terminated with an electro-tactile stimulation serving as US delivered 0.2s before CS+ offset which corresponds to a 100% reinforcement rate. During reinstatement, which followed the extinction training and preceded the reinstatement-test, three USs were delivered unannounced with 5s intervals between each US.

Two light gray fractals served as CSs (RGB [230, 230, 230]), 492*492 pixels), which were pseudo-randomly presented for 6 – 8s (mean: 7s) with no more than two identical stimuli in succession. CS presentations were interleaved by an intertrial-interval (ITI) during which a white fixation cross was displayed for 10 – 16s (mean: 13s). ITIs before and after the delivery of the reinstatement USs were 10s and 13s respectively.

The electro-tactile US was delivered via a Digitimer DS7A constant current stimulator (Welwyn Garden City, Hertfordshire, UK) as a train of three 2ms rectangular pulses interleaved by a 50ms interpulse interval. It was administered via a 1cm diameter platinum pin surface electrode which was fixated on the back of the right hand or more precisely, between the metacarpal bones of the middle and the index finger. Prior to the start of the experiment on day 1, the US was individually calibrated in a standardized stepwise procedure instructed by the experimenter. On a verbal scale ranging from zero (= stimulus was not unpleasant at all) to ten (= stimulus is the worst imaginable in the context of this study), it was aimed at an US aversiveness rating between 7 and 8 (unpleasant, but tolerable level). Participants were, however, not informed about this target level. The US amplitude was kept constant across experimental days within one measurement time point but was re-calibrated at T1.

Presentation Software (2010, Version 14.8, Neurobehavioral Systems, Inc., Albany California, USA) was used for stimulus presentation with all stimuli being displayed on a gray background (RGB [100, 100, 100]). In order to prevent renewal effects, the background color was kept constant across experimental phases including reinstatement-test and the ITI (Haaker et al., 2014). Allocation of CS types to CS+ and CS- and the presentation of the first CS type during acquisition and extinction training were counterbalanced across participants, but stimuli were identical for all participants.

2.4 Fear ratings and contingency awareness

Prior and subsequent to acquisition and extinction training as well as after the reinstatement-test, participants were asked to rate “how much stress, fear, and tension” they experienced when CS+ and CS- were last presented. These fear ratings of the CSs were obtained on a VAS with the poles zero (answer = none) and 100 (answer = maximum). After reinstatement-test, participants underwent two fear ratings with one rating referring retrospectively to the first presentations of the CSs directly after reinstatement and the other rating referring to the last presentation of each CS type during reinstatement-test. Prior to the experimental session, participants were familiarized with the rating procedure to ensure correct handling. Participants also had to rate the aversiveness of the US after acquisition training on day 1 and after reinstatement-test on day 2. Answers were given on the identical VAS as described for fear ratings. For analysis purposes, all ratings were transformed to a scale ranging from zero to 25. All ratings had to be confirmed via button press. Unconfirmed ratings were excluded from all analyses.

After the experimental procedures on day 1 and day 2, CS-US contingency awareness was assessed in a standardized post-experimental interview (adapted from Bechara et al., 1995). Based on this interview, participants were classified as aware, unaware, or uncertain of CS-US contingency.

2.5 Skin conductance response

During each phase of the experiment, SCRs were recorded continuously by a BIOPAC MP 100 amplifier (BIOPAC Systems, Inc., Goleta, California, USA) and Spike 2 software (Cambridge Electronic Design, Cambridge, UK). Data were converted from analog to digital using a CED2502-SA (Cambridge Electronic Design Limited, Cambridge, UK). Two self-adhesive hydrogel Ag/AgCl-sensor recording SCR electrodes (diameter = 55 mm) were fixed on the distal and proximal hypothenar of the left hand. A gain of 5 Ω and a 10 Hz lowpass filter were used. Data were acquired at 1000 Hz and afterwards down-sampled to 10 Hz.

Semi-manual scoring of the SCRs was performed using the custom-made software EDA View (developed by Prof. Dr. Matthias Gamer, University of Würzburg, Germany), which is based on SCR quantification using the trough-to-peak method. The settings of the program were specified as the trough emerging in the range of 0.9 to 3.5s after CS onset or 0.9 to 2.5 s after US onset (Boucsein et al., 2012; Sjouwerman & Lonsdorf, 2019) with the peak occurring within subsequent 5 s (maximum rise time; Boucsein et al., 2012).

Confounded SCRs (e.g. due to recording artifacts caused by electrode detachment) and SCRs exceeding the defined time window were classified as missing values and excluded from analyses. SCRs under 0.01 μ S occurring in the specified time window were classified as zero-responses. Non-responding on day 1 was defined as zero-responses in more than two-thirds of responses to the US (i.e., more than 9 out of 14). On day 2, participants who did not respond to any of the three reinstatement USs were classified as non-responders. SCR data were preprocessed for the following analyses with MATLAB (2016, Mathworks, Natick, Massachusetts, USA) version R2016b. In **Study I** and **Study III**, raw SCRs were log-transformed ($\log_{10}[1 + \text{raw amplitude}]$) and range corrected to account for inter-individual variability by dividing each SCR by the maximum SCR per participant and day (Lykken & Venables, 1971), whereas in **Study IV**, SCRs were solely log-transformed. In **Study II**, different SCR data transformation types were implemented (i.e., none, log-transformation, log-transformation and range correction).

2.6 Brain imaging

Brain imaging data were recorded during experimental sessions at T0 and T1 (functional and structural data) as well as at T5 (structural data only). Aside from the fear conditioning task, participants underwent a resting state scan (T0 and T1) and a face-matching task (T0, T1, and T5; Hariri et al., 2002). However, these tasks were not part of this thesis project and will hence not be discussed in detail.

2.6.1 Study I: MRI data acquisition and analysis

On day 2, T1-weighted structural images ($1 \times 1 \times 1$ mm) were recorded with a 3T PRISMA whole body scanner (Siemens Medical Solutions, Erlangen, Germany) by using magnetization prepared rapid gradient echo (MPRAGE) sequence (TR = 2300 ms, TE = 2.98 ms, field of view = 192×256 mm, 240 slices, slice thickness: 1 mm) and a 64-channel head coil.

The brain imaging software Freesurfer 6.0.1 (<https://surfer.nmr.mgh.harvard.edu/>) was used to reconstruct the volume of subcortical brain regions and cortical thickness. Regions of interest (ROIs) in **Study I** correspond to those regions as defined in Freesurfer (for a visualization, see <https://surfer.nmr.mgh.harvard.edu/>). The subcortical or volume-based stream included an initial Talairach registration, initial volumetric labeling, bias field correction, nonlinear volumetric atlas registration, and volumetric labeling of subcortical structures (Fischl et al., 2002). Likewise, the surface-based stream obtaining quantifications of cortical thickness encompassed an initial Talairach registration, then bias field correction, skull stripping, white matter classification, surface generation, and gyral labeling (Dale et al., 1999). Parcellation of the cortex was accomplished according to Freesurfer's Desikan-Killiany cortical atlas (Desikan et al., 2006).

2.6.2 Study II: fMRI data acquisition and analysis

For functional data acquisition, the identical 3 Tesla PRISMA whole body scanner (Siemens Medical Solutions, Erlangen, Germany), a 64-channel head coil, and an echo planar imaging (EPI) sequence (repetition time (TR): 1980 ms, echo time (TE): 30 ms, number of slices: 54, slice thickness: 1.7 mm (1 mm gap), field of view = 132×132 mm) were used. For analyses of functional data, SPM12 (Wellcome Department of Neuroimaging, London, United Kingdom), and MATLAB (2019, Mathworks, Natick, Massachusetts, USA) were used. Functional data were preprocessed by involving realignment, coregistration, normalization to a group-specific DARTEL template, and smoothing (6 mm full width at half maximum, FWHM).

Since participants could not have learned the CS-US association during the first CS trials due to delay conditioning (i.e., the US occurring subsequent to the CS+), separate regressors for the first CS+ and CS- trials and the subsequent trials were implemented for first-level acquisition training analyses. Motion parameters, habituation trials, US, and rating onset were integrated as nuisance regressors. Similarly, extinction could not have occurred during the first two CS trials of extinction training. Thus, separate regressors of interest were defined for the first CS+ and CS- trial and all subsequent extinction training trials. Motion parameters, US presentation, and fear ratings served as regressors of no interest. Different statistical analyses were conducted on first-level models only (see section 2.9). Thus, no second-level analysis was carried out.

The eleven ROIs encompassed the amygdala, caudate nucleus, dACC, dlPFC, hippocampus, bilateral anterior insula, nucleus accumbens (NAcc), pallidum, putamen, thalamus, and vmPFC. Amygdala, hippocampus, caudate nucleus, pallidum, putamen, NAcc, and thalamus anatomical masks were applied based on the Harvard-Oxford atlas (Desikan et al., 2006) by using a 0.5 maximum probability threshold. The anterior insula was designated to the alignment of a box of size 60 x 30 x 60 mm centered around MNIxyz = 0, 30, 0 extracted from anatomical subdivisions (Nieuwenhuys, 2012) and the 0.5 thresholded anatomical mask from the Harvard Oxford atlas.

The cortical ROIs dACC and dlPFC were defined by creating a box of size 20 x 16 x 16 mm around peak voxels as derived from a meta-analysis (Fullana et al., 2016). For the dACC, the x coordinate was set to 0: dACC: MNIxyz = 0, 18, 42; left dlPFC: MNIxyz = -36, 44, 22, right dlPFC: MNIxyz = 34, 44, 32 (Fullana et al., 2016). As described in previous work (Lonsdorf et al., 2014), the cortical vmPFC was defined by building a 20 x 16 x 16 mm box based on peak coordinates applied in prior fear learning studies (vmPFC: MNIxyz = 0, 40, -12, e.g., Kalisch et al., 2006; Milad, Wright, et al., 2007). The x coordinate was set to 0 to symmetrize masks along the midline.

2.7 Further outcome measures of no interest

2.7.1 Questionnaires

Participants filled in the state scale of the State-Trait Inventory (STAI-S; Spielberger et al., 1983) as paper-pencil version on all days involving an experimental session (day 1 and day 2 of T0 and T1 as well as on T5). At T0 and T1, participants also completed the paper-pencil version of the Pittsburgh Sleep Quality Index (PSQI; Buysse et al., 1989). Additionally, at T0, T1 and T2, participants completed a battery of the following questionnaires, always presented in the identical

order computerized via LimeSurvey including the Kurzer Fragebogen zur Erfassung von Belastungen (KFB; Flor, 1991), the German versions of the Social Support Appraisal Scale (SS-A-d; Laireiter, 1993), the Trier Inventory for Chronic Stress (TICS; Schulz et al., 2004), Stressverarbeitungsfragebogen (SVF 78; Janke & Erdmann, 2002), Berliner Social-Support Scales (BSSS; Schwarzer & Schulz, 2003), brief Coping Orientation to Problems Experienced Inventory (Brief COPE; Carver, 1997), List of Threatening Experiences (LTE, modified version; Brugha et al., 1985), Cognitive Emotion Regulation Questionnaire (CERQ short; Garnefski & Kraaij, 2006), Generalized Self-Efficacy Scale (GSE; Schwarzer & Jerusalem, 1995), Life Events Checklist (LEC, modified version; Canli et al., 2006; Caspi et al., 1996), Perceived Stress Questionnaire (PSQ; Fliege et al., 2009) and self-constructed questions addressing cortisol. At T3 and T4, the questionnaire battery was complemented by the trait scale of the STAI (STAI-T; Spielberger et al., 1983) and the Beck Depression Inventory 2nd Edition (BDI-II; Beck et al., 1996). At T5, the participants filled in the following questionnaires: BSSS, LTE, CERQ short, GSE, LEC, PCQ, STAI-T, BDI-II, Intolerance of Uncertainty Scale (IUS; Freeston et al., 1994), NEO-Five-Factor-Inventory neuroticism (NEO-FFI neuroticism; Kanning, 2009) and the PROMIS-scales (Patient-Reported Outcomes Measurement Information System; <https://www.healthmeasures.net/explore-measurement-systems/promis>). For approximately half of the subjects, the last T5 appointment fell during the lockdown due to the COVID-19 pandemic. These participants additionally completed a self-constructed questionnaire about their experiences during the lockdown. At T0, T1, and T5, participants used the computer in the laboratory, whereas at T2 – T4 participants completed the questionnaires remotely.

2.7.2 Other physiological outcomes

Further physiological outcome measures that were obtained during the experimental days at T0 and T1 were saliva samples, hair samples, and blood samples. During all experimental sessions within the MR scanner, respiration, pulse, and eye tracking data were recorded (EyeLink 1000 device, SR Research Ltd., Mississauga, Ontario, Canada). To track sleep overnight, participants were equipped with wrist-activity monitors (Actiwatch; Philips Respironics, 2009) after the experiment on day 1 which they returned on day 2.

2.8 Additional complementary data sets

Two additional data sets were used to complement the data set on which this thesis is primarily based. In **Study III**, the additional data set was used as a case example to illustrate the association between trait anxiety and CS discrimination observed in this data set, and the consequential sample bias that results from performance-based exclusion of participants. In **Study IV**, the additional data set was used for the same purpose as the main data set: to illustrate the extent to which the application of multiple statistical models impacts on the results and to demonstrate the application of the new R package *multifear* in more than one data set.

2.8.1 Additional data set used in Study III

In **Study III**, the additional data set which included 268 participants (female_N = 195, male_N = 73, age_M = 25, age_{SD} = 4) originates from a study investigating the impact of individual emotional negativity, as assessed by various questionnaires, on CS discrimination across several outcome measures (Sjouwerman et al., 2020). All participants provided written informed consent. The study was conducted according to the Declaration of Helsinki and the ethical approval was granted by the Ethical Review Board of the German Psychological Association (DGPS).

Participants completed a battery of questionnaires, including the STAI prior to the experiment. The conducted fear conditioning paradigm included acquisition and extinction training as well as a return of fear phase. However, exclusively data from acquisition training were analyzed in **Study III**. CSs were a black colored rectangle and ellipse presented via Presentation Software (Version 14.8, Neurobehavioral Systems, Inc, Albany California, USA) on a yellow, green, blue, or purple computer screen for 6 s and 9 times each with not more than two identical stimuli in a row. Background colors served as context, but had no further meaning for the acquisition training. The reinforcement rate was 100%. Allocation of shapes to CS+ and CS-, order appearance of CS+/CS- and background color were counterbalanced across participants. A white fixation cross presented for 11.5 s (± 1.5 s) on a black background served as ITI.

The constitution and apparatus of the US/US delivery and its calibration correspond to the description in section 2.3. The apparatus and procedure to acquire and preprocess SCR data as well as the criteria of non-responding were identical to those described in section 2.5 except that the scoring window ranged from 0.9 to 4.0 s (Boucsein et al., 2012) and SCRs $< 0.02 \mu\text{S}$ were classified as zero responses. SCRs were log-transformed and range corrected to approach a normal distribution.

2.8.2 Additional data set used in Study IV

The additional data set used in **Study IV** originated from a study in which the impact of dopamine-induced prefrontal reactivations on extinction learning was investigated (Gerlicher et al., 2018). Participants (N = 40 male subjects, $age_M = 28.1$ years, $age_{SD} = 2.7$ years) provided written informed consent and the local ethics committee had approved the experimental protocol (Ethics Committee of the State Medical Association, Rheinland-Pfalz, Germany).

A black square and a black rhombus were shown for 4.5 s as CSs in front of two different contexts A and B (living room or kitchen) by using Presentation Software (Version 14.8, Neurobehavioral Systems, Inc, Albany California, USA). Allocation of geometric shapes to CS+ and CS- and contexts were randomized across participants. Three square-wave pulses of 2 ms (50 ms interstimulus interval; generated by a DS7A Digitimer, Weybridge) served as electro-tactile US which occurred simultaneously with the CS+ and was delivered through a surface electrode with a platinum pin on the back of the right hand. The US was calibrated individually aiming at a painful, but tolerable level. CS presentations were interleaved by 17 – 19 s ITIs (mean = 18.5 s). CSs were pseudo-randomly presented with not more than two trials in succession.

The paradigm was conducted in MR environment and encompassed three days including fear acquisition training on day 1 (presentation of ten CS+ and CS- each in context A, reinforcement rate = 50%), extinction training (presentation of 15 CS+ and CS- each in context B) and a following drug administration on day 2 as well as a drug manipulation test on day 3.

SCRs were captured via self-adhesive Ag/AgCl electrodes from the thenar and hypothenar of the non-dominant hand. Data were obtained by using the BIOPAC MP150 with EDA100C. An offline low-pass filter was applied (second-order Butterworth filter, cut-off frequency: 1 Hz) by using MATLAB (Mathworks, Natick, Massachusetts, USA). Mostly, the response quantification followed that of the main data set except that the identification of the response trough was determined 0.9 – 4 s after stimulus onset (Boucsein et al., 2012) and the usage of a 0.02 μ S minimum amplitude criterion. Two participants were excluded due to recording artifacts during acquisition and extinction training resulting in a final sample of N = 38 participants for both days.

2.9 Statistical analyses

This paragraph only covers the primary analyses. For more comprehensive information, please consult the published studies. To test associations of cortical thickness (dACC, insula, and mODF) or subcortical volume (amygdala) and differential SCRs (averaged across phases) and fear

ratings (operationalized as post-acquisition and pre- minus post-extinction) in **Study I**, simple linear regressions were calculated by using the T0 data. Moreover, Bayes factors were calculated for all analyses to complement the traditional null hypothesis significance testing (NHST). For reasons of robustness, the last two analyses were repeated with the inclusion of solely the first or second half of phases (SCR) or ratings prior to/after phases, as well as the inclusion of sex and total intracranial volume (TIV) as covariates, and with the exclusion of outliers (± 3 SDs above/below the mean). To assess the moderating role of CS-US contingency awareness in the putative association of dACC thickness, exploratory moderated regression analyses were performed. Partial correlations were computed to investigate the relationship between amygdala volume and trait as well as state anxiety.

To calculate reliability as well as predictability across experimental phases in **Study II**, SCRs, fear ratings, and BOLD fMRI at time points T0 and T1 were taken into account: At both time points, internal consistency was calculated by applying the odd-even approach. To determine longitudinal reliability at the individual level, intra-class correlation coefficients (ICCs) were calculated. Moreover, within-subject similarity (i.e., the correlation of responses within each individual across T0 and T1) was compared to between-subject similarity (i.e., the averaged correlation of responses of one individual at T0 and responses of all other individuals at T1), and the overlap of individual significant voxels as measured with BOLD fMRI at T0 and T1 were examined. For investigations of longitudinal reliability at the group level, the overlap of group averaged significant voxels and the explained variance in SCRs at T1 by the variance of T0 (i.e., R^2) were investigated for BOLD fMRI and for SCRs respectively. To assess if responding in a given experimental phase can be predicted by responding in a preceding experimental phase, several simple linear regressions were calculated with the given phase as the dependent variable and the preceding phase as the independent variable. All reliability measures were calculated involving different data specifications encompassing stimulus type (CS+, CS-, CS discrimination, US), different operationalizations of experimental phases (e.g., average acquisition training, last two trials of acquisition training) and different data transformations such as ranking or log-transformation and range correction of the data (exclusively SCR data).

Study III aimed to investigate the impact of methodological heterogeneity in the definition and exclusion of ‘non-learners’ and ‘non-responders’. To extract criteria of ‘non-learners’ and ‘non-responders’ based on SCR performance, a systematic literature search was conducted following the PRISMA guidelines (Moher et al., 2009) including all publications within a 6 months period prior to the beginning of the study.

Exclusion rates due to ‘non-learning’ and ‘non-responding’ were extracted from the literature. To statistically compare CS discrimination between groups of the sample that were built on the basis of different exclusion ‘non-learner’ criteria in the literature (here referred to as ‘exclusion groups’), a mixed ANOVA with CS discrimination in SCRs and fear ratings as dependent variables and the between-subjects factor ‘Exclusion group’ as well as the within-subject factor ‘CS type’ was conducted. For this analysis, the main data set 1 (i.e., T0 data) was taken into account. It is important to note that this calculation involves a degree of circularity, as the exclusion groups were based on CS discrimination abilities. However, it is even more important to check whether there was in fact no CS discrimination according to the criteria often used in the literature. Post hoc t-tests were performed to investigate which exclusion groups differed from each other. To examine how the exclusion of ‘non-learners’ impacts the participant pre-selection for specific individual differences such as trait anxiety, a univariate ANOVA including STAI-T score serving as dependent and ‘Exclusion group’ as independent variable was performed by using the complementary data set 2. To compare STAI-T scores in different exclusion groups, post hoc pairwise t-tests were computed. To assess the distribution of ‘non-responding’ across stimuli, percentages of ‘non-responses’ were extracted for CS+, CS-, US, and CS discrimination. Furthermore, Spearman rank correlations were calculated to examine the relationship between the non-responses to the US and the non-responses to the CSs.

To investigate the heterogeneity in the statistical model selection in the field of fear conditioning, the first step in **Study IV** was to extract relevant statistical models and data reduction approaches (i.e., processing of the data prior to inclusion in the statistical model, such as averaging all trials or specific blocks of trials of an experimental phase) typically applied in the field. Therefore, a systematic literature search was conducted according to the PRISMA guidelines (Moher et al., 2009) covering a time period of six months prior to the beginning of the project. For calculations of the multiverse analyses, the R package *multifear* was used, which was also introduced in **Study IV**. Mean and median p-values were calculated for NHST analyses (significance level $\alpha = .05$) while Bayes factors were computed for Bayesian analyses. Forest plots were created to illustrate the effect sizes of all the different frequentist ANOVAs and t-tests including the different types of data reduction approaches.

All statistical analyses were conducted using different software versions of R (R Core Team) and MATLAB (Mathworks, Natick, Massachusetts, USA).

3 Results and brief discussions

3.1 Study I: Revisiting potential associations between brain morphology, fear acquisition, and extinction through new data and a literature review

The ongoing challenges in obtaining robust and replicable results (Open Science Collaboration, 2015; Stroebe & Strack, 2014) underscores the necessity for additional attempts to replicate empirical findings – also in the field of fear conditioning. In line with this, recent findings have challenged the robustness of structural-brain-behavior associations, whose investigation have a long-standing tradition in psychological and neuroscientific research, with low replication rates observed across various psychological measures (Boekel et al., 2015; Genon et al., 2017; Kharabian Masouleh et al., 2019). Research into how individual differences in defensive responding correspond to variations in brain structure is limited involving few studies investigating fear acquisition and extinction learning, as well as retention test, with inconsistent results reported (Cacciaglia et al., 2015; Hartley et al., 2011; Milad, Quirk, et al., 2007; Pohlack et al., 2012; Winkelmann et al., 2016). Most studies have small sample sizes but report surprisingly strong correlations (see **Figure 3**). Since the robustness of structural brain-behavior associations has recently been called into question (Kharabian Masouleh et al., 2019; Masouleh et al., 2020), **Study I** attempted to conceptually replicate previous findings of associations between cortical thickness/subcortical volume and defensive responding in SCRs and fear ratings, both in a large sample and within a single study.

Based on previously reported associations in the literature, cortical thickness of the dACC and insula as well as the amygdala volume was assessed during acquisition training, and for extinction training, mOFC thickness, and amygdala volume were included in the analyses. Exploratory analyses focused on the moderating role of CS-US contingency awareness regarding putative associations of dACC thickness and differential responding and on the replication attempt of previously reported associations of amygdala volume and self-reported trait as well as state anxiety.

Neither NHST nor Bayesian results provided evidence for an association of differential responding and cortical thickness or subcortical volume. In fact, there was evidence to the contrary: Bayes factors yielded mostly moderate to strong relative evidence in favor of the null hypotheses indicating that such relationships are likely absent. This was true for SCRs as well as fear rating during acquisition (see **Figure 4**) and extinction training and for all brain regions under

investigation. Additional robustness analyses including solely the first or second half (SCR) or the pre/post ratings instead of the complete phases and gender or TIV as covariates did not yield substantially different results. Furthermore, CS-US contingency awareness did not have a moderating role in the putative association of dACC thickness and differential responding, and no significant relationships between amygdala volume and trait as well as state anxiety were observed.

In sum, previous findings on the association between individual differences in brain morphology and differential physiological or subjective conditioned responding during acquisition and extinction could not be replicated (Abend et al., 2020; Cacciaglia et al., 2015; Hartley et al., 2011; Milad, Quirk, et al., 2007; Rauch et al., 2005; Winkelmann et al., 2016). However, it should be highlighted that **Study I** represents a conceptual replication attempt with procedural differences in comparison with previous work such as i) the application of a partial instead of 100% reinforcement rate, ii) the number of trials included, iii) the implementation of immediate vs. 24-h-delayed extinction training, or iv) differences in data analyses, such as the usage of different software tools. These procedural and methodological differences might have contributed to these divergent results and may represent boundary conditions under which a significant effect does not occur. This would imply that the generalizability of results may be limited.

Yet, there might be another often overlooked explanation for the lack of associations observed in **Study I**: While the reliability of structural MRI was reported to be excellent (Elliott et al., 2020; Han et al., 2006), the reliability of SCRs and fear ratings is alarmingly understudied (see next section on **Study II**). Most importantly, reliability puts an upper bound to correlations as the correlation between two variables cannot exceed the correlation within these two variables (i.e., the reliability; Spearman, 1910). Thus, the absence of a significant relationship between brain morphology and conditioned responding might be also due to unreliable physiological and behavioral measures (i.e., SCR and fear ratings). Yet, little is known about the reliability of measures commonly used in fear conditioning research. To fill this gap, measurement reliability was investigated in **Study II**.

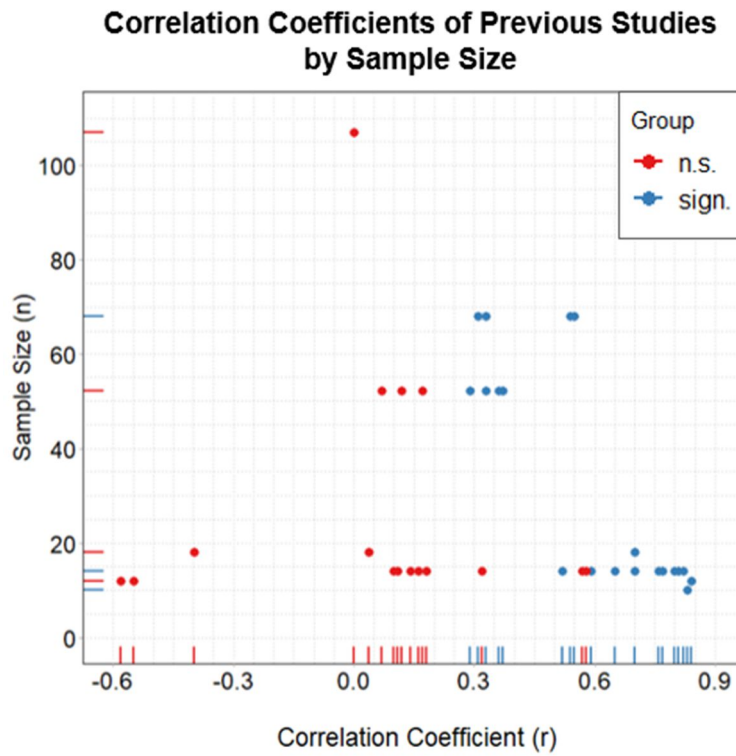


Figure 3 Illustration of effect sizes as reported in previous and current work as a function of sample size. Note that while some previous studies used regressions instead of correlations, all effect size measures were converted to correlation coefficients for comparability. Blue dots represent significant results while red dots represent non-significant ones. Note also that some studies report multiple associations, and thus are depicted by multiple dots.

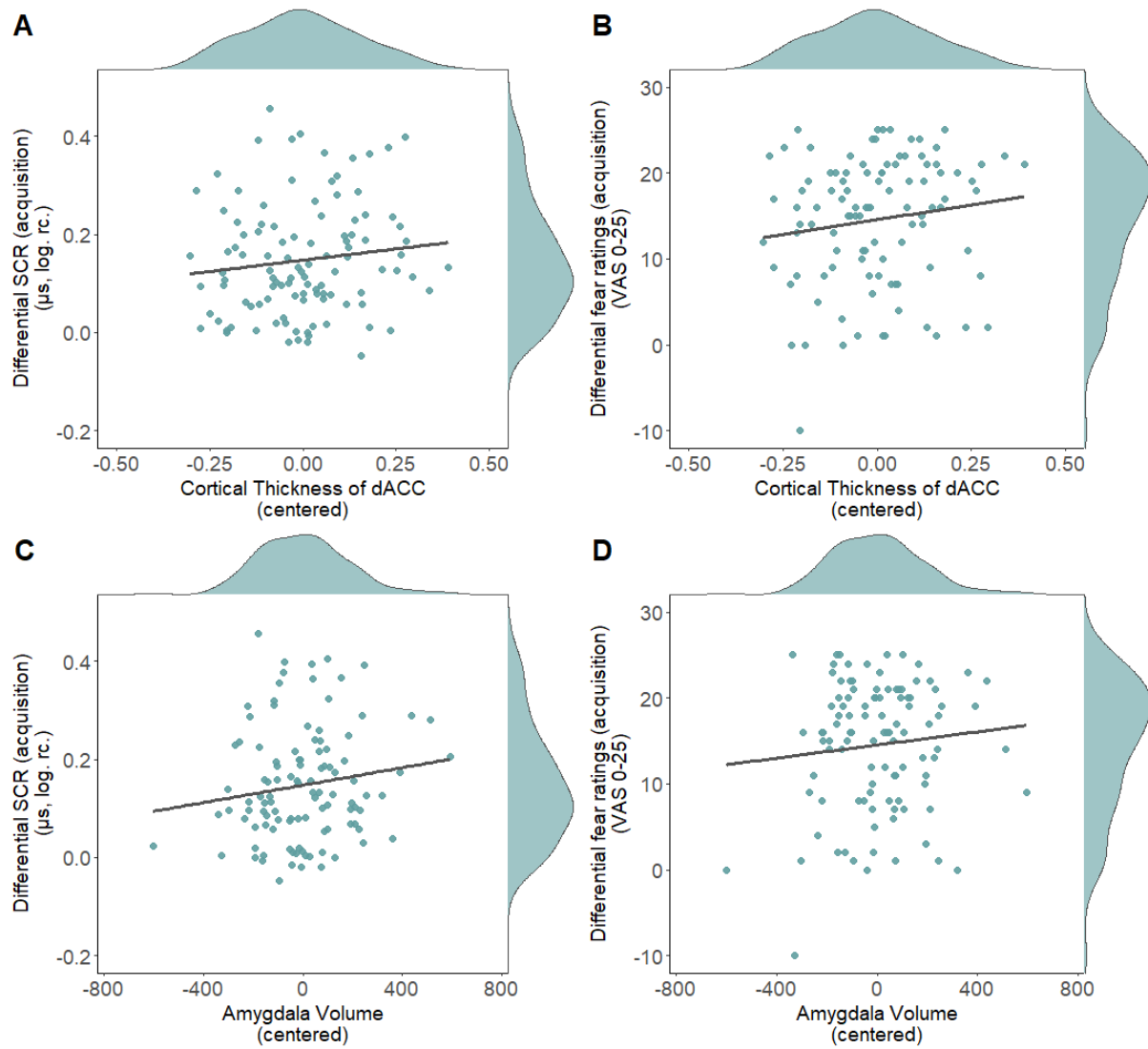


Figure 4 Illustration of the (absence of) a relationship of cortical thickness of the dACC (A and B) and amygdala volume (C and D) with CS discrimination $[(CS+) - (CS-)]$ in SCRs (A and C) and fear ratings (B and D) during acquisition training across participants. The curves above and to the right of the scatterplots illustrate marginal densities.

3.2 Study II: Robust group- but limited individual-level (longitudinal) reliability and insights into cross-phases response prediction of conditioned fear

Whereas the scientific focus in the past has been primarily on basic, generic mechanisms underlying conditioned responding and drawing conclusions about different groups, there is currently a shift in emphasis toward individual difference research (Lonsdorf & Merz, 2017). To answer important clinical questions, such as why some individuals respond to treatments while others do not, or why some individuals relapse, measures that model these processes in the laboratory are required that allow prediction at an individual level beyond the group mean (Fröhner et al., 2019; Hedge et al., 2018; Lonsdorf & Merz, 2017). An important prerequisite for this is that the employed measures are reliable. However, the reliability of conditioned responding in fear conditioning paradigms is heavily understudied with only five studies addressing this topic (Cooper, Dunsmoor, et al., 2022; Fredrikson et al., 1993; Ridderbusch et al., 2021; Torrents-Rodas et al., 2014; Zeidan et al., 2012) and hence requires more attention.

Similarly, there is also a lack of empirical investigations into the relationship of conditioned responding at different time points of very short intervals, i.e. across different experimental phases. This relationship is often implicitly assumed by researchers who ‘control’ responding in later experimental phases for responding in earlier phases (e.g., see Milad et al., 2009, critically discussed in Lonsdorf et al., 2019). However, these associations of conditioned responding are rarely directly investigated and lack consistency in findings: Evidence from animal studies or indirectly from patient research supports both the presence (Foa et al., 1983; Gershman & Hartley, 2015; Rauch et al., 2004) and absence (Bouton et al., 2006; Kozak et al., 1988; Pitman et al., 1996; Plendl & Wotjak, 2010; Prenoveau et al., 2013; Riley et al., 1995; Shumake et al., 2014) of these associations.

Study II aimed to fill these gaps in the literature by investigating i) longitudinal reliability of conditioned responding at the individual and the group level, ii) internal consistency, and iii) the association of conditioned responding across different experimental phases. To account for methodological heterogeneity in the literature, we followed a manyverse-inspired approach and included different outcome measures (i.e., BOLD fMRI, SCR, and fear ratings) as well as data specifications such as phase operationalizations (e.g., responding averaged across the phase or the last two trials), data transformations (e.g., log-transformation and range correction of the data), different stimulus types (i.e., CS+, CS-, US, and CS discrimination), and reliability measures (i.e., ICCs, similarity, or overlap).

Across most data specifications, internal consistency of SCRs (see **Figure 5A-B**) as well as longitudinal reliability at the group level for SCRs (see **Figure 6**) and BOLD fMRI were robust whereas individual-level longitudinal reliability of SCRs, fear ratings (see **Figure 5C-F**) and BOLD fMRI was somewhat limited. The latter was evident in more traditional approaches such as ICCs, but also in more advanced reliability measures such as overlap or similarity (see **Figure 7**) with the exception of individual BOLD activation patterns at T0 being more similar to their own activation patterns at T1 than to those of others. The limited reliability was particularly true for responding during extinction training and was apparent across different data specifications such as ranking or log-transformation and range correction of the data (SCR only) of the data. Adding more trials did not significantly improve reliability. Taken together, these findings suggest that the individual-level longitudinal reliability remains relatively consistent, regardless of changes in data transformations or paradigm specifications. This simplifies the integration of prior research that employed different time intervals, reliability metrics, and paradigms.

Regarding predictions over relatively brief time intervals, significant associations between responses across different experimental phases were observed for SCR, fear ratings, and BOLD fMRI. Higher responses in previous phases were generally linked with slightly higher responses in subsequent phases for all outcome measures. The strength of predictions, however, depended on data specifications. More precisely, several predictions were not significant, specifically for CS discrimination in SCRs and BOLD fMRI. This may be attributed to less reliable difference scores (i.e., CS+ minus CS-; Infantolino et al., 2018; Lynam et al., 2006), where meaningful variance is subtracted (Moriarity & Alloy, 2021).

In sum, predictions at the individual level might be more feasible for shorter time intervals while at the group level, predictions may be also plausible for longer time intervals. However, the benchmarks used in this study originate from psychometric research and should be interpreted with caution as the definition of “good” reliability in an experimental context is still unclear (Parsons et al., 2019). Nevertheless, these findings prompt the inquiry of whether conditioned responses are adequate for making long-term individual predictions and what steps could be taken to obtain more robust results.

A promising solution for the limited individual long-term predictions might be the application of homogenous (latent) subgroups that share a certain similarity in aspects such as rapid or slow response profiles (Galatzer-Levy et al., 2013). This approach exploits the advantage of robust group reliability but also allows conclusions to be drawn about individual differences. Furthermore, reliability might be enhanced by utilizing more sophisticated approaches such as intra-

individual neural response variability (Månsson et al., 2021) or multivariate imaging techniques, which have been proposed to be more reliable than traditional analysis methods (Kragel et al., 2021; Marek et al., 2020; Noble et al., 2021; Visser et al., 2021).

Measures that lack reliability represent a challenge for the robustness and replicability of results as reliability ensures that the employed measures generate stable data, i.e. producing the same results for the same participants over time (Heale & Twycross, 2015). However, we need to deepen our understanding of which methodological choices lead to more robust results. One such choice involves the exclusion of specific participants, which can have major consequences for the generalizability and robustness of results, as demonstrated in **Study III**.

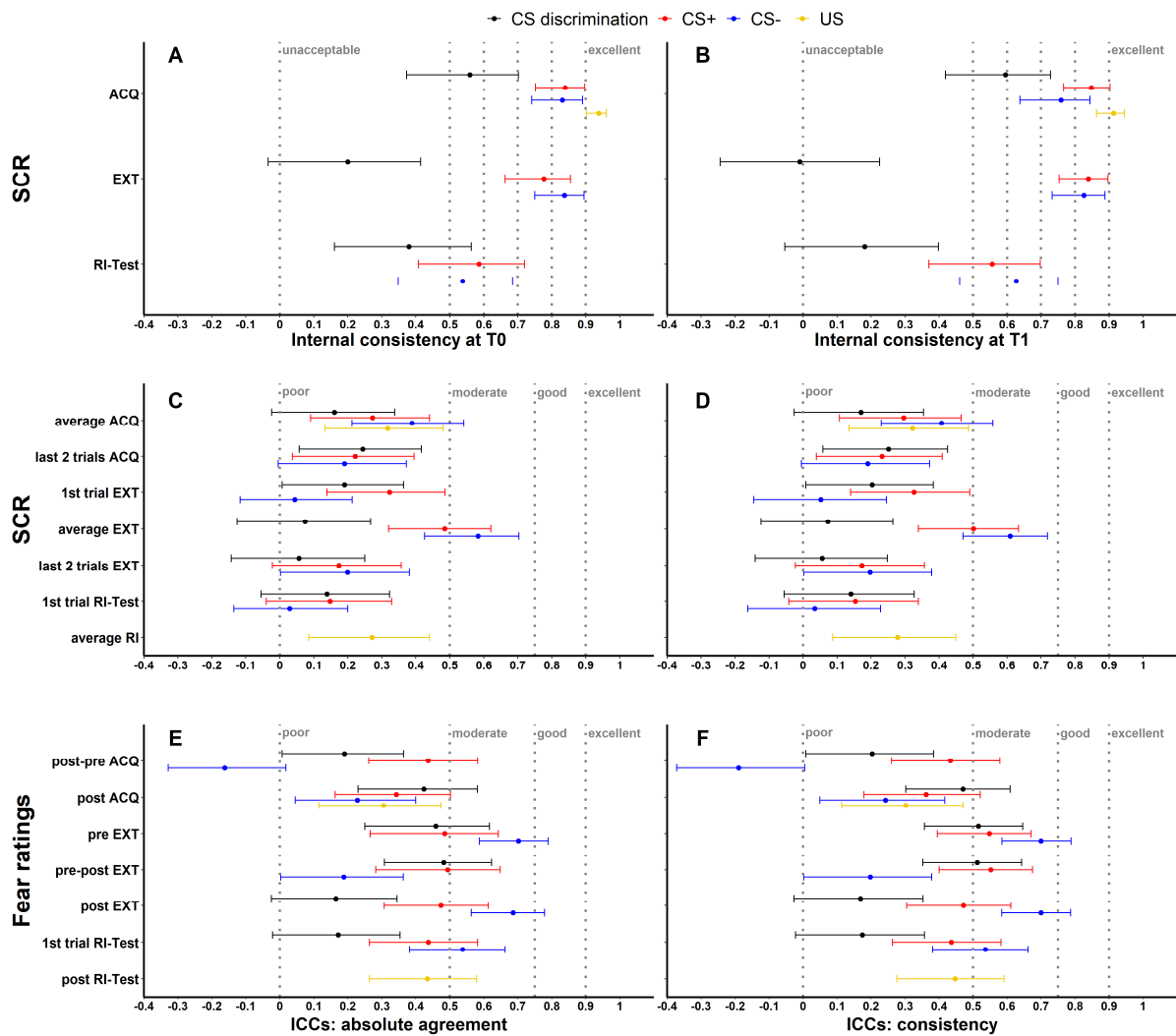


Figure 5 Demonstration of internal consistency (at T0: A and at T1: B), absolute agreement ICCs (C and E) as well as consistency ICCs (D and F) for SCRs (A, B, C, and E) and fear ratings (D and F). Internal consistency mirrors the reliability within one time point, whereas ICCs mirror the reliability across time points. Colors indicate different stimulus types (red = CS+, blue = CS-, yellow = US) or CS discrimination (black). The y-axes comprise different (operationalizations of)

experimental phases. Error bars represent 95% confidence intervals which indicate significance when zero is not included. For internal consistency, benchmarks are depicted as determined by Kline (2013) with cut-offs for unacceptable (<0.5), poor (>0.5 but <0.6), questionable (>0.6 but <0.7), acceptable (>0.7 but <0.8), good (>0.8 but <0.9), and excellent (≥ 0.9). Accordingly, the benchmarks for ICCs were poor (<0.5), moderate (>0.5 but <0.75), good (>0.75 but <0.9), and excellent (≥ 0.9) as stipulated by Koo and Li (2016). It should be noted that these benchmarks origin from psychometric work on questionnaires and thus, completely different benchmarks could be applicable for fear ratings and especially for SCRs. Thus, these benchmarks rather serve as rough guidance and should not be overinterpreted for experimental data which are inherently more noisy (Parsons, 2020).

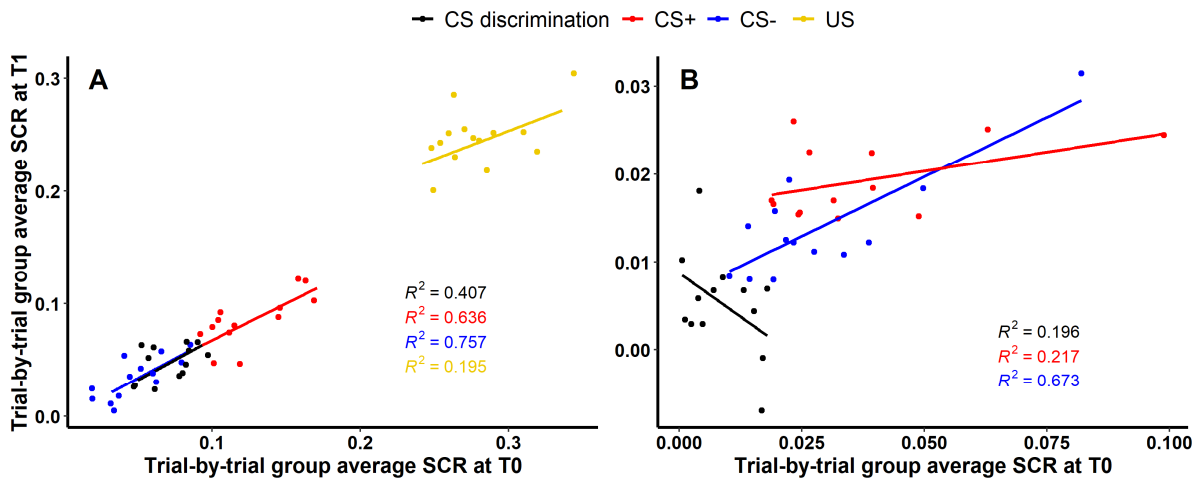


Figure 6 Illustration of the longitudinal reliability at the group level of SCRs during acquisition (A) and extinction training (B) color coded for stimulus type (red = CS+, blue = CS-, yellow = US) and CS discrimination (black). Longitudinal reliability at the group level of SCRs was determined by R squared as extracted from simple linear regression including group average trials-by-trial SCRs at T0 as predictor and group average trial-by-trial SCRs at T1 as criterion.

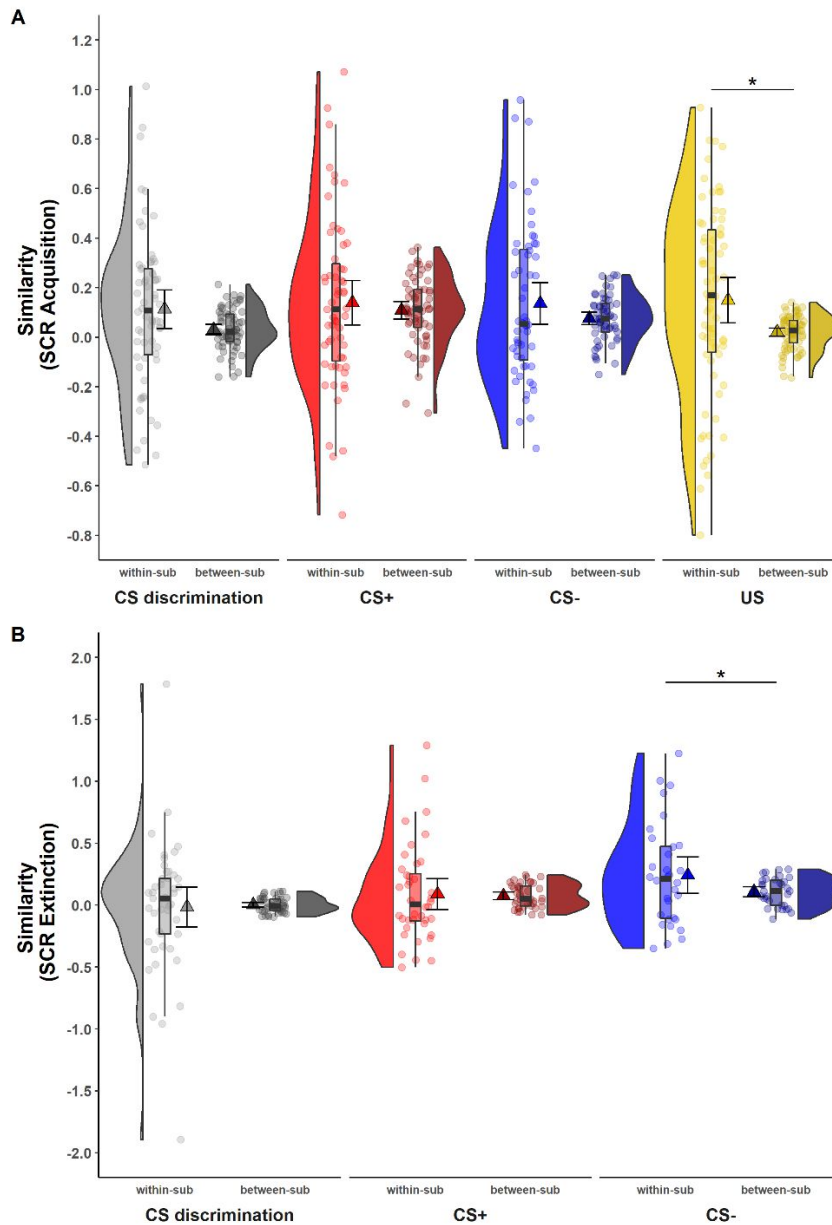


Figure 7 Comparison of within- and between-subject similarity for raw SCRs during (A) acquisition and (B) extinction training separately for CS discrimination (gray), CS+ (red), CS- (blue), and US (yellow). Each data point represents the correlation between trial-by-trial SCRs of each participant at T0 and T1. The within-subject similarity is the correlation between the same subject's SCRs at T0 and T1, while the between-subject similarity is the average correlation between one subject's SCRs at T0 and all other subjects' SCRs at T1. The triangles represent the mean correlations, and the error bars indicate the 95% confidence intervals. The boxplots show the distribution of the data, with the median represented by a bold line and the interquartile range (IQR) indicated by the box. The whiskers extend to the minimum and maximum values within the range of 25th/75th percentiles ± 1.5 IQR. The densities next to the boxplots represent the data distributions. One data point was excluded from the figure because its similarity was above 3.5 (within-subject similarity for the CS+). The variance differs considerably between within- and between-subject similarity because between-subject similarity is based on correlations averaged across participants, while within-subject similarity is based on non-averaged correlations calculated for each participant. within-sub = within-subject; between-sub = between-subject. * $p < 0.05$.

3.3 Study III: Navigating the garden of forking paths for data exclusions in fear conditioning research

In fear conditioning paradigms, a possible option for participant exclusion pertains to how to handle individuals who are labeled as SCR 'non-learners' or 'non-responders', which refer to participants who are classified as physiologically showing no or “insufficient” CS discrimination (i.e., ‘non-learners’) or stimulus-driven responses (i.e., ‘non-responders’). Although they are often routinely excluded due to the common belief that investigations within fear conditioning paradigms require a strong conditioned response, consistent definitions of ‘non-learners’ and ‘non-responders’ are lacking. Related challenges also involve their exclusion often being based on a single outcome measure (i.e., SCRs), while excluding them from analyses of *all* outcome measures. Moreover, CS discrimination – as an important criterion for ‘non-learning’ – appears to vary as a function of individual differences such as trait anxiety (e.g., Gazendam et al., 2013; Sjouwerman et al., 2020; Staples-Bradley et al., 2018), and thus, the exclusion of ‘non-learning’ subgroups might result in a substantial sample bias.

To address these theoretical concerns empirically in two data sets, the aims of **Study III** were to i) identify ‘non-learner’/‘non-responder’ criteria based on a systematic literature search, ii) showcase the impact of these criteria on results and their interpretation and iii) illustrate which empirical recommendations can be derived.

The systematic literature search revealed substantial heterogeneity in exclusion criteria of ‘non-learners’ with exclusion rates ranging from 2% up to 74% (see **Figure 8A**). Definitions of what constitutes a ‘non-learner’ involved variables such as the number of included trials or a varying pre-specified CS discrimination cutoff (see **Figure 8B**). The cutoffs at which participants were excluded for ‘non-learning’ ranged from $< 0 \mu\text{S}$ to $< 0.1 \mu\text{S}$ with the majority of participants being excluded when SCRs to the CS- were equal to or greater than responses to the CS+ (i.e., cutoff ≤ 0).

Moreover, we illustrate the different portions of the sample in data set 1 which would be excluded according to these different cutoffs (see **Figure 9**) including two case examples (ID#1 and ID#2), which differ in their SCR magnitude of CS discrimination, but – most importantly – exhibit the identical CS discrimination ratio of .04. Thus, cutoffs that require a strong CS discrimination (e.g., < 0.1) might disregard individuals who show weak SCR amplitudes, but similar CS discrimination ratios. Of note, most of the exclusion groups as defined by different CS discrimination cutoffs showed indeed significant CS discrimination in SCRs when tested

statistically, and – even more strikingly – *all* exclusion groups discriminated in fear ratings, although they would have been excluded as non-learners based on SCRs.

The examination of how the exclusion of ‘non-learners’ affects the pre-selection of participants for particular individual differences, such as trait anxiety, uncovered that individuals with higher levels of trait anxiety exhibited poorer CS discrimination. Consequently, these individuals are more prone to exclusion as a function of increasing cut-offs (data set 2). This is of great relevance as i) this procedure may result in limited generalizability of findings due to a generally restricted sample and ii) response patterns of anxious individuals such as deficient safety learning appear to be similar to those of anxiety patients (Duits et al., 2015; Gazendam et al., 2013). Hence, their exclusion might be a threat to the translation of empirical findings into clinical applications.

The definitions of ‘non-responding’ (i.e., lacking physiological responses), which varied substantially according to the criteria of i) stimulus type (i.e., CS+, CS-, or US), ii) minimum SCR amplitude and iii) the percentage of trials which met these criteria, led to exclusions of 0% to 14% of participants in the systemically identified studies. The number of SCR non-responses to the US was substantially lower compared to non-responses to the CSs (see **Figure 10A**). Furthermore, US non-responses significantly predicted CS non-responses (see **Figure 10B**), but not vice versa, suggesting that specifically US non-responses may be more appropriate for classifying physiological SCR ‘non-responders’ than CS responses. Yet, it is difficult to formulate a generally valid criterion as its definition also hinges on experimental variables such as design (e.g., single- vs. multiple-day paradigm), hardware, or sampling rate.

However, several recommendations can be derived from the results of **Study III**: First, researchers should critically evaluate whether ‘non-learning’ and ‘non-responding’ criteria correspond to their specific study context or consider whether they could empirically derive criteria from their own study. Second, as the majority of participants indeed showed significant discrimination by applying different criteria of ‘non-learning’, a post-hoc check could be conducted to assure that the applied criteria worked as intended. Third, the identified ‘non-learners’ particularly discriminated between CSs in fear ratings. Thus, participants should not be excluded on the basis of one outcome measure alone. Fourth, while it might be meaningful to exclude physical ‘non-responders’ as they do not react in general in SCR, other exclusions can be a threat to the generalizability of findings. Finally, for reasons of transparency, we recommend that researchers should also report results in the absence of the application of these criteria – at least in the supplementary material. Fearing divergent results is unwarranted, as they hold the informative

potential to reveal possible boundary conditions under which the effects emerge (or not). In sum, an informed approach to treating ‘non-learners’ and ‘non-responders’ might protect against relying on an effect that is only present in a specific subgroup or missing a significant effect because a relevant subgroup was excluded.

To navigate the garden of forking paths, many decisions must be carefully made, including how to handle ‘non-responders’ and ‘non-learners’. This key decision point concerns the pre-processing of data which was addressed in **Study III** in a small (data) manyverse. Another crucial decision following the pre-processing steps involves the selection of an appropriate statistical analysis model – the selection of which is similarly rendered difficult by a substantial number of options and heterogeneity in the literature. A model multiverse, as used in **Study IV**, is an effective tool for exploring the impact of this heterogeneity on results.

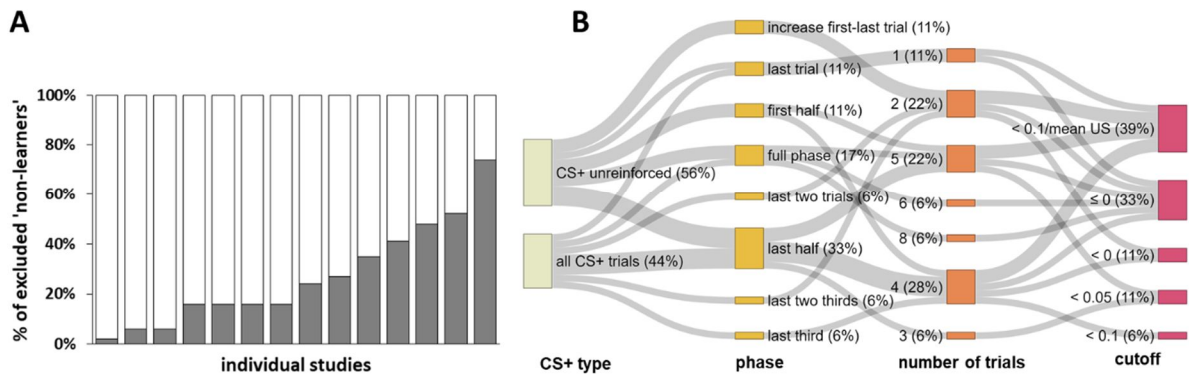


Figure 8 Illustration of the percentage of participants excluded as ‘non-learners’ across 14 individual studies derived from 11 different records (A). The percentages shown do not necessarily reflect the total number of participants excluded per study as some studies excluded participants on the basis of both, ‘non-learning’ and ‘non-responding’, e.g. the study with the highest percentage of exclusions (74%). The sankey plot (B) illustrates the different paths and combinations of CS+ types used, experimental phases, number of trials, and SCR CS discrimination cut-off scores defining ‘non-learners’ across studies. The width of the paths is scaled according to the frequency of the combinations used. Some percentages for certain combinations may not add up to 100% due to rounding.

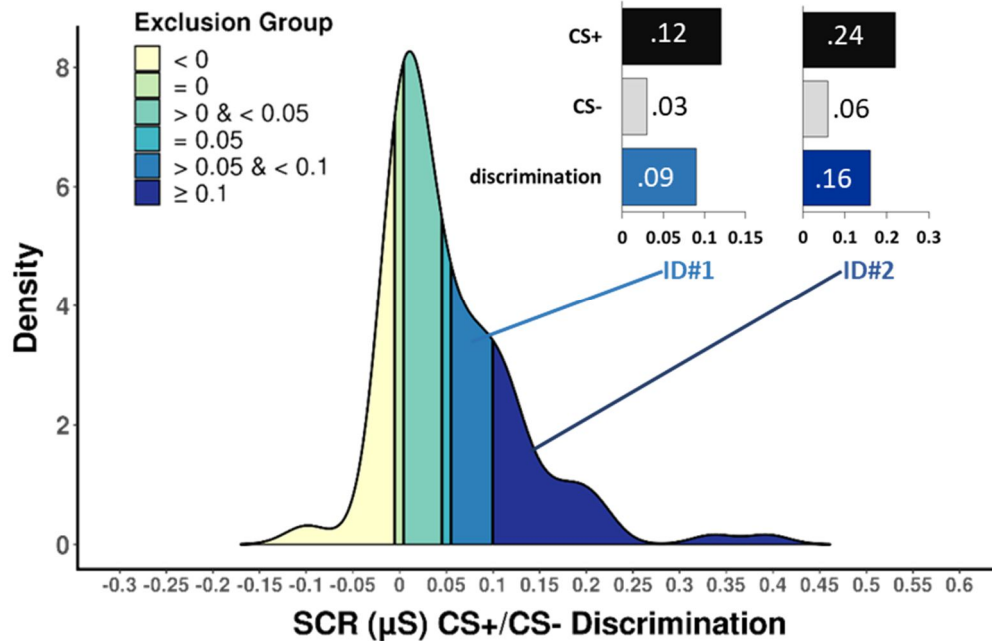


Figure 9. Illustration of the density distribution of CS discrimination [(CS+) – (CS-)] in raw SCRs (data set 1). Different colored areas represent parts of the sample that would be excluded by applying different cutoff criteria (see legend). Two participants are depicted as case examples (ID#1 and ID#2) who are part of different exclusion groups but – most importantly – show an identical discrimination ratio (4:1) indicating the higher probability of participants exhibiting stronger SCR amplitudes to remain in the final sample. Please note that in practice, the groups are cumulative, implying that the groups depicted in lighter shades always include the groups illustrated in darker shades. They are, however, considered distinct groups for illustrative purposes.

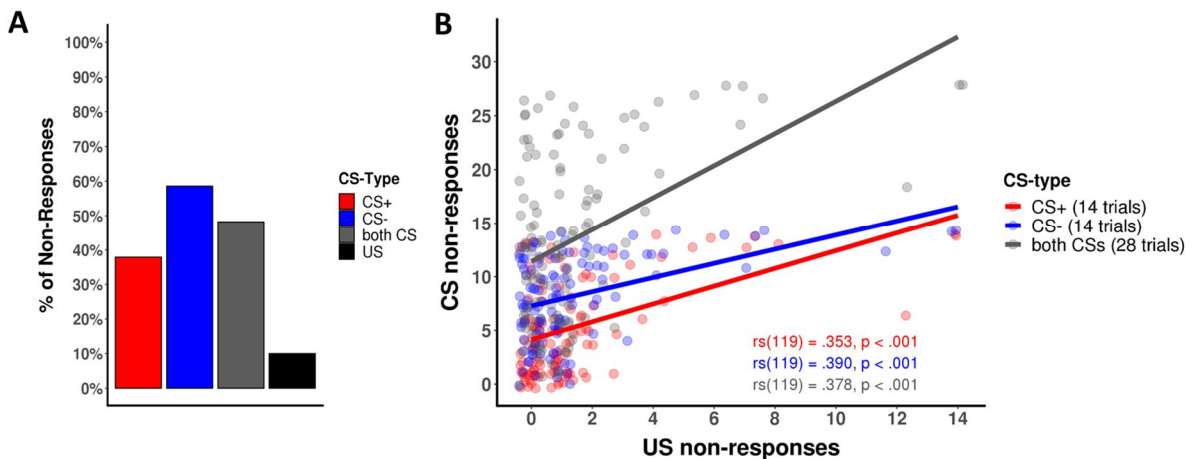


Figure 10 Illustration of percentage of SCR ‘non-responses’ across stimuli (A) and the correlation (Spearman) of US ‘non-responses’ and CS ‘non-responses’ (B) across all participants (data set 1). Colors indicate the type of stimulus (red = CS+, blue = CS-, gray = both CSs, black = US). In this study, ‘non-responses’ were defined as responses with an amplitude < 0.01 mS.

3.4 Study IV: Multiverse analyses in fear conditioning research

In fear conditioning research, the exploration of the garden of forking paths mainly focused on data manyverses or multiverses including assessments of the effects of different definitions of ‘non-learners’/‘non-responders’ (see **Study III**), ‘extinction retention’ (Lonsdorf et al., 2019) or SCR quantification approaches on results (Kuhn et al., 2022; Sjouwerman et al., 2022). However, there is also a wide variety of approaches to aggregate and analyze data by using different data reduction procedures and statistical models (e.g., t-tests, ANOVAs, or mixed models) respectively. Data reduction procedures refer to data processing steps such as the inclusion of single trial data, averaging across all trials, or specific blocks of an experimental phase prior to the data integration into a statistical model. These data reduction procedures and statistical models are often equally plausible and justifiable but have been shown to produce substantially different results and conclusions (Botvinik-Nezer et al., 2020; Dutilh et al., 2019; Kuhn et al., 2022; Silberzahn et al., 2018).

Presumably, researchers are provided with more degrees of freedom in choosing statistical models due to a lack of formalization of psychological theories as these are often verbally described as latent constructs and their associations (Farrell & Lewandowsky, 2018; Lewandowsky & Farrell, 2010). The specific assumptions and details of the theory may be left to the researcher, which can result in the translation of a single theory into different statistical models. To formalize verbal theories in fear conditioning, understanding how specific methods – or more precisely specific statistic models – affect results is crucial. A multiverse approach, in which all possible statistical models are applied in one step in the context of a single analysis, shows promise for improving this understanding. Thus, the aims of **Study IV** were to i) introduce the model multiverse concept in the field of fear conditioning, ii) illustrate the impact of utilizing various statistical models (as identified previously through a systematic literature review) on the outcomes of two distinct data sets; and iii) present the user-friendly R software package *multifear* which allows for the execution of multiverse analyses with a single line of code.

In the literature, a wide variety of data reduction approaches (depicted in **Figure 12**) as well as statistical models were identified. The identified statistical models include frequentist as well as Bayesian approaches (i.e., ANOVA, t-tests, and mixed models). Among these models, repeated measures ANOVA was the most commonly used.

All different combinations of these data reduction procedures and statistical models identified in the literature were included in the following multiverse analysis, also referred to as

model multiverse, which was calculated for SCRs during acquisition and extinction training. The mean p-values across all identified statistical models and data reduction procedures (see **Figure 11**) indicated significant CS discrimination during acquisition in data set 1 but was only marginal in data set 2. For extinction training, the mean p-values were non-significant in both data sets. Similarly, for acquisition training, the proportions of the Bayes factor above 1 were 70 – 100%, indicating that there was stronger evidence for the alternative hypothesis than for the null hypothesis, while for extinction training, these proportions were 50% or lower. The inspection of CS discrimination (i.e., “CS effect”) during acquisition training yielded medium to large effect sizes while the interaction of CS type and time showed small to medium effects in data set 1 (see **Figure 12**). Contrarily, in data set 2, effect sizes for CS discrimination were rather inconsistent and ranged from approaching zero to large effects. The addition of the time factor (CS by time interaction) resulted in (very) low effect sizes. For extinction training, effect sizes were expectedly low in data set 2 as SCRs to the CS+ typically wane throughout the extinction phase resulting in lower CS discrimination. Surprisingly, substantial CS discrimination was observed in data set 1 for the majority of the models. This might be attributed to the use of different reinforcement rates in data set 1 (partial reinforcement) and data set 2 (continuous pairing) as partial reinforcement is linked to a slower process of extinction learning (Dunsmoor et al., 2007; Haselgrove et al., 2004).

Size and direction (i.e., significant vs. non-significant) of effects appeared to hinge on the selection of the statistical model and the choice of included (blocks of) trials. Hence, while multiple valid options may exist that could be equally justifiable, choosing one particular analytical option over the other might result in different findings and conclusions questioning the robustness of effects.

In future fear conditioning studies, reporting multiverse analyses, at minimum in the supplement, could enhance our comprehension of experimental boundary conditions (e.g., the inclusion of specific trial numbers or covariates), that significantly affect the magnitude of the effect under investigation. Thus, multiverse studies can aid to unravel which model generates robust effects. In this context, the outcomes obtained from conducting multiverse analyses might offer an ideal foundation for constructing more sophisticated formal theoretical frameworks (Oberauer & Lewandowsky, 2019). These formal theories can then be tested more directly by implementing a specific statistical model and thereby reducing methodological heterogeneity and producing more consistent results. As fear conditioning procedures hold significant translational relevance in shaping future clinical intervention and prevention programs, it is crucial to align analytical techniques across studies. This would enable comparisons between study results, enhance

replicability, and ultimately accelerate the translation of fear conditioning research into clinical practice.

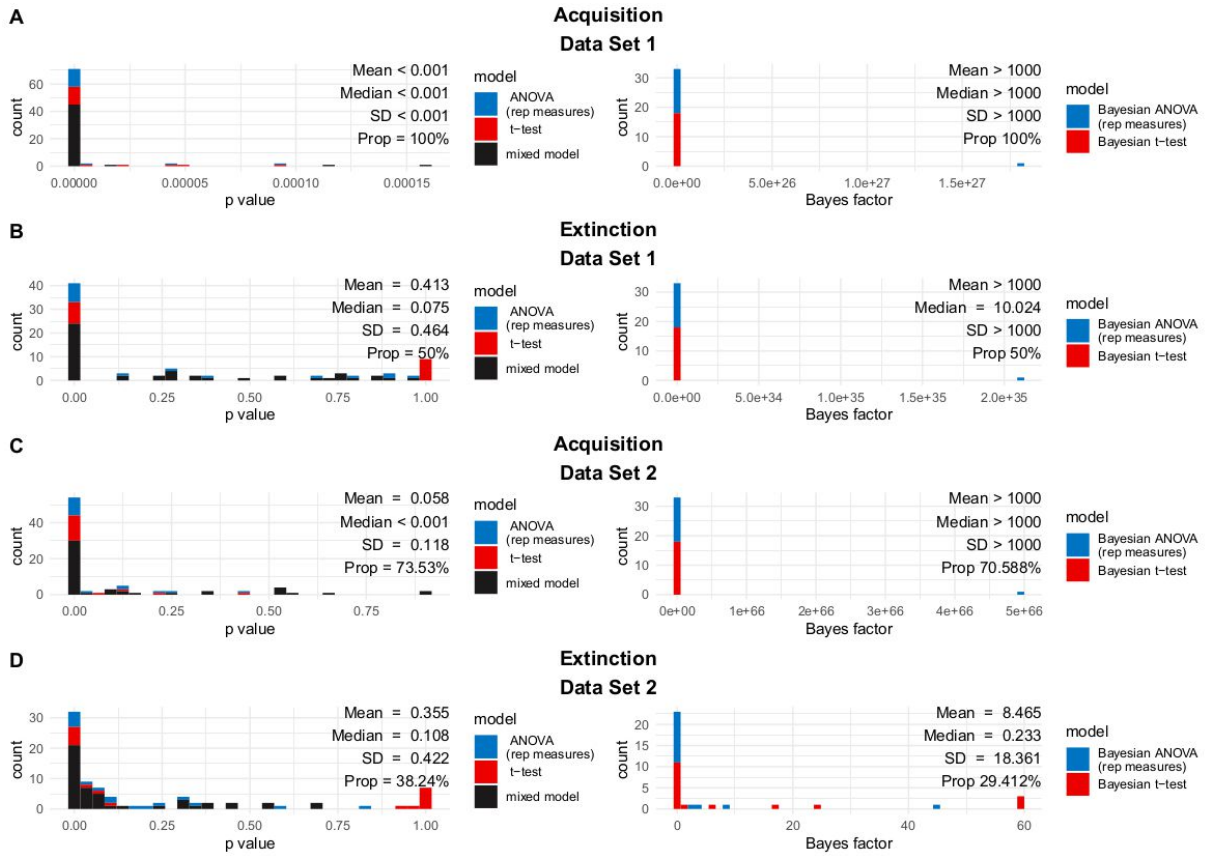


Figure 11 Histograms illustrating the amount of various p-values resulting from the classic NHST (left panel) and Bayes factors according to the Bayesian approach (right panel) resulting from the multiverse calculations for data set 1 (A: acquisition training and B: extinction training) and data set 2 (C: acquisition training and D: extinction training). Bayes factors above 1 point toward evidence for the alternative hypothesis relative to the null hypothesis while for factors below 1 the opposite is true.

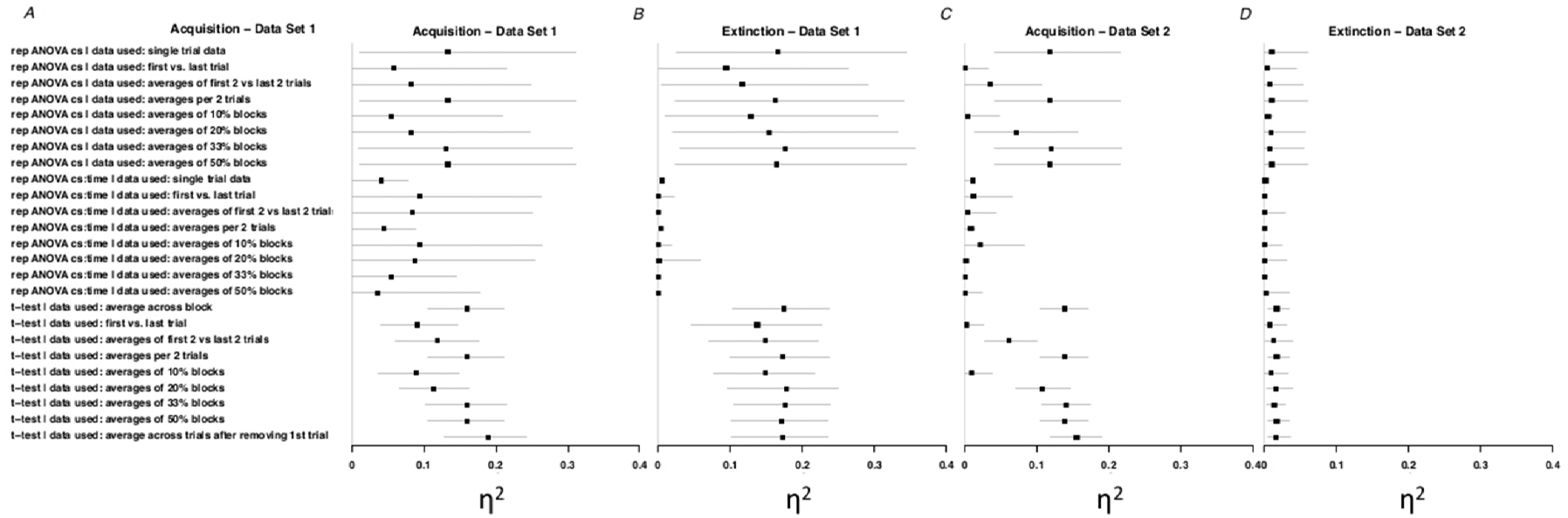


Figure 12 Forest plot illustrating the effect sizes that result from different analyses as conducted within the multiverse analyses for acquisition training (data set 1: A, data set 2: C) and extinction training (data set 1: B, data set 2: D). To render effect sizes more comparable, effect sizes resulting from t-tests were transformed into eta squared as extracted from ANOVAs. Error bars represent 90% confidence intervals indicating significance when zero is not included.

4 General discussion

The overall objective of this thesis was to deepen our understanding of the methodological heterogeneity involved in fear conditioning research, with the ultimate goal of contributing to the ongoing generation of cumulative knowledge on result robustness. This was accomplished through a comprehensive series of four studies. Most of the data analyzed in the four presented studies originate from a longitudinal investigation in a large sample that spanned six measurement time points. Among these time points, two comprised a two-day differential fear conditioning paradigm in which several outcome measures were acquired including physiological (e.g., fMRI and SCR) and self-report data (e.g., fear ratings and questionnaires).

Although the studies collectively indicated limited robustness of results, they offer valuable insights into the complexities of the topic. These insights can be leveraged to apply existing or to develop novel methods that promote greater robustness in future research. While **Study I** did raise questions about the robustness of specific brain-behavior interactions, as they could not be conceptually replicated, these insights serve as an important starting point for further investigation. Moreover, individual differences in conditioned responding in outcome measures that are commonly used in the field are robust at short time intervals but less so over longer time intervals (**Study II**). Furthermore, the exclusion of specific subsets of participants (**Study III**) and the choice of specific statistical models (**Study IV**) substantially affected the results and their interpretation. Importantly, all of the studies included in this thesis revealed substantial methodological heterogeneity in the literature at several steps of the research process with regard to experimental design (**Study I**), operationalization of experimental phases, data specifications, and reliability measures (**Study II**), the criteria for excluding specific participants (**Study III**), and the statistical approaches employed (**Study IV**). Through the application of these various methodological approaches in each study, it has become clear that there are opportunities for improving the robustness and facilitating the integration and literature embedding of study outcomes. By addressing methodological heterogeneity and robustness, researchers can work toward developing a more comprehensive understanding of these important topics. These efforts will ultimately contribute to increased comparability, integrability, and generalizability of findings, supporting greater replication and advancing our knowledge in fear conditioning research.

The findings of the studies included in this thesis, as outlined above, align well with previous work demonstrating that methodological heterogeneity can result in significant variations in outcomes and their interpretation when identical data was submitted to various analysts, despite initially sharing identical raw data (Botvinik-Nezer et al., 2020; Silberzahn et al., 2018). In fear

conditioning research, there have been empirical investigations and discussions on definitions and applications of analytical approaches as in **Study IV** (Ney et al., 2020, 2022), different ‘extinction retention indices’ (Lonsdorf et al., 2019), and SCR quantification approaches (Kuhn et al., 2022; Sjouwerman et al., 2022) as well as the procedural heterogeneity and robustness in studies scrutinizing reinstatement-dependent ROF (Haaker et al., 2014; Sjouwerman & Lonsdorf, 2020).

These studies have collectively highlighted that greater methodological homogeneity may be worth striving for in order to promote the robustness of results. To date, only a few steps in the scientific process of conducting and analyzing fear conditioning experiments have been investigated meta-scientifically (Nebe et al., 2023). As such, my thesis aims to contribute to the meta-scientific understanding of fear conditioning research and provides implications and potential remedies that address both challenges of decreasing methodological heterogeneity and improving the robustness of results. These remedies, which are detailed below, include the direct and indirect enhancement of robustness by increasing reliability and precision, reducing methodological heterogeneity, developing and refining theories, promoting transparency in reporting and in visual data presentation, and, in a broader sense, a change of the scientific culture (Ney et al., 2018, 2020; Nosek et al., 2022; Zorowitz & Niv, 2023).

In the recent past, there have been several suggestions to increase robustness of fear conditioning results directly (Ney et al., 2018), which are also frequently recommended for increasing reliability, such as augmenting the sample size (Ney et al., 2020; Zorowitz & Niv, 2023). In **Study I**, in which we could not conceptually replicate specific brain-morphology interactions, the sample was quite large. However, we cannot rule out the possibility that the brain-morphology interaction would have been significant in an even larger sample. Yet, this would be rather unlikely, as an inspection of the literature has shown that significant effects were mainly observed in studies using smaller samples (e.g., Hartley et al., 2011; Rauch et al., 2005). Future replication studies could show, whether increasing the sample size may also enhance the robustness of the absence of effects.

A further proposal to enhance robustness and reliability would be to optimize the experimental design, e.g. by increasing the number of trials (Ney et al., 2020; Zorowitz & Niv, 2023). However, we did not observe a significant effect of raising the number of trials on reliability in **Study II**. Zorowitz and Niv (2023) proposed another approach to improve reliability which involves increasing the variability between participants. One way to achieve this in SCRs might be refraining from adjustments for individual differences (i.e., range correction) as suggested by Lykken and Venables (1971). In **Study II**, however, the utilization of different SCR transformation types, including range correction, in estimating reliability did not result in significant changes in

reliability. In sum, the results of **Study II** do not indicate a benefit from increasing the number of trials or enhancing the between-participant variance by the use of specific data transformations. Nevertheless, these findings are encouraging for the fear conditioning field, as they suggest that previous research using different experimental designs and data transformations can still be integrated.

Another proposal to boost reliability is optimizing parameter estimation through the use of advanced analytical methods (Zorowitz & Niv, 2023), which is demonstrated to be beneficial in **Study II**: While more traditional reliability estimates for fMRI suggested limited reliability, more sophisticated techniques such as the similarity approach (Fröhner et al., 2019) indicated higher reliability. It may be generally advantageous for the enhancement of robustness to move away from traditional methods, as has been proposed for the practice of routinely discarding trials (Ney, 2018), but which might also apply to other routines: **Study III** demonstrates that routine practices of participant exclusion can overlook significant individuals, thereby reducing the generalizability of the results, and hence should be avoided. Additional suggestions for improving robustness comprise reducing the level of noise present in psycho(physio)logical data or utilizing appropriate estimates of individual differences (Ney et al., 2018). However, these recommendations remain suggestions in the first place and require further exploration and analysis on how to implement them and whether they work.

Yet another suggestion, which was actually formulated to increase replicability, but which also applies to robustness due to their interdependence, involves strengthening the methods utilized to amplify the signal and minimize error (Nosek et al., 2022). This may be accomplished by, for instance, using more robust manipulations, enhancing experimental effects, or choosing the most appropriate analytical approach (see **Study IV**; Nosek et al., 2022; Smith & Little, 2018; Vazire et al., 2020; Zorowitz & Niv, 2023). However, identifying the most effective and appropriate methods to implement these proposals can be challenging due to the substantial methodological heterogeneity in psychological research. How this methodological heterogeneity can be addressed is discussed below, subsequent to another important and related, but mostly neglected, remedy to gain more robust results: enhancing the precision of measures (Nebe et al., 2023). For instance, when analyzing continuously recorded stimulus-evoked skin conductance responses (SCRs), several procedures must be carried out, all of which can affect the precision of the measurements. These include determining the magnitude of the responses (see Kuhn et al., 2022; Pineles et al., 2009; Sjouwerman et al., 2016) and the aforementioned adjustment for inter-individual differences (such as range correction; Lykken & Venables, 1971) to allow for comparisons between participants. However, only a small number of these procedures have been systematically

investigated for their impact on measurement precision (Nebe et al., 2023). Therefore, it is crucial to assess how these proposed strategies for improving precision and reliability can be adapted to fear conditioning research and if they offer any benefits. Thus, an additional solution to foster robustness might be the decrease of methodological heterogeneity in the long run. How can this be accomplished? The answer lies in exploring it. Although it may seem paradoxical, the initial step would be to enhance methodological heterogeneity – but within a single study. In other words, we need to thoroughly investigate this heterogeneity through approaches such as the many- or multiverse analyses as conducted in **Studies II, III, and IV**. It is these meta-science practices that provide information on how we can create research practices that are more efficient, and less prone to bias, yielding reliable and replicable results (Ioannidis, 2018; Ioannidis et al., 2015). These approaches can be thought of as massive and comprehensive robustness analyses (see **Study IV**).

Another solution that aids to explore methodological heterogeneity is the use of specification curve approaches (Simonsohn et al., 2020). In this approach, identical to the multiverse technique, the results of all combinations of reasonable paths – or specifications in this approach – are calculated and graphed in a so-called specification curve in which the effect sizes of all results are plotted next to each other in ascending order of magnitude. This specification curve allows specific patterns to be identified – for example, it indicates which specifications lead to higher effect sizes and thus, presumably, which specifications might boost result robustness.

With the development of multiverse and specification curve approaches, the response to the call for the development of more advanced statistical tools (Ioannidis, 2014) had already begun. Additionally, it may be beneficial to introduce and enhance the utilization of existing advanced techniques, much like in **Study IV** where the multiverse approach was employed for the first time in fear conditioning research. More sophisticated analyses such as multivariate imaging techniques (Kragel et al., 2021; Noble et al., 2021; Visser et al., 2021) or intra-individual neural response variability (Månsson et al., 2021) were demonstrated to result in comparably higher reliability estimates which in turn should foster Open field needs to enhance the understanding of the extent to which common, but also novel analytical approaches are robust. Once we have achieved this, we can devote ourselves to the development and refinement of theories.

Methodological flexibility appears to originate from ambiguity to find the “best” solution or might be more related to the lack of formal theories in the field, rather than to Questionable Research Practices (QRP) such as “p-hacking” (Simmons et al., 2011). If formal theories are absent, it is unclear how they can be translated into statistical models and directly tested. Multiverse or specification curve approaches focus on the consequences of this flexibility: The comprehensive

collection of all possible and putatively equally justifiable models resulting from the multiverse analyses might be used as a basis not only for calibration studies (Bach et al., 2020), but also improved theoretical frameworks (Oberauer & Lewandowsky, 2019) as well as computational models (Krypotos et al., 2020). Formal theories can then be tested with a more specific selection of methodologies. Consequently, we can offer particular recommendations on how to design, conduct, and analyze fear conditioning studies. This could lead to greater methodological consistency in fear conditioning research by providing empirical evidence for the “best” methods to be employed in a specific application and thus to a more successful comparison, integration, and replication of results.

A further proposal as a means to promote the ‘R-terms’ is to increase openness, which includes both transparency and accessibility (Nosek & Bar-Anan, 2012). This transparency and accessibility can be realized, if applied properly, by open science tools, such as pre-registration of studies, registered reports, reporting standards, sharing of experimental protocols, data, and analytical code as applied in **Studies I – IV** (Ioannidis, 2014; Nebe et al., 2023; Nosek et al., 2022). There is already empirical evidence suggesting that pre-registration of studies may enhance replication success (Protzko et al., 2020). The use of these tools has grown in popularity in recent years (Wallach et al., 2018), but for their application, there is still some room for improvement.

Transparency does not only concern the written report, but also the visual representation of the data which should be refined (Larson-Hall, 2017; Weissgerber et al., 2019): To present a more comprehensive view of the data and uncover patterns that may be obscured by averaged data, it is important to include not only summary statistics such as in bar graphs, but also trial-by-trial data (Ney et al., 2020), or individual data points and their distributions as realized, for instance, in beeswarm- and piratplots (Larson-Hall, 2017), or in scatter- or box plots (Weissgerber et al., 2015), particularly with included marginal densities (see **Study I**). A combination of different relevant graphs is achieved in so-called rain cloud plots, which integrate not only mean and standard deviation (or CI), but also box plots with median and quartiles, density, and individual data points (Allen et al., 2021, see also **Study II**). Such figures including multiple illustrations of the data might also contribute to the transparency of scientific results.

From a broader perspective, fostering the ‘R-terms’ also entails addressing the structural, social, and individual factors that impede their implementation (Nosek et al., 2022). These structural and social factors may include the absence of incentives for replication work and the tendency for novelty to overshadow “boring” replications. Consequently, there is a need for incentives and a shift in culture toward valuing transparency and scientific rigor over the pure number of publications. Individual factors, such as confirmation bias (Nickerson, 1998) and outcome bias (Baron &

Hershey, 1988; Nosek & Errington, 2020), may also hinder progress. These refer to the tendency of researchers to selectively focus on evidence that supports their existing beliefs (i.e., confirmation bias) or to assess replication designs based on whether the results align with their desired outcomes (i.e., outcome bias) respectively and need to be addressed in a personal endeavor. Nonetheless, there are signs of a changing research culture, with a growing ease of sharing primary data and code on platforms like Open Science Framework (OSF, osf.org), or Zenodo (zenodo.org; Nosek, 2022), and an increasing availability of open data sets, also in the field of fear conditioning (Ehlers & Lonsdorf, 2022).

However, this thesis has its limitations as it does not incorporate some of the commonly used outcome measures in fear conditioning research, apart from SCR, fear ratings, and fMRI as included here. Further important measures involve FPS, different rating types (e.g., arousal, valence, or contingency ratings), heart rate, and pupillary response (Lonsdorf et al., 2017). Although this thesis did not encompass all outcome measures, it still included several crucial ones ensuring that defensive responding was captured as a multidimensional construct, as different outcome measures are thought to touch on different mechanisms of stress- and fear-related processes (Hamm & Weike, 2005; Lipp, 2006; Vrana et al., 1988).

Another related limitation is that the functionality of the *multifear* package is currently restricted to SCRs and experimental phases such as acquisition and extinction training, but should be rapidly extended to include other outcome measures such as FPS or ratings, and other experimental phases such as fear generalization as well as other conditioning procedures, different data transformations, different participant exclusion criteria, and the integration of covariates. In addition, model multiverses as implemented in the package could be complemented by other multiverses such as data multiverses or even – and even more demanding – design multiverses (Harder, 2020).

An additional constraint of this thesis is, that it focused exclusively on a single path in the garden of forking paths up to a certain decision point (i.e., the exclusion of participants, see **Study III**) from which the many- or multiverses branched off. Ideally, however, a full multiverse should incorporate all equally reasonable alternatives for each decision point, such as study design, data collection, and SCR pre-processing. However, this approach may pose some computational challenges, and the task of identifying and determining the relative justifiability of all reasonable alternatives remains a complex and challenging one – but the acceptance of which might show us how to achieve the most robust results possible.

However, even though a full multiverse was not carried out within the framework of this thesis, it is not only dedicated to meta-science but also the work conducted in the context of this thesis employed a considerable number of meta-science and also open science tools. Studies **I**, **II**, and **III** were pre-registered and the data as well as code that was used for data analysis of **Study III** and **IV** are openly shared online (**Study I**: OSF, **Study II**: Zenodo, **Study III**: OSF). The R package *multifear*, introduced in Study IV, is available for downloading and testing on GitHub (<https://github.com/AngelosPsy/multifear>), where a sample code is also provided. Furthermore, manuscripts of **Study II** and **Study IV** were written as reproducible manuscripts in R Markdown, which can be accessed on Zenodo (**Study II**). Even though a full multiverse was only used in **Study IV**, the methodological heterogeneity in the literature was taken into account with manyverse like approaches (**Studies II and III**), and the robustness of the results was examined in additional analyses (**Study I**). Although I have addressed several meta-scientific issues in this thesis, there are still several future directions we could go from here.

If resources and time were infinite, my vision would be to rerun all the studies included in this thesis with outcome measures that have not yet been incorporated, such as FPS and heart rate. It would be particularly intriguing to investigate the reliability of these measures, given the limited amount of research on this topic. Additionally, since there is currently no published study on reproducibility of fear conditioning research to my knowledge, I would like to conduct a meta-study to determine whether previous fear conditioning results can be (computationally) reproduced by using the same data and methodology employed in the original studies. This would not only strengthen the credibility of previous findings but also provide information on which methodological approaches are particularly promising for further increasing reproducibility.

In terms of following up on the individual studies included in this thesis, my vision would be to perform a direct replication of previously reported links between brain morphology and individual differences in defensive responding in a large sample, rather than a conceptual replication as was done in **Study I**, to inspect whether the effect holds under (almost) identical conditions. Regarding **Study II**, I propose to incorporate a multiverse analysis with several conceivable time spans between measurement time points to assess at what point the reliability starts to be low. Furthermore, I would aim to integrate a multiverse of results obtained by excluding each exclusion group as identified in **Study III** from analyses of a specific data set and comparing these findings. In continuation of **Study IV**, in which a model multiverse analysis was conducted, I would intend to utilize the specification curve approach by considering all plausible specifications of a fear conditioning paradigm, including experimental design, outcome measures, pre-processing steps,

and statistical analyses, to determine which specifications yield the strongest and therefore presumably the most robust effects.

In conclusion, the fear conditioning paradigm is considered to hold strong potential for successfully translating empirical discoveries into clinical practices (Anderson & Insel, 2006; Beckers et al., 2023) but the current challenges posed by methodological heterogeneity and limited robustness of findings in the fear conditioning provide opportunities for improvement. Addressing these challenges through further research on fear conditioning research and can help reduce methodological heterogeneity, enhance robustness, and pave the way for successful replication. By embracing open science and meta-science practices more widely, we can empower scientific progress and the advancement of knowledge. Finally, reproducible, robust, and replicable findings in the field that contribute to the cumulative generation of knowledge may promote what is ultimately important: the accelerated translation of empirical fear conditioning findings into clinical applications in order to improve existing and create novel successful interventions.

5 List of Abbreviations

ANOVA	Analysis of Variance
BNST	Bed Nucleus of the Stria Terminalis
BOLD	Blood Oxygenation Level Dependent
CBT	Cognitive Behavioral Therapy
CR	Conditioned Reaction
CS	Conditioned Stimulus
dACC	dorsal Anterior Cingulate Cortex
dIPFC	dorsolateral Prefrontal Cortex
EMG	Electromyography
FPS	Fear Potentiated Startle
fMRI	functional Magnetic Resonance Imaging
HR	Heart Rate
ICC	Intra-class correlation coefficient
ITI	Inter-Trial Interval
MPRAGE	Magnetization Prepared Rapid Gradient Echo
MRI	Magnetic Resonance Imaging
NAcc	Nucleus Accumbens
NHST	Null Hypothesis Significance Testing
NS	Neutral Stimulus
OFC	Orbitofrontal Cortex
QRP	Questionable Research Practices
ROF	Return of Fear
SCL	Skin Conductance Level

SCR	Skin Conductance responses
SD	Standard Deviation
TIV	Total Intracranial Volume
UR	Unconditioned reaction
US	Unconditioned Stimulus
VAS	Visual Analogue Scale
vmPFC	ventromedial Prefrontal Cortex

6 References

- Abend, R., Gold, A. L., Britton, J. C., Michalska, K. J., Shechner, T., Sachs, J. F., Winkler, A. M., Leibenluft, E., Averbeck, B. B., & Pine, D. S. (2020). Anticipatory Threat Responding: Associations With Anxiety, Development, and Brain Structure. *Biological Psychiatry*, *87*(10), 916–925. <https://doi.org/10.1016/j.biopsych.2019.11.006>
- Ahmed, O., & Lovibond, P. F. (2019). Rule-based processes in generalisation and peak shift in human fear conditioning. *Quarterly Journal of Experimental Psychology*, *72*(2), 118–131. <https://doi.org/10.1177/1747021818766461>
- Ahrens, L. M., Pauli, P., Reif, A., Mühlberger, A., Langs, G., Aalderink, T., & Wieser, M. J. (2016). Fear conditioning and stimulus generalization in patients with social anxiety disorder. *Journal of Anxiety Disorders*, *44*, 36–46. <https://doi.org/10.1016/j.janxdis.2016.10.003>
- Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J., & Kievit, R. A. (2021). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, *4*, 63. <https://doi.org/10.12688/wellcomeopenres.15191.2>
- Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., Cheung, F., Christopherson, C. D., Cordes, A., Cremata, E. J., Della Penna, N., Estel, V., Fedor, A., Fitneva, S. A., Frank, M. C., Grange, J. A., Hartshorne, J. K., Hasselman, F., Henninger, F., ... Zuni, K. (2016). Response to Comment on “Estimating the reproducibility of psychological science”. *Science*, *351*(6277), 1037–1037. <https://doi.org/10.1126/science.aad9163>
- Anderson, K. C., & Insel, T. R. (2006). The promise of extinction research for the prevention and treatment of anxiety disorders. *Biological Psychiatry*, *60*(4), 319–321. <https://doi.org/10.1016/j.biopsych.2006.06.022>
- Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. <https://doi.org/10.1037/met0000365>
- Atlas, L. Y., Sandman, C. F., & Phelps, E. A. (2022). Rating expectations can slow aversive reversal learning. *Psychophysiology*, *59*(3). <https://doi.org/10.1111/psyp.13979>
- Baas, J. M. P., van Ooijen, L., Goudriaan, A., & Kenemans, J. L. (2008). Failure to condition to a cue is associated with sustained contextual fear. *Acta Psychologica*, *127*(3), 581–592. <https://doi.org/10.1016/j.actpsy.2007.09.009>
- Bach, D. R., Melinščak, F., Fleming, S. M., & Voelkle, M. C. (2020). Calibrating the experimental measurement of psychological attributes. *Nature Human Behaviour*, *4*(12), 1229–1235. <https://doi.org/10.1038/s41562-020-00976-8>
- Bandelow, B., & Michaelis, S. (2015). Epidemiology of anxiety disorders in the 21st century. *Dialogues in Clinical Neuroscience*, *17*(3), 327–335. <https://doi.org/10.31887/DCNS.2015.17.3/bbandelow>

- Barlow, D. H., Allen, L. B., & Basden, S. L. (2007). Psychological Treatments for Panic Disorders, Phobias, and Generalized Anxiety Disorder. In D. H. Barlow, L. B. Allen, & S. L. Basden, *A Guide to Treatments that Work* (S. 351–394). Oxford University Press. <https://doi.org/10.1093/med:psych/9780195304145.003.0013>
- Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569–579. <https://doi.org/10.1037/0022-3514.54.4.569>
- Bauer, E. A., MacNamara, A., Sandre, A., Lonsdorf, T. B., Weinberg, A., Morriss, J., & van Reekum, C. M. (2020). Intolerance of uncertainty and threat generalization: A replication and extension. *Psychophysiology*, *57*(5), e13546. <https://doi.org/10.1111/psyp.13546>
- Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., & Damasio, A. R. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science (New York, N.Y.)*, *269*(5227), 1115–1118. <https://doi.org/10.1126/science.7652558>
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck depression inventory (BDI-II)* (Bd. 10). Pearson.
- Beckers, T., Hermans, D., Lange, I., Luyten, L., Scheveneels, S., & Vervliet, B. (2023). Understanding clinical fear and anxiety through the lens of human fear conditioning. *Nature Reviews Psychology*. <https://doi.org/10.1038/s44159-023-00156-1>
- Bernstein, D. P., & Fink, L. (1998). Childhood Trauma Questionnaire: A retrospective self-report manual. *San Antonio, TX: The Psychological Corporation*.
- Bitsios, P., Szabadi, E., & Bradshaw, C. M. (2004). The fear-inhibited light reflex: Importance of the anticipation of an aversive event. *International Journal of Psychophysiology*, *52*(1), 87–95. <https://doi.org/10.1016/j.ijpsycho.2003.12.006>
- Blumenthal, T. D., Cuthbert, B. N., Filion, D. L., Hackley, S., Lipp, O. V., & Van Boxtel, A. (2005). Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology*, *42*(1), 1–15. <https://doi.org/10.1111/j.1469-8986.2005.00271.x>
- Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, *66*, 115–133. <https://doi.org/10.1016/j.cortex.2014.11.019>
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., ... Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. <https://doi.org/10.1038/s41586-020-2314-9>
- Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E., Filion, D. L., & Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*(8), 1017–1034. <https://doi.org/10.1111/j.1469-8986.2012.01384.x>

- Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin*, *114*(1), 80–99. <https://doi.org/10.1037/0033-2909.114.1.80>
- Bouton, M. E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry*, *52*(10), 976–986. [https://doi.org/10.1016/S0006-3223\(02\)01546-9](https://doi.org/10.1016/S0006-3223(02)01546-9)
- Bouton, M. E. (2004). Context and Behavioral Processes in Extinction: Table 1. *Learning & Memory*, *11*(5), 485–494. <https://doi.org/10.1101/lm.78804>
- Bouton, M. E., & Bolles, R. C. (1979a). Role of conditioned contextual stimuli in reinstatement of extinguished fear. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*(4), 368–378. <https://doi.org/10.1037/0097-7403.5.4.368>
- Bouton, M. E., & Bolles, R. C. (1979b). Contextual control of the extinction of conditioned fear. *Learning and Motivation*, *10*(4), 445–466. [https://doi.org/10.1016/0023-9690\(79\)90057-2](https://doi.org/10.1016/0023-9690(79)90057-2)
- Bouton, M. E., García-Gutiérrez, A., Zilski, J., & Moody, E. W. (2006). Extinction in multiple contexts does not necessarily make extinction less vulnerable to relapse. *Behaviour Research and Therapy*, *44*(7), 983–994. <https://doi.org/10.1016/j.brat.2005.07.007>
- Bouton, M. E., & Swartzentruber, D. (1989). Slow reacquisition following extinction: Context, encoding, and retrieval mechanisms. *Journal of Experimental Psychology: Animal Behavior Processes*, *15*(1), 43–53. <https://doi.org/10.1037/0097-7403.15.1.43>
- Bradley, M. M., & Lang, P. J. (2000). Measuring Emotion: Behavior, feeling, and physiology. In R. D. Lane, L. Nadel, G. L. Ahern, J. Allen, & A. W. Kaszniak (Hrsg.), *Cognitive Neuroscience of Emotion*. (S. 25–49). Oxford University Press.
- Brown, J. S., Kalish, H. I., & Farber, I. E. (1951). Conditioned fear as revealed by magnitude of startle response to an auditory stimulus. *Journal of Experimental Psychology*, *41*(5), 317–328. <https://doi.org/10.1037/h0060166>
- Brugha, T., Bebbington, P., Tennant, C., & Hurry, J. (1985). The List of Threatening Experiences: A subset of 12 life event categories with considerable long-term contextual threat. *Psychological Medicine*, *15*(1), 189–194. <https://doi.org/10.1017/s003329170002105x>
- Bublitzky, F., Guerra, P. M., Pastor, M. C., Schupp, H. T., & Vila, J. (2013). Additive effects of threat-of-shock and picture valence on startle reflex modulation. *PLoS One*, *8*(1), e54003. <https://doi.org/10.1371/journal.pone.0054003>
- Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Research*, *28*(2), 193–213. [https://doi.org/10.1016/0165-1781\(89\)90047-4](https://doi.org/10.1016/0165-1781(89)90047-4)
- Cacciaglia, R., Pohlack, S. T., Flor, H., & Nees, F. (2015). Dissociable roles for hippocampal and amygdalar volume in human fear conditioning. *Brain Structure & Function*, *220*(5), 2575–2586. <https://doi.org/10.1007/s00429-014-0807-8>
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., ... Wu, H. (2018). Evaluating the

- replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637–644. <https://doi.org/10.1038/s41562-018-0399-z>
- Canli, T., Qiu, M., Omura, K., Congdon, E., Haas, B. W., Amin, Z., Herrmann, M. J., Constable, R. T., & Lesch, K. P. (2006). Neural correlates of epigenesis. *Proceedings of the National Academy of Sciences*, 103(43), 16033–16038. <https://doi.org/10.1073/pnas.0601674103>
- Cannon, W. B. (1929). *Bodily Changes in Pain, Hunger, Fear and Rage: An Account of Recent Research Into the Function of Emotional Excitement* (2. Aufl.). Appleton-Century-Crofts.
- Carver, C. S. (1997). You want to measure coping but your protocol' too long: Consider the brief cope. *International Journal of Behavioral Medicine*, 4(1), 92–100. https://doi.org/10.1207/s15327558ijbm0401_6
- Caspi, A., Moffitt, T. E., Thornton, A., Freedman, D., Amell, J. W., Harrington, H., Smeijers, J., & Silva, P. A. (1996). The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research*, 6(2), 101–114. [https://doi.org/10.1002/\(SICI\)1234-988X\(199607\)6:2<101::AID-MPR156>3.3.CO;2-E](https://doi.org/10.1002/(SICI)1234-988X(199607)6:2<101::AID-MPR156>3.3.CO;2-E)
- Castegnetti, G., Tzovara, A., Staib, M., Paulus, P. C., Hofer, N., & Bach, D. R. (2016). Modeling fear-conditioned bradycardia in humans: Modeling fear-conditioned bradycardia in humans. *Psychophysiology*, 53(6), 930–939. <https://doi.org/10.1111/psyp.12637>
- Chalkia, A., Schroyens, N., Leng, L., Vanhasbroeck, N., Zenses, A.-K., Van Oudenhove, L., & Beckers, T. (2020). No persistent attenuation of fear memories in humans: A registered replication of the reactivation-extinction effect. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, 129, 496–509. <https://doi.org/10.1016/j.cortex.2020.04.017>
- Chang, C., & Maren, S. (2009). Early extinction after fear conditioning yields a context-independent and short-term suppression of conditional freezing in rats. *Learning & Memory (Cold Spring Harbor, N.Y.)*, 16(1), 62–68. <https://doi.org/10.1101/lm.1085009>
- Choi, B. C. K., & Pak, A. W. P. (2005). A catalog of biases in questionnaires. *Preventing Chronic Disease*, 2(1), A13.
- Cooper, S. E., Dunsmoor, J. E., Koval, K., Pino, E., & Steinman, S. (2022). *Test-Retest Reliability of Human Threat Conditioning and Generalization* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/84uqz>
- Cooper, S. E., van Dis, E. A. M., Hagens, M. A., Kryptos, A.-M., Nemeroff, C. B., Lissek, S., Engelhard, I. M., & Dunsmoor, J. E. (2022). A meta-analysis of conditioned fear generalization in anxiety-related disorders. *Neuropsychopharmacology*. <https://doi.org/10.1038/s41386-022-01332-2>
- Craske, M. G., Hermans, D., & Vervliet, B. (2018). State-of-the-art and future directions for extinction as a translational model for fear and anxiety. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 373(1742), 20170025. <https://doi.org/10.1098/rstb.2017.0025>

- Craske, M. G., & Mystkowski, J. L. (2006). Exposure Therapy and Extinction: Clinical Studies. In M. G. Craske, D. Hermans, & D. Vansteenwegen (Hrsg.), *Fear and learning: From basic processes to clinical implications*. (S. 217–233). American Psychological Association. <https://doi.org/10.1037/11474-011>
- Craske, M. G., Wolitzky-Taylor, K. B., Mineka, S., Zinbarg, R., Waters, A. M., Vrshek-Schallhorn, S., Epstein, A., Naliboff, B., & Ornitz, E. (2012). Elevated responding to safe conditions as a specific risk factor for anxiety versus depressive disorders: Evidence from a longitudinal investigation. *Journal of Abnormal Psychology, 121*(2), 315–324. <https://doi.org/10.1037/a0025738>
- Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage, 9*(2), 179–194. <https://doi.org/10.1006/nimg.1998.0395>
- Davis, M., & Astrachan, D. I. (1978). Conditioned fear and startle magnitude: Effects of different footshock or backshock intensities used in training. *Journal of Experimental Psychology: Animal Behavior Processes, 4*(2), 95–103. <https://doi.org/10.1037/0097-7403.4.2.95>
- Davis, M., Walker, D. L., Miles, L., & Grillon, C. (2010). Phasic vs sustained fear in rats and humans: Role of the extended amygdala in fear vs anxiety. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology, 35*(1), 105–135. <https://doi.org/10.1038/npp.2009.109>
- Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Hrsg.), *Handbook of psychophysiology* (S. 159–181). Cambridge University Press.
- Del Giudice, M., & Gangestad, S. W. (2021). A Traveler’s Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Advances in Methods and Practices in Psychological Science, 4*(1), 251524592095492. <https://doi.org/10.1177/2515245920954925>
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage, 31*(3), 968–980. <https://doi.org/10.1016/j.neuroimage.2006.01.021>
- Dimberg, U. (1987). Facial reactions, autonomic activity and experienced emotion: A three component model of emotional conditioning. *Biological Psychology, 24*(2), 105–122. [https://doi.org/10.1016/0301-0511\(87\)90018-4](https://doi.org/10.1016/0301-0511(87)90018-4)
- Dirikx, T., Beckers, T., Muyls, C., Eelen, P., Vansteenwegen, D., Hermans, D., & D’hooge, R. (2007). Differential Acquisition, Extinction, and Reinstatement of Conditioned Suppression in Mice. *Quarterly Journal of Experimental Psychology, 60*(10), 1313–1320. <https://doi.org/10.1080/17470210701515785>
- Dirikx, T., Vansteenwegen, D., Eelen, P., & Hermans, D. (2009). Non-differential return of fear in humans after a reinstatement procedure. *Acta Psychologica, 130*(3), 175–182. <https://doi.org/10.1016/j.actpsy.2008.12.002>

- Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., van den Hout, M. A., & Baas, J. M. P. (2015). UPDATED META-ANALYSIS OF CLASSICAL FEAR CONDITIONING IN THE ANXIETY DISORDERS: Review: Updated Meta-Analysis of Fear Conditioning in Anxiety Disorders. *Depression and Anxiety*, 32(4), 239–253. <https://doi.org/10.1002/da.22353>
- Dunsmoor, J. E., Bandettini, P. A., & Knight, D. C. (2007). Impact of continuous versus intermittent CS-UCS pairing on human brain activation during Pavlovian fear conditioning. *Behavioral Neuroscience*, 121(4), 635–642. <https://doi.org/10.1037/0735-7044.121.4.635>
- Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking Extinction. *Neuron*, 88(1), 47–63. <https://doi.org/10.1016/j.neuron.2015.09.028>
- Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Kryptos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., van Maanen, L., ... Donkin, C. (2019). The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models. *Psychonomic Bulletin & Review*, 26(4), 1051–1069. <https://doi.org/10.3758/s13423-017-1417-2>
- Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00621>
- Ehlers, M. R., & Lonsdorf, T. B. (2022). Data sharing in experimental fear and anxiety research: From challenges to a dynamically growing database in 10 simple steps. *Neuroscience & Biobehavioral Reviews*, 143, 104958. <https://doi.org/10.1016/j.neubiorev.2022.104958>
- Ekman, P., & Cordaro, D. (2011). What is Meant by Calling Emotions Basic. *Emotion Review*, 3(4), 364–370. <https://doi.org/10.1177/1754073911410740>
- Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, 31(7), 792–806. <https://doi.org/10.1177/0956797620916786>
- Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.
- Fava, G. A., Rafanelli, C., Grandi, S., Conti, S., Ruini, C., Mangelli, L., & Belluardo, P. (2001). Long-term outcome of panic disorder with agoraphobia treated by exposure. *Psychological Medicine*, 31(5), 891–898. <https://doi.org/10.1017/S0033291701003592>
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole Brain Segmentation. *Neuron*, 33(3), 341–355. [https://doi.org/10.1016/S0896-6273\(02\)00569-X](https://doi.org/10.1016/S0896-6273(02)00569-X)
- Fliege, H., Rose, M., Arck, P., Levenstein, S., & Klapp, B. F. (2009). *PSQ - Perceived Stress Questionnaire*. <https://doi.org/10.23668/PSYCHARCHIVES.5138>

- Flor, H. (1991). *Psychobiologie des Schmerzes: Empirische Untersuchungen zur Psychobiologie, Diagnostik und Therapie chronischer Schmerzsyndrome der Skelettmuskulatur* (1. Aufl). Huber.
- Foa, E. B., Grayson, J. B., Steketee, G. S., Doppelt, H. G., Turner, R. M., & Latimer, P. R. (1983). Success and failure in the behavioral treatment of obsessive-compulsives. *Journal of Consulting and Clinical Psychology, 51*(2), 287–297. <https://doi.org/10.1037/0022-006X.51.2.287>
- Fox, P. T., & Raichle, M. E. (1986). Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proceedings of the National Academy of Sciences, 83*(4), 1140–1144. <https://doi.org/10.1073/pnas.83.4.1140>
- Fox, P. T., Raichle, M. E., Mintun, M. A., & Dence, C. (1988). Nonoxidative Glucose Consumption During Focal Physiologic Neural Activity. *Science, 241*(4864), 462–464. <https://doi.org/10.1126/science.3260686>
- Fredrikson, M., Annas, P., Georgiades, A., Hursti, T., & Tersman, Z. (1993). Internal consistency and temporal stability of classically conditioned skin conductance responses. *Biological Psychology, 35*(2), 153–163. [https://doi.org/10.1016/0301-0511\(93\)90011-V](https://doi.org/10.1016/0301-0511(93)90011-V)
- Freeston, M. H., Rhéaume, J., Letarte, H., Dugas, M. J., & Ladouceur, R. (1994). Why do people worry? *Personality and Individual Differences, 17*(6), 791–802. [https://doi.org/10.1016/0191-8869\(94\)90048-5](https://doi.org/10.1016/0191-8869(94)90048-5)
- Fröhner, J. H., Teckentrup, V., Smolka, M. N., & Kroemer, N. B. (2019). Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects. *NeuroImage, 195*, 174–189. <https://doi.org/10.1016/j.neuroimage.2019.03.053>
- Fullana, M. A., Albajes-Eizagirre, A., Soriano-Mas, C., Vervliet, B., Cardoner, N., Benet, O., Radua, J., & Harrison, B. J. (2018). Fear extinction in the human brain: A meta-analysis of fMRI studies in healthy participants. *Neuroscience & Biobehavioral Reviews, 88*, 16–25. <https://doi.org/10.1016/j.neubiorev.2018.03.002>
- Fullana, M. A., Albajes-Eizagirre, A., Soriano-Mas, C., Vervliet, B., Cardoner, N., Benet, O., Radua, J., & Harrison, B. J. (2019). Amygdala where art thou? *Neuroscience & Biobehavioral Reviews, 102*, 430–431. <https://doi.org/10.1016/j.neubiorev.2018.06.003>
- Fullana, M. A., Dunsmoor, J. E., Schruers, K. R. J., Savage, H. S., Bach, D. R., & Harrison, B. J. (2020). Human fear conditioning: From neuroscience to the clinic. *Behaviour Research and Therapy, 124*, 103528. <https://doi.org/10.1016/j.brat.2019.103528>
- Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., & Radua, J. (2016). Neural signatures of human fear conditioning: An updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry, 21*(4), 500–508. <https://doi.org/10.1038/mp.2015.88>
- Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (Second edition). SAGE.
- Galatzer-Levy, I. R., Bonanno, G. A., Bush, D. E. A., & Ledoux, J. E. (2013). Heterogeneity in threat extinction learning: Substantive and methodological considerations for identifying

- individual difference in response to stress. *Frontiers in Behavioral Neuroscience*, 7, 55. <https://doi.org/10.3389/fnbeh.2013.00055>
- Garnefski, N., & Kraaij, V. (2006). Cognitive emotion regulation questionnaire – development of a short 18-item version (CERQ-short). *Personality and Individual Differences*, 41(6), 1045–1053. <https://doi.org/10.1016/j.paid.2006.04.010>
- Gazendam, F. J., Kamphuis, J. H., & Kindt, M. (2013). Deficient safety learning characterizes high trait anxious individuals. *Biological Psychology*, 92(2), 342–352. <https://doi.org/10.1016/j.biopsycho.2012.11.006>
- 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*. Columbia Statistics. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
- Genon, S., Wensing, T., Reid, A., Hoffstaedter, F., Caspers, S., Grefkes, C., Nickl-Jockschat, T., & Eickhoff, S. B. (2017). Searching for behavior relating to grey matter volume in a-priori defined right dorsal premotor regions: Lessons learned. *NeuroImage*, 157, 144–156. <https://doi.org/10.1016/j.neuroimage.2017.05.053>
- Gerlicher, A. M. V., Tüscher, O., & Kalisch, R. (2018). Dopamine-dependent prefrontal reactivations explain long-term benefit of fear extinction. *Nature Communications*, 9(1), 4294. <https://doi.org/10.1038/s41467-018-06785-y>
- Gershman, S. J., & Hartley, C. A. (2015). Individual differences in learning predict the return of fear. *Learning & Behavior*, 43(3), 243–250. <https://doi.org/10.3758/s13420-015-0176-z>
- Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science”. *Science*, 351(6277), 1037–1037. <https://doi.org/10.1126/science.aad7243>
- Glotzbach-Schoon, E., Andreatta, M., Mühlberger, A., & Pauli, P. (2015). Reinstatement of contextual anxiety in humans: Effects of state anxiety. *International Journal of Psychophysiology*, 98(3), 557–566. <https://doi.org/10.1016/j.ijpsycho.2015.07.013>
- Golkar, A., & Öhman, A. (2012). Fear extinction in humans: Effects of acquisition–extinction delay and masked stimulus presentations. *Biological Psychology*, 91(2), 292–301. <https://doi.org/10.1016/j.biopsycho.2012.07.007>
- Graham, B. M., & Milad, M. R. (2011). The study of fear extinction: Implications for anxiety disorders. *The American Journal of Psychiatry*, 168(12), 1255–1265. <https://doi.org/10.1176/appi.ajp.2011.11040557>
- Granholm, E., & Steinhauer, S. R. (2004). Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, 52(1), 1–6. <https://doi.org/10.1016/j.ijpsycho.2003.12.001>
- Grillon, C. (2008). Models and mechanisms of anxiety: Evidence from startle studies. *Psychopharmacology*, 199(3), 421–437. <https://doi.org/10.1007/s00213-007-1019-1>

- Haaker, J., Golkar, A., Hermans, D., & Lonsdorf, T. B. (2014). A review on human reinstatement studies: An overview and methodological challenges. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *21*(9), 424–440. <https://doi.org/10.1101/lm.036053.114>
- Hamm, A. O., Greenwald, M. K., Bradley, M. M., & Lang, P. J. (1993). Emotional learning, hedonic change, and the startle probe. *Journal of Abnormal Psychology*, *102*(3), 453–465. <https://doi.org/10.1037/0021-843X.102.3.453>
- Hamm, A. O., & Weike, A. I. (2005). The neuropsychology of fear learning and fear regulation. *International Journal of Psychophysiology*, *57*(1), 5–14. <https://doi.org/10.1016/j.ijpsycho.2005.01.006>
- Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., & Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, *32*(1), 180–194. <https://doi.org/10.1016/j.neuroimage.2006.02.051>
- Harder, J. A. (2020). The Multiverse of Methods: Extending the Multiverse Analysis to Address Data-Collection Decisions. *Perspectives on Psychological Science*, *15*(5), 1158–1177. <https://doi.org/10.1177/1745691620917678>
- Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., deMayo, B. E., Long, B., Yoon, E. J., & Frank, M. C. (2021). Analytic reproducibility in articles receiving open data badges at the journal *Psychological Science*: An observational study. *Royal Society Open Science*, *8*(1), 201494. <https://doi.org/10.1098/rsos.201494>
- Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., & Weinberger, D. R. (2002). The amygdala response to emotional stimuli: A comparison of faces and scenes. *NeuroImage*, *17*(1), 317–323. <https://doi.org/10.1006/nimg.2002.1179>
- Härter, M. C., Conway, K. P., & Merikangas, K. R. (2003). Associations between anxiety disorders and physical illness. *European Archives of Psychiatry and Clinical Neuroscience*, *253*(6), 313–320. <https://doi.org/10.1007/s00406-003-0449-y>
- Hartley, C. A., Fischl, B., & Phelps, E. A. (2011). Brain Structure Correlates of Individual Differences in the Acquisition and Inhibition of Conditioned Fear. *Cerebral Cortex*, *21*(9), 1954–1962. <https://doi.org/10.1093/cercor/bhq253>
- Hartley, C. A., Gorun, A., Reddan, M. C., Ramirez, F., & Phelps, E. A. (2014). Stressor controllability modulates fear extinction in humans. *Neurobiology of Learning and Memory*, *113*, 149–156. <https://doi.org/10.1016/j.nlm.2013.12.003>
- Haselgrove, M., Aydin, A., & Pearce, J. M. (2004). A Partial Reinforcement Extinction Effect Despite Equal Rates of Reinforcement During Pavlovian Conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*(3), 240–250. <https://doi.org/10.1037/0097-7403.30.3.240>
- Häuser, W., Schmutzer, G., & Glaesmer, H. (2011). Maltreatment in childhood and adolescence. *Deutsches Ärzteblatt International*, *108*(17), 287–294.

- Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence Based Nursing, 18*(3), 66–67. <https://doi.org/10.1136/eb-2015-102129>
- Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166–1186. <https://doi.org/10.3758/s13428-017-0935-1>
- Hermans, D., Dirikx, T., Vansteenwegen, D., Baeyens, F., Van den Bergh, O., & Eelen, P. (2005). Reinstatement of fear responses in human aversive conditioning. *Behaviour Research and Therapy, 43*(4), 533–551. <https://doi.org/10.1016/j.brat.2004.03.013>
- Hofmann, S. G., & Smits, J. A. J. (2008). Cognitive-behavioral therapy for adult anxiety disorders: A meta-analysis of randomized placebo-controlled trials. *The Journal of Clinical Psychiatry, 69*(4), 621–632. <https://doi.org/10.4088/jcp.v69n0415>
- Holland, P. C. (1992). Occasion Setting in Pavlovian Conditioning. In *Psychology of Learning and Motivation* (Bd. 28, S. 69–125). Elsevier. [https://doi.org/10.1016/S0079-7421\(08\)60488-0](https://doi.org/10.1016/S0079-7421(08)60488-0)
- Huff, N. C., Hernandez, J. A., Blanding, N. Q., & LaBar, K. S. (2009). Delayed extinction attenuates conditioned fear renewal and spontaneous recovery in humans. *Behavioral Neuroscience, 123*(4), 834–843. <https://doi.org/10.1037/a0016511>
- Indovina, I., Robbins, T. W., Núñez-Elizalde, A. O., Dunn, B. D., & Bishop, S. J. (2011). Fear-Conditioning Mechanisms Associated with Trait Vulnerability to Anxiety in Humans. *Neuron, 69*(3), 563–571. <https://doi.org/10.1016/j.neuron.2010.12.034>
- Infantolino, Z. P., Luking, K. R., Sauder, C. L., Curtin, J. J., & Hajcak, G. (2018). Robust is not necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons. *NeuroImage, 173*, 146–152. <https://doi.org/10.1016/j.neuroimage.2018.02.024>
- Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision medicine for psychiatry. *The American Journal of Psychiatry, 171*(4), 395–397. <https://doi.org/10.1176/appi.ajp.2014.14020138>
- Ioannidis, J. P. A. (2014). How to make more published research true. *PLoS Medicine, 11*(10), e1001747. <https://doi.org/10.1371/journal.pmed.1001747>
- Ioannidis, J. P. A. (2018). Meta-research: Why research on research matters. *PLOS Biology, 16*(3), e2005468. <https://doi.org/10.1371/journal.pbio.2005468>
- Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLOS Biology, 13*(10), e1002264. <https://doi.org/10.1371/journal.pbio.1002264>
- Jacobi, F., Höfler, M., Siegert, J., Mack, S., Gerschler, A., Scholl, L., Busch, M. A., Hapke, U., Maske, U., Seiffert, I., Gaebel, W., Maier, W., Wagner, M., Zielasek, J., & Wittchen, H.-U. (2014). Twelve-month prevalence, comorbidity and correlates of mental disorders in Germany: The Mental Health Module of the German Health Interview and Examination Survey for Adults (DEGS1-MH): 12-Month Prevalence of Mental Disorders in Germany.

- International Journal of Methods in Psychiatric Research*, 23(3), 304–319.
<https://doi.org/10.1002/mpr.1439>
- Janke, W., & Erdmann, G. (2002). *SVF 78: Eine kurzform des stressverarbeitungsfragebogens SVF 120; kurzbeschreibung und grundlegende kennwerte; manual*. Hogrefe, Verlag für Psychologie.
- Jiang, Y.-Z., Liu, Y., Xiao, Y., Hu, X., Jiang, L., Zuo, W.-J., Ma, D., Ding, J., Zhu, X., Zou, J., Verschraegen, C., Stover, D. G., Kaklamani, V., Wang, Z.-H., & Shao, Z.-M. (2021). Molecular subtyping and genomic profiling expand precision medicine in refractory metastatic triple-negative breast cancer: The FUTURE trial. *Cell Research*, 31(2), 178–186. <https://doi.org/10.1038/s41422-020-0375-9>
- Johnson, D. C., Ho, W., Uddin, B., Tetteh-Quarshie, S., & Morriss, J. (2022). *Evidence for different roles of inhibitory and prospective intolerance of uncertainty during threat discrimination learning* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/qwspe>
- Kalisch, R., Korenfeld, E., Stephan, K. E., Weiskopf, N., Seymour, B., & Dolan, R. J. (2006). Context-dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 26(37), 9503–9511. <https://doi.org/10.1523/JNEUROSCI.2021-06.2006>
- Kanning, U. P. (2009). NEO-Fünf-Faktoren-Inventar nach costa und McCrae (NEO-FFI). *Zeitschrift für Arbeits- und Organisationspsychologie A&O*, 53(4), 194–198.
- Kastrati, G., Rosén, J., Fredrikson, M., Chen, X., Kuja-Halkola, R., Larsson, H., Jensen, K. B., & Åhs, F. (2022). Genetic influences on central and peripheral nervous system activity during fear conditioning. *Translational Psychiatry*, 12(1), 95. <https://doi.org/10.1038/s41398-022-01861-w>
- Kharabian Masouleh, S., Eickhoff, S. B., Hoffstaedter, F., Genon, S., & Alzheimer's Disease Neuroimaging Initiative. (2019). Empirical examination of the replicability of associations between brain structure and psychological variables. *ELife*, 8, e43464. <https://doi.org/10.7554/eLife.43464>
- Kim, J. H., & Richardson, R. (2007). A developmental dissociation in reinstatement of an extinguished fear response in rats. *Neurobiology of Learning and Memory*, 88(1), 48–57. <https://doi.org/10.1016/j.nlm.2007.03.004>
- Kline, P. (2013). *Handbook of Psychological Testing* (0 Aufl.). Routledge. <https://doi.org/10.4324/9781315812274>
- Klingelhöfer-Jens, M., Morriss, J., & Lonsdorf, T. B. (2022). Effects of intolerance of uncertainty on subjective and psychophysiological measures during fear acquisition and delayed extinction. *International Journal of Psychophysiology*, 177, 249–259. <https://doi.org/10.1016/j.ijpsycho.2022.05.006>
- Klumpers, F., Kroes, M. C., Heitland, I., Everaerd, D., Akkermans, S. E. A., Oosting, R. S., van Wingen, G., Franke, B., Kenemans, J. L., Fernández, G., & Baas, J. M. P. (2015). Dorsomedial Prefrontal Cortex Mediates the Impact of Serotonin Transporter Linked

- Polymorphic Region Genotype on Anticipatory Threat Reactions. *Biological Psychiatry*, 78(8), 582–589. <https://doi.org/10.1016/j.biopsych.2014.07.034>
- Kong, E., Monje, F. J., Hirsch, J., & Pollak, D. D. (2014). Learning not to Fear: Neural Correlates of Learned Safety. *Neuropsychopharmacology*, 39(3), 515–527. <https://doi.org/10.1038/npp.2013.191>
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kozak, M. J., Foa, E. B., & Steketee, G. (1988). Process and outcome of exposure treatment with obsessive-compulsives: Psychophysiological indicators of emotional processing. *Behavior Therapy*, 19(2), 157–169. [https://doi.org/10.1016/S0005-7894\(88\)80039-X](https://doi.org/10.1016/S0005-7894(88)80039-X)
- Kragel, P. A., Han, X., Kravynak, T. E., Gianaros, P. J., & Wager, T. D. (2021). Functional MRI Can Be Highly Reliable, but It Depends on What You Measure: A Commentary on Elliott et al. (2020). *Psychological Science*, 32(4), 622–626. <https://doi.org/10.1177/0956797621989730>
- Krypotos, A.-M., Crombez, G., Meulders, A., Claes, N., & Vlaeyen, J. W. S. (2020). Decomposing conditioned avoidance performance with computational models. *Behaviour Research and Therapy*, 133, 103712. <https://doi.org/10.1016/j.brat.2020.103712>
- Kuhn, M., Gerlicher, A. M. V., & Lonsdorf, T. B. (2022). Navigating the manyverse of skin conductance response quantification approaches – A direct comparison of TROUGH-TO-PEAK, baseline correction, and model-based approaches in Ledalab and PSPM. *Psychophysiology*. <https://doi.org/10.1111/psyp.14058>
- Kuhn, M., Mertens, G., & Lonsdorf, T. B. (2016). State anxiety modulates the return of fear. *International Journal of Psychophysiology*, 110, 194–199. <https://doi.org/10.1016/j.ijpsycho.2016.08.001>
- Kull, S., Müller, B. H., Blechert, J., Wilhelm, F. H., & Michael, T. (2012). Reinstatement of fear in humans: Autonomic and experiential responses in a differential conditioning paradigm. *Acta Psychologica*, 140(1), 43–49. <https://doi.org/10.1016/j.actpsy.2012.02.007>
- LaBar, K. S., & Phelps, E. A. (2005). Reinstatement of Conditioned Fear in Humans Is Context Dependent and Impaired in Amnesia. *Behavioral Neuroscience*, 119(3), 677–686. <https://doi.org/10.1037/0735-7044.119.3.677>
- Laireiter, A. (1993). SS-A-d: Fragebogen zur Erfassung wahrgenommener Sozialer Unterstützung: Deutschsprachige Version der SS-A-Skala von Vaux. In G. Westhoff (Hrsg.), *Handbuch empirischer Meßinstrumente* (S. 826–829). Hogrefe Verlag für Psychologie.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, 97(3), 377–395. <https://doi.org/10.1037/0033-295X.97.3.377>
- Lang, P. J., Davis, M., & Öhman, A. (2000). Fear and anxiety: Animal models and human cognitive psychophysiology. *Journal of Affective Disorders*, 61(3), 137–159. [https://doi.org/10.1016/S0165-0327\(00\)00343-8](https://doi.org/10.1016/S0165-0327(00)00343-8)

- Larson-Hall, J. (2017). Moving Beyond the Bar Plot and the Line Graph to Create Informative and Attractive Graphics 1. *The Modern Language Journal*, *101*(1), 244–270. <https://doi.org/10.1111/modl.12386>
- LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. <https://doi.org/10.1177/2515245918787489>
- LeBel, E. P., & Paunonen, S. V. (2011). Sexy But Often Unreliable: The Impact of Unreliability on the Replicability of Experimental Findings With Implicit Measures. *Personality and Social Psychology Bulletin*, *37*(4), 570–583. <https://doi.org/10.1177/0146167211400619>
- LeDoux, J. (2003). The Emotional Brain, Fear, and the Amygdala. *Cellular and Molecular Neurobiology*, *23*(4/5), 727–738. <https://doi.org/10.1023/A:1025048802629>
- Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE publications.
- Lipp, O. V. (2006). Human fear learning: Contemporary procedures and measurement. In M. G. Craske, D. E. Hermans, & D. E. Vansteenwegen (Hrsg.), *Fear and learning: From basic processes to clinical implications* (S. 37–52). American Psychological Association.
- Lipp, O. V., & Vaitl, D. (1990). Reaction time task as unconditional stimulus: Comparing aversive and nonaversive unconditional stimuli. *The Pavlovian Journal of Biological Science*, *25*(2), 77–83. <https://doi.org/10.1007/BF02964606>
- Lissek, S. (2012). TOWARD AN ACCOUNT OF CLINICAL ANXIETY PREDICATED ON BASIC, NEURALLY MAPPED MECHANISMS OF PAVLOVIAN FEAR-LEARNING: THE CASE FOR CONDITIONED OVERGENERALIZATION: Special Article: The Case for Conditioned Overgeneralization. *Depression and Anxiety*, *29*(4), 257–263. <https://doi.org/10.1002/da.21922>
- Lissek, S., Baas, J. M. P., Pine, D. S., Orme, K., Dvir, S., Nugent, M., Rosenberger, E., Rawson, E., & Grillon, C. (2005). Airpuff startle probes: An efficacious and less aversive alternative to white-noise. *Biological Psychology*, *68*(3), 283–297. <https://doi.org/10.1016/j.biopsycho.2004.07.007>
- 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. *Behaviour Research and Therapy*, *43*(11), 1391–1424. <https://doi.org/10.1016/j.brat.2004.10.007>
- Lissek, S., Rabin, S. J., McDowell, D. J., Dvir, S., Bradford, D. E., Geraci, M., Pine, D. S., & Grillon, C. (2009). Impaired discriminative fear-conditioning resulting from elevated fear responding to learned safety cues among individuals with panic disorder. *Behaviour Research and Therapy*, *47*(2), 111–118. <https://doi.org/10.1016/j.brat.2008.10.017>
- Liu, J., Wei, W., Kuang, H., Zhao, F., & Tsien, J. Z. (2013). Changes in Heart Rate Variability Are Associated with Expression of Short-Term and Long-Term Contextual and Cued Fear Memories. *PLoS ONE*, *8*(5), e63590. <https://doi.org/10.1371/journal.pone.0063590>

- Lonsdorf, T. B., & Baas, J. M. P. (2017). Genetics in Experimental Psychopathology: From Laboratory Models to Therapygenetics. Where do we go from Here? *Psychopathology Review*, *a4*(2), 169–188. <https://doi.org/10.5127/pr.037915>
- Lonsdorf, T. B., Haaker, J., & Kalisch, R. (2014). Long-term expression of human contextual fear and extinction memories involves amygdala, hippocampus and ventromedial prefrontal cortex: A reinstatement study in two independent samples. *Social Cognitive and Affective Neuroscience*, *9*(12), 1973–1983. <https://doi.org/10.1093/scan/nsu018>
- Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., Heitland, I., Hermann, A., Kuhn, M., Kruse, O., Meir Drexler, S., Meulders, A., Nees, F., Pittig, A., Richter, J., Römer, S., Shibani, Y., Schmitz, A., Straube, B., ... 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. *Neuroscience and Biobehavioral Reviews*, *77*, 247–285. <https://doi.org/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. *Neuroscience and Biobehavioral Reviews*, *80*, 703–728. <https://doi.org/10.1016/j.neubiorev.2017.07.007>
- Lonsdorf, T. B., Merz, C. J., & Fullana, M. A. (2019). Fear Extinction Retention: Is It What We Think It Is? *Biological Psychiatry*, *85*(12), 1074–1082. <https://doi.org/10.1016/j.biopsych.2019.02.011>
- Lueken, U., Straube, B., Reinhardt, I., Maslowski, N. I., Wittchen, H.-U., Ströhle, A., Wittmann, A., Pfleiderer, B., Konrad, C., Ewert, A., Uhlmann, C., Arolt, V., Jansen, A., & Kircher, T. (2014). Altered top-down and bottom-up processing of fear conditioning in panic disorder with agoraphobia. *Psychological Medicine*, *44*(2), 381–394. <https://doi.org/10.1017/S0033291713000792>
- Luyten, L., & Beckers, T. (2017). A preregistered, direct replication attempt of the retrieval-extinction effect in cued fear conditioning in rats. *Neurobiology of Learning and Memory*, *144*, 208–215. <https://doi.org/10.1016/j.nlm.2017.07.014>
- Lykken, D. T., & Venables, P. H. (1971). Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology*, *8*(5), 656–672. <https://doi.org/10.1111/j.1469-8986.1971.tb00501.x>
- Lynam, D. R., Hoyle, R. H., & Newman, J. P. (2006). The perils of partialling: Cautionary tales from aggression and psychopathy. *Assessment*, *13*(3), 328–341. <https://doi.org/10.1177/1073191106290562>
- Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*, *15*(5), e0233107. <https://doi.org/10.1371/journal.pone.0233107>
- Machlin, L., Miller, A. B., Snyder, J., McLaughlin, K. A., & Sheridan, M. A. (2019). Differential Associations of Deprivation and Threat With Cognitive Control and Fear Conditioning in Early Childhood. *Frontiers in Behavioral Neuroscience*, *13*, 80. <https://doi.org/10.3389/fnbeh.2019.00080>

- Månsson, K. N. T., Waschke, L., Manzouri, A., Furmark, T., Fischer, H., & Garrett, D. D. (2021). Moment-to-moment brain signal variability reliably predicts psychiatric treatment outcome. *Biological Psychiatry*, S0006322321016644. <https://doi.org/10.1016/j.biopsych.2021.09.026>
- Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., Moore, L. A., Conan, G., Uriarte, J., ... Dosenbach, N. U. F. (2020). *Towards Reproducible Brain-Wide Association Studies* [Preprint]. Neuroscience. <https://doi.org/10.1101/2020.08.21.257758>
- Maren, S. (2014). Nature and causes of the immediate extinction deficit: A brief review. *Neurobiology of Learning and Memory*, 113, 19–24. <https://doi.org/10.1016/j.nlm.2013.10.012>
- Masouleh, S. K., Eickhoff, S. B., & Genon, S. (2020). *Searching for replicable associations between cortical thickness and psychometric variables in healthy adults: Empirical facts* [Preprint]. Neuroscience. <https://doi.org/10.1101/2020.01.10.901181>
- MATLAB. (2016). *MATLAB* (9.1(R2016b)). The MathWorks, Inc. <https://www.mathworks.com/>
- MATLAB. (2019). *MATLAB* (9.6(R2019a)). The MathWorks, Inc. <https://www.mathworks.com/>
- McIntosh, R. D., & Chambers, C. D. (2020). The three R's of scientific integrity: Replicability, reproducibility, and robustness. *Cortex*, 129, A4–A7. <https://doi.org/10.1016/j.cortex.2020.04.019>
- Mertens, G., De Wolf, N., Bouwman, V., & Engelhard, I. M. (2022). The relationship between Intolerance of Uncertainty and conditioned fear acquisition: Evidence from a large sample. *International Journal of Psychophysiology*, 177, 67–75. <https://doi.org/10.1016/j.ijpsycho.2022.04.011>
- Merz, C. J., Hamacher-Dang, T. C., & Wolf, O. T. (2016). Immediate extinction promotes the return of fear. *Neurobiology of Learning and Memory*, 131, 109–116. <https://doi.org/10.1016/j.nlm.2016.03.013>
- Milad, M. R., Pitman, R. K., Ellis, C. B., Gold, A. L., Shin, L. M., Lasko, N. B., Zeidan, M. A., Handwerker, K., Orr, S. P., & Rauch, S. L. (2009). Neurobiological basis of failure to recall extinction memory in posttraumatic stress disorder. *Biological Psychiatry*, 66(12), 1075–1082. <https://doi.org/10.1016/j.biopsych.2009.06.026>
- Milad, M. R., & Quirk, G. J. (2012). Fear Extinction as a Model for Translational Neuroscience: Ten Years of Progress. *Annual Review of Psychology*, 63(1), 129–151. <https://doi.org/10.1146/annurev.psych.121208.131631>
- Milad, M. R., Quirk, G. J., Pitman, R. K., Orr, S. P., Fischl, B., & Rauch, S. L. (2007). A Role for the Human Dorsal Anterior Cingulate Cortex in Fear Expression. *Biological Psychiatry*, 62(10), 1191–1194. <https://doi.org/10.1016/j.biopsych.2007.04.032>
- 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. *Biological Psychiatry*, 62(5), 446–454.
<https://doi.org/10.1016/j.biopsych.2006.10.011>
- Mineka, S., & Oehlberg, K. (2008). The relevance of recent developments in classical conditioning to understanding the etiology and maintenance of anxiety disorders. *Acta Psychologica*, 127(3), 567–580. <https://doi.org/10.1016/j.actpsy.2007.11.007>
- Mineka, S., & Zinbarg, R. (2006). A contemporary learning theory perspective on the etiology of anxiety disorders: It's not what you thought it was. *American Psychologist*, 61(1), 10–26. <https://doi.org/10.1037/0003-066X.61.1.10>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Moratti, S., & Keil, A. (2005). Cortical activation during Pavlovian fear conditioning depends on heart rate response patterns: An MEG study. *Cognitive Brain Research*, 25(2), 459–471. <https://doi.org/10.1016/j.cogbrainres.2005.07.006>
- Moriarty, D. P., & Alloy, L. B. (2021). Back to Basics: The Importance of Measurement Properties in Biological Psychiatry. *Neuroscience & Biobehavioral Reviews*, 123, 72–82. <https://doi.org/10.1016/j.neubiorev.2021.01.008>
- Morriss, J., Chapman, C., Tomlinson, S., & van Reekum, C. M. (2018). Escape the bear and fall to the lion: The impact of avoidance availability on threat acquisition and extinction. *Biological Psychology*, 138, 73–80. <https://doi.org/10.1016/j.biopsycho.2018.08.017>
- Morriss, J., Zuj, D. V., & Mertens, G. (2021). The role of intolerance of uncertainty in classical threat conditioning: Recent developments and directions for future research. *International Journal of Psychophysiology*, 166, 116–126. <https://doi.org/10.1016/j.ijpsycho.2021.05.011>
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3(3), 221–229. <https://doi.org/10.1038/s41562-018-0522-1>
- Myers, K. M., & Davis, M. (2002). Behavioral and Neural Analysis of Extinction. *Neuron*, 36(4), 567–584. [https://doi.org/10.1016/S0896-6273\(02\)01064-4](https://doi.org/10.1016/S0896-6273(02)01064-4)
- Myers, K. M., & Davis, M. (2007). Mechanisms of fear extinction. *Molecular Psychiatry*, 12(2), 120–150. <https://doi.org/10.1038/sj.mp.4001939>
- National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. National Academy Press.
- Nebe, S., Reutter, M., Baker, D. H., Bölte, J., Domes, G., Gamer, M., Gärtner, A., Gießing, C., Mann, C. G. née, Hilger, K., Jawinski, P., Kulke, L., Lischke, A., Markett, S., Meier, M., Merz, C. J., Popov, T., Puhmann, L., Quintana, D. S., ... Feld, G. (2023). *Enhancing Precision in Human Neuroscience* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/m8c4k>
- Nees, F., Heinrich, A., & Flor, H. (2015). A mechanism-oriented approach to psychopathology: The role of Pavlovian conditioning. *International Journal of Psychophysiology*, 98(2), 351–364. <https://doi.org/10.1016/j.ijpsycho.2015.05.005>

- Ney, L. J., Laing, P. A. F., Steward, T., Zuj, D. V., Dymond, S., & Felmingham, K. L. (2020). Inconsistent analytic strategies reduce robustness in fear extinction via skin conductance response. *Psychophysiology*, *57*(11). <https://doi.org/10.1111/psyp.13650>
- Ney, L. J., Laing, P. A. F., Steward, T., Zuj, D. V., Dymond, S., Harrison, B., Graham, B., & Felmingham, K. L. (2022). Methodological implications of sample size and extinction gradient on the robustness of fear conditioning across different analytic strategies. *PLOS ONE*, *17*(5), e0268814. <https://doi.org/10.1371/journal.pone.0268814>
- Ney, L. J., Wade, M., Reynolds, A., Zuj, D. V., Dymond, S., Matthews, A., & Felmingham, K. L. (2018). Critical evaluation of current data analysis strategies for psychophysiological measures of fear conditioning and extinction in humans. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *134*, 95–107. <https://doi.org/10.1016/j.ijpsycho.2018.10.010>
- Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
- Nieuwenhuys, R. (2012). The insular cortex. In *Progress in Brain Research* (Bd. 195, S. 123–163). Elsevier. <https://doi.org/10.1016/B978-0-444-53860-4.00007-6>
- Noble, S., Scheinost, D., & Constable, R. T. (2021). A guide to the measurement and interpretation of fMRI test-retest reliability. *Current Opinion in Behavioral Sciences*, *40*, 27–32. <https://doi.org/10.1016/j.cobeha.2020.12.012>
- Norrholm, S. D., Vervliet, B., Jovanovic, T., Boshoven, W., Myers, K. M., Davis, M., Rothbaum, B., & Duncan, E. J. (2008). Timing of extinction relative to acquisition: A parametric analysis of fear extinction in humans. *Behavioral Neuroscience*, *122*(5), 1016–1030. <https://doi.org/10.1037/a0012604>
- Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia: I. Opening Scientific Communication. *Psychological Inquiry*, *23*(3), 217–243. <https://doi.org/10.1080/1047840X.2012.692215>
- Nosek, B. A., & Errington, T. M. (2020). The best time to argue about what a replication means? Before you do it. *Nature*, *583*(7817), 518–520. <https://doi.org/10.1038/d41586-020-02142-6>
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. <https://doi.org/10.3758/s13423-019-01645-2>
- Öhman, A. (2009). Human fear conditioning and the amygdala. In P. J. Whalen & E. A. Phelps (Hrsg.), *The human amygdala* (S. 118–154). The Guilford Press.

- Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483–522.
<https://doi.org/10.1037/0033-295X.108.3.483>
- Olatunji, B. O., Davis, M. L., Powers, M. B., & Smits, J. A. J. (2013). Cognitive-behavioral therapy for obsessive-compulsive disorder: A meta-analysis of treatment outcome and moderators. *Journal of Psychiatric Research*, *47*(1), 33–41.
<https://doi.org/10.1016/j.jpsychires.2012.08.020>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- 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.
<https://doi.org/10.1371/journal.pone.0105054>
- Parsons, S. (2020). *Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability* [Preprint]. PsyArXiv.
<https://doi.org/10.31234/osf.io/y6tcz>
- Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395.
<https://doi.org/10.1177/2515245919879695>
- Pavlov, I. P. (1927). *Conditioned reflexed*. Oxford University Press.
- Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction Learning in Humans. *Neuron*, *43*(6), 897–905. <https://doi.org/10.1016/j.neuron.2004.08.042>
- Pineles, S. L., Orr, M. R., & Orr, S. P. (2009). An alternative scoring method for skin conductance responding in a differential fear conditioning paradigm with a long-duration conditioned stimulus. *Psychophysiology*, *46*(5), 984–995. <https://doi.org/10.1111/j.1469-8986.2009.00852.x>
- Pitman, R. K., Orr, S. P., Altman, B., Longpre, R. E., Poiré, R. E., Macklin, M. L., Michaels, M. J., & Steketee, G. S. (1996). Emotional processing and outcome of imaginal flooding therapy in vietnam veterans with chronic posttraumatic stress disorder. *Comprehensive Psychiatry*, *37*(6), 409–418. [https://doi.org/10.1016/S0010-440X\(96\)90024-3](https://doi.org/10.1016/S0010-440X(96)90024-3)
- Plendl, W., & Wotjak, C. T. (2010). Dissociation of within- and between-Session Extinction of Conditioned Fear. *Journal of Neuroscience*, *30*(14), 4990–4998.
<https://doi.org/10.1523/JNEUROSCI.6038-09.2010>
- Pohlack, S. T., Nees, F., Liebscher, C., Cacciaglia, R., Diener, S. J., Ridder, S., Woermann, F. G., & Flor, H. (2012). Hippocampal but not amygdalar volume affects contextual fear conditioning in humans. *Human Brain Mapping*, *33*(2), 478–488.
<https://doi.org/10.1002/hbm.21224>

- Prenoveau, J. M., Craske, M. G., Liao, B., & Ornitz, E. M. (2013). Human fear conditioning and extinction: Timing is everything...or is it? *Biological Psychology*, *92*(1), 59–68. <https://doi.org/10.1016/j.biopsycho.2012.02.005>
- Presentation® software. (2010). *Presentation® software* (14.8). Neurobehavioral Systems, Inc. www.neurobs.com
- Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., & Schooler, J. (2020). *High Replicability of Newly-Discovered Social-behavioral Findings is Achievable* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/n2a9x>
- Rachman, S. (1989). The return of fear: Review and prospect. *Clinical Psychology Review*, *9*(2), 147–168. [https://doi.org/10.1016/0272-7358\(89\)90025-1](https://doi.org/10.1016/0272-7358(89)90025-1)
- Ramirez-Moreno, D. F., & Sejnowski, T. J. (2012). A computational model for the modulation of the prepulse inhibition of the acoustic startle reflex. *Biological Cybernetics*, *106*(3), 169–176. <https://doi.org/10.1007/s00422-012-0485-7>
- Rauch, S. A. M., Foa, E. B., Furr, J. M., & Filip, J. C. (2004). Imagery vividness and perceived anxious arousal in prolonged exposure treatment for PTSD. *Journal of Traumatic Stress*, *17*(6), 461–465. <https://doi.org/10.1007/s10960-004-5794-8>
- Rauch, S. L., Milad, M. R., Orr, S. P., Quinn, B. T., Fischl, B., & Pitman, R. K. (2005). Orbitofrontal thickness, retention of fear extinction, and extraversion. *NeuroReport*, *16*(17), 1909–1912. <https://doi.org/10.1097/01.wnr.0000186599.66243.50>
- R Core Team (various years). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Reinhard, G., Lachnit, H., & Konig, S. (2006). Tracking stimulus processing in Pavlovian pupillary conditioning. *Psychophysiology*, *43*(1), 73–83. <https://doi.org/10.1111/j.1469-8986.2006.00374.x>
- Rescorla, R. A. (1993). Inhibitory associations between S and R in extinction. *Animal Learning & Behavior*, *21*(4), 327–336. <https://doi.org/10.3758/BF03197998>
- Rescorla, R. A. (2001). Experimental extinction. In R. R. Mowrer & S. B. Klein (Hrsg.), *Handbook of contemporary learning theories* (S. 119–154). Erlbaum.
- Ridderbusch, I. C., Wroblewski, A., Yang, Y., Richter, J., Hollandt, M., Hamm, A. O., Wittchen, H.-U., Ströhle, A., Arolt, V., Margraf, J., Lueken, U., Herrmann, M. J., Kircher, T., & Straube, B. (2021). Neural adaptation of cingulate and insular activity during delayed fear extinction: A replicable pattern across assessment sites and repeated measurements. *NeuroImage*, *237*, 118157. <https://doi.org/10.1016/j.neuroimage.2021.118157>
- Riley, W. T., McCormick, M. G. F., Simon, E. M., Stack, K., Pushkin, Y., Overstreet, M. M., Carmona, J. J., & Magakian, C. (1995). Effects of alprazolam dose on the induction and habituation processes during behavioral panic induction treatment. *Journal of Anxiety Disorders*, *9*(3), 217–227. [https://doi.org/10.1016/0887-6185\(95\)00003-7](https://doi.org/10.1016/0887-6185(95)00003-7)

- Rosen, J. B., & Schulkin, J. (1998). From normal fear to pathological anxiety. *Psychological Review*, *105*(2), 325–350. <https://doi.org/10.1037/0033-295X.105.2.325>
- Sánchez-Meca, J., Rosa-Alcázar, A. I., Marín-Martínez, F., & Gómez-Conesa, A. (2010). Psychological treatment of panic disorder with or without agoraphobia: A meta-analysis☆. *Clinical Psychology Review*, *30*(1), 37–50. <https://doi.org/10.1016/j.cpr.2009.08.011>
- Sareen, J., Cox, B. J., Clara, I., & Asmundson, G. J. G. (2005). The relationship between anxiety disorders and physical disorders in the U.S. National Comorbidity Survey. *Depression and Anxiety*, *21*(4), 193–202. <https://doi.org/10.1002/da.20072>
- Scharfenort, R., & Lonsdorf, T. B. (2016). Neural correlates of and processes underlying generalized and differential return of fear. *Social Cognitive and Affective Neuroscience*, *11*(4), 612–620. <https://doi.org/10.1093/scan/nsv142>
- Scharfenort, R., Menz, M., & Lonsdorf, T. B. (2016). Adversity-induced relapse of fear: Neural mechanisms and implications for relapse prevention from a study on experimentally induced return-of-fear following fear conditioning and extinction. *Translational Psychiatry*, *6*, e858. <https://doi.org/10.1038/tp.2016.126>
- 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. *Behaviour Research and Therapy*, *86*, 87–94. <https://doi.org/10.1016/j.brat.2016.08.015>
- Schmajuk, N. A., & Holland, P. C. (Hrsg.). (1998). *Occasion setting: Associative learning and cognition in animals*. American Psychological Association. <https://doi.org/10.1037/10298-000>
- Schmidt, S. (2009). Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology*, *13*(2), 90–100. <https://doi.org/10.1037/a0015108>
- Schulz, P., Schlotz, W., & Becker, P. (2004). *Trierer inventar zum chronischen stress (TICS)[Trier inventory for chronic stress (TICS)]*.
- Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. *J. Weinman, S. Wright, & M. Johnston, Measures in health psychology: A user's portfolio. Causal and control beliefs*, *35*, 37.
- Schwarzer, R., & Schulz, U. (2003). Soziale unterstützung bei der krankheitsbewältigung: Die berliner social support skalen (BSSS). *Diagnostica*, *49*(2), 73–82.
- Shumake, J., Fergusson-Moreira, S., & Monfils, M. H. (2014). Predictability and heritability of individual differences in fear learning. *Animal Cognition*, *17*(5), 1207–1221. <https://doi.org/10.1007/s10071-014-0752-1>
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., ... Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. <https://doi.org/10.1177/2515245917747646>

- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214. <https://doi.org/10.1038/s41562-020-0912-z>
- Sjouwerman, R., Illius, S., Kuhn, M., & Lonsdorf, T. B. (2022). A data multiverse analysis investigating non-model based SCR quantification approaches. *Psychophysiology*. <https://doi.org/10.1111/psyp.14130>
- Sjouwerman, R., & Lonsdorf, T. B. (2019). Latency of skin conductance responses across stimulus modalities. *Psychophysiology*, 56(4), e13307. <https://doi.org/10.1111/psyp.13307>
- Sjouwerman, R., & Lonsdorf, T. B. (2020). Experimental boundary conditions of reinstatement-induced return of fear in humans: Is reinstatement in humans what we think it is? *Psychophysiology*, 57(5). <https://doi.org/10.1111/psyp.13549>
- 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: Startle probe and rating: Impact on fear learning. *Psychophysiology*, 53(12), 1889–1899. <https://doi.org/10.1111/psyp.12761>
- Sjouwerman, R., Scharfenort, R., & Lonsdorf, T. B. (2020). Individual differences in fear acquisition: Multivariate analyses of different emotional negativity scales, physiological responding, subjective measures, and neural activation. *Scientific Reports*, 10(1), 15283. <https://doi.org/10.1038/s41598-020-72007-5>
- Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, 25(6), 2083–2101. <https://doi.org/10.3758/s13423-018-1451-8>
- Soto, C. J. (2019). How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project. *Psychological Science*, 30(5), 711–727. <https://doi.org/10.1177/0956797619831612>
- Spearman, C. (1910). CORRELATION CALCULATED FROM FAULTY DATA. *British Journal of Psychology*, 1904-1920, 3(3), 271–295. <https://doi.org/10.1111/j.2044-8295.1910.tb00206.x>
- Spielberger, C. D., Gorsuch, R. L., Lushene, R. E., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Consulting Psychologists Press.
- SPM12. Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology. <https://www.fil.ion.ucl.ac.uk/spm/>
- Staples-Bradley, L. K., Treanor, M., & Craske, M. G. (2018). Discrimination between safe and unsafe stimuli mediates the relationship between trait anxiety and return of fear. *Cognition and Emotion*, 32(1), 167–173. <https://doi.org/10.1080/02699931.2016.1265485>
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science: A Journal of the*

- Association for Psychological Science*, 11(5), 702–712.
<https://doi.org/10.1177/1745691616658637>
- Stegmann, Y., Andreatta, M., & Wieser, M. J. (2022). The effect of inherently threatening contexts on visuocortical engagement to conditioned threat. *Psychophysiology*.
<https://doi.org/10.1111/psyp.14208>
- Stroebe, W., & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science*, 9(1), 59–71.
<https://doi.org/10.1177/1745691613514450>
- Torrents-Rodas, D., Fullana, M. A., Bonillo, A., Andi3n, O., Molinuevo, B., Caseras, X., & Torrubia, R. (2014). Testing the temporal stability of individual differences in the acquisition and generalization of fear: Stability acquisition and generalization of fear. *Psychophysiology*, 51(7), 697–705. <https://doi.org/10.1111/psyp.12213>
- Torrents-Rodas, D., Fullana, M. A., Bonillo, A., Caseras, X., Andi3n, O., & Torrubia, R. (2013). No effect of trait anxiety on differential fear conditioning or fear generalization. *Biological Psychology*, 92(2), 185–190. <https://doi.org/10.1016/j.biopsycho.2012.10.006>
- Tzovara, A., Korn, C. W., & Bach, D. R. (2018). Human Pavlovian fear conditioning conforms to probabilistic learning. *PLOS Computational Biology*, 14(8), e1006243.
<https://doi.org/10.1371/journal.pcbi.1006243>
- Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2020). *Credibility Beyond Replicability: Improving the Four Validities in Psychological Science* [Preprint]. PsyArXiv.
<https://doi.org/10.31234/osf.io/bu4d3>
- Vervliet, B., Craske, M. G., & Hermans, D. (2013). Fear extinction and relapse: State of the art. *Annual Review of Clinical Psychology*, 9, 215–248. <https://doi.org/10.1146/annurev-clinpsy-050212-185542>
- Visser, R. M., Bathelt, J., Scholte, H. S., & Kindt, M. (2021). Robust BOLD Responses to Faces But Not to Conditioned Threat: Challenging the Amygdala’s Reputation in Human Fear and Extinction Learning. *The Journal of Neuroscience*, 41(50), 10278–10292.
<https://doi.org/10.1523/JNEUROSCI.0857-21.2021>
- Vrana, S. R., Spence, E. L., & Lang, P. J. (1988). The startle probe response: A new measure of emotion? *Journal of Abnormal Psychology*, 97(4), 487–491. <https://doi.org/10.1037/0021-843X.97.4.487>
- Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLOS Biology*, 13(4), e1002128.
<https://doi.org/10.1371/journal.pbio.1002128>
- Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., Savic, M. D., Garovic, V. D., & Milic, N. M. (2019). Reveal, Don’t Conceal: Transforming Data Visualization to Improve Transparency. *Circulation*, 140(18), 1506–1518. <https://doi.org/10.1161/CIRCULATIONAHA.118.037777>
- Westbrook, R. F., Iordanova, M., McNally, G., Richardson, R., & Harris, J. A. (2002). Reinstatement of fear to an extinguished conditioned stimulus: Two roles for context.

- Journal of Experimental Psychology: Animal Behavior Processes*, 28(1), 97–110.
<https://doi.org/10.1037/0097-7403.28.1.97>
- Winkelmann, T., Grimm, O., Pohlack, S. T., Nees, F., Cacciaglia, R., Dinu-Biringer, R., Steiger, F., Wicking, M., Ruttorf, M., Schad, L. R., & Flor, H. (2016). Brain morphology correlates of interindividual differences in conditioned fear acquisition and extinction learning. *Brain Structure & Function*, 221(4), 1927–1937. <https://doi.org/10.1007/s00429-015-1013-z>
- Wolitzky-Taylor, K. B., Horowitz, J. D., Powers, M. B., & Telch, M. J. (2008). Psychological approaches in the treatment of specific phobias: A meta-analysis. *Clinical Psychology Review*, 28(6), 1021–1037. <https://doi.org/10.1016/j.cpr.2008.02.007>
- Wroblewski, A., Hollandt, M., Yang, Y., Ridderbusch, I. C., Pietzner, A., Szeska, C., Lotze, M., Wittchen, H.-U., Heinig, I., Pittig, A., Arolt, V., Koelkebeck, K., Rothkopf, C. A., Adolph, D., Margraf, J., Lueken, U., Pauli, P., Herrmann, M. J., Winkler, M. H., ... Richter, J. (2022). Sometimes I feel the fear of uncertainty: How intolerance of uncertainty and trait anxiety impact fear acquisition, extinction and the return of fear. *International Journal of Psychophysiology*, 181, 125–140.
<https://doi.org/10.1016/j.ijpsycho.2022.09.001>
- Yonkers, K. A., Bruce, S. E., Dyck, I. R., & Keller, M. B. (2003). Chronicity, relapse, and illness?course of panic disorder, social phobia, and generalized anxiety disorder: Findings in men and women from 8 years of follow-up. *Depression and Anxiety*, 17(3), 173–179.
<https://doi.org/10.1002/da.10106>
- Zeidan, M. A., Lebron-Milad, K., Thompson-Hollands, J., Im, J. J. Y., Dougherty, D. D., Holt, D. J., Orr, S. P., & Milad, M. R. (2012). Test–Retest Reliability during Fear Acquisition and Fear Extinction in Humans. *CNS Neuroscience & Therapeutics*, 18(4), 313–317.
<https://doi.org/10.1111/j.1755-5949.2011.00238.x>
- Zorowitz, S., & Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, S2451902223000423. <https://doi.org/10.1016/j.bpsc.2023.02.004>
- Zuo, X.-N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature Human Behaviour*, 3(8), 768–771. <https://doi.org/10.1038/s41562-019-0655-x>

7 Study I

This article was published in *Scientific Reports*, 10, Ehlers, M. R., Nold, J., Kuhn, M., Klingelhöfer-Jens, M., & Lonsdorf, T. B., Revisiting potential associations between brain morphology, fear acquisition and extinction through new data and a literature review, 19894, License: CC BY 4.0 DEED, <https://creativecommons.org/licenses/by/4.0/>, no changes have been implemented, Springer Nature (2020).



OPEN

Revisiting potential associations between brain morphology, fear acquisition and extinction through new data and a literature review

Mana R. Ehlers¹✉, Janne Nold¹, Manuel Kuhn^{1,2}, Maren Klingelhöfer-Jens¹ & Tina B. Lonsdorf¹

Inter-individual differences in defensive responding are widely established but their morphological correlates in humans have not been investigated exhaustively. Previous studies reported associations with cortical thickness of the dorsal anterior cingulate cortex, insula and medial orbitofrontal cortex as well as amygdala volume in fear conditioning studies. However, these associations are partly inconsistent and often derived from small samples. The current study aimed to replicate previously reported associations between physiological and subjective measures of fear acquisition and extinction and brain morphology. Structural magnetic resonance imaging was performed on 107 healthy adults who completed a differential cued fear conditioning paradigm with 24 h delayed extinction while skin conductance response (SCR) and fear ratings were recorded. Cortical thickness and subcortical volume were obtained using the software Freesurfer. Results obtained by traditional null hypothesis significance testing and Bayesian statistics do not support structural brain-behavior relationships: Neither differential SCR nor fear ratings during fear acquisition or extinction training could be predicted by cortical thickness or subcortical volume in regions previously reported. In summary, the current pre-registered study does not corroborate associations between brain morphology and inter-individual differences in defensive responding but differences in experimental design and analyses approaches compared to previous work should be acknowledged.

Marked inter-individual differences in defensive responding have been suggested to be the result of underlying neurobiological differences that manifest as stable trait-like characteristics (rodents¹, humans²). Defensive conditioned responding can be investigated in the laboratory by means of fear conditioning protocols.

Generally, the fear conditioning procedure comprises different experimental phases²: throughout acquisition training an innately aversive stimulus, the unconditioned stimulus (US), is paired with an initially neutral stimulus, the conditioned stimulus (CS+), producing a conditioned response (CR) to the CS+ while a second control stimulus (CS−) is never paired with the US. Hence, a fear memory is formed as the CS+ gains predictive power of the appearance of a US and comes to elicit a defensive conditioned response by itself. In the laboratory, different outcome measures such as skin conductance response (SCR), fear potentiated startle (FPS), ratings of fear and/or US expectancy as well as BOLD fMRI can be used as proxies thereof. The difference in responding to the CS+ and the CS− (i.e., CS discrimination) is taken as an approximation for the strength of fear learning. During extinction training, the CS+ is no longer coupled with the US and a plethora of results suggest that an inhibitory extinction memory is formed as a consequence³. As a result, conditioned responding is reduced. When at a later time exposed to the CS+ (i.e., ‘retention test’ or ‘return of fear test’, for an overview see³) one can either observe a ‘retention of the extinction memory’ indicating dominance of the extinction memory or the return of conditioned responding (i.e., ‘return of fear’) indicating dominance of the fear over the extinction memory.

¹Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Martinistrasse 52, W34, 20246 Hamburg, Germany. ²Department of Psychiatry, Harvard Medical School, and Center for Depression, Anxiety and Stress Research, McLean Hospital, Belmont, MA 02478, USA. ✉email: ma.ehlers@uke.de

While the basic mechanisms of fear conditioning and extinction and the importance of inter-individual differences in defensive responding are well recognized, research concerning a potential mapping of such inter-individual differences onto variability in brain morphology is sparse. Structural-brain-behavior associations (i.e., associations between inter-individual variability in brain morphology and behavior or physiology) have a long history in psychology and neuroscience^{4,5}. In *in vivo* human studies inter-individual variability in brain structure is commonly extracted from anatomical scans acquired through magnetic resonance imaging (MRI). The most common methods include measures of grey matter tissue such as grey matter volume using the Computational Anatomy Toolbox⁶ and measures of cortical thickness and subcortical volume using the software FreeSurfer^{7–10}. Yet, structural-brain-behavior associations were recently scrutinized as it was shown in a large sample of healthy adults that significant associations are rare and also show low replication rates across a range of psychological measures^{11–13}.

Previous work in fear conditioning research has reported individual differences in brain morphology to be associated with differences in conditioned responding during fear and extinction learning as well as retention of extinction. Most of these studies have focused on skin conductance response (SCR) while fewer studies investigated associations with ratings of valence, arousal or CS–US contingency awareness^{14–16} and a single study with fear potentiated startle¹⁷ (see Table 1). Of note, all areas that have been reported to show structural associations with inter-individual differences in defensive responding during fear and/or extinction learning have been linked to group averages in functional brain activation as assessed by BOLD fMRI during learning and expression of fear and extinction^{18,19}; the amygdala, insula and prefrontal areas (dorsal anterior cingulate cortex (dACC), and medial orbitofrontal cortex (mOFC) in cue conditioning as well as the hippocampus in context conditioning.

Morphological variability within the amygdala has been positively related to average differential responding during acquisition training in SCR [(CS+) – (CS–)] but not ratings of arousal and valence¹⁶ or CS–US contingency^{14,16}. More precisely, this association was reported for the volume of the *right* amygdala during early acquisition in sample 1 but during late acquisition in sample 2 despite the identical experimental protocol¹⁶ while a smaller earlier study with a largely overlapping sample reported a positive correlation with *left* amygdala volume in early but not late acquisition¹⁴. These discrepancies might be explained by differences in analyses such as segmentation approaches, different SCR quantification approaches, different scoring criteria for SCR as well as the inclusion of a large number of covariates as well as correction of differential SCR responding by responding during preceding experimental phases (see Table 1). In contrast to these studies, others did not find significant positive associations between differential [(CS+) – (CS–)] autonomic conditioned responding during acquisition training and amygdala volume but report an insignificant *negative* relationship in two small samples for both right and left amygdala²⁰. It should be noted however that due to different numbers of trials included in these studies (see Table 1), the full acquisition phase in this study, corresponds largely to the first half of acquisition training in the studies by Cacciaglia et al.¹⁴ and Winkelmann et al.¹⁶.

In addition to the amygdala, the volume in the right posterior insula/posterior operculum was reported to show a positive association with differential SCR during fear acquisition training in two samples—although this did not survive correction for multiple comparisons in the smaller sample²⁰.

Furthermore, SCR to the CS+, but not to the CS– or differential SCR responding during acquisition training were reported to be *positively* correlated with thickness of the dorsal anterior cingulate cortex (dACC)²¹ in a 100% reinforcement protocol which was, however, not replicated in two samples in a study employing partial reinforcement²⁰. A recent study¹⁷ with a large sample of anxiety patients and healthy controls (N = 351) including children and adults, reported the dorsomedial/dorsolateral prefrontal cortex (dm/dlPFC)—a region located substantially more lateral than the area identified by Milad et al.²¹—to be *negatively* correlated with a measure of general SCR averaged across both CS types (i.e., CS+ and CS–) and experimental phases (i.e., fear acquisition and extinction training). The interpretation of this aggregate SCR measure, however, is not straightforward with respect to associative learning processes. While Abend et al. interpret the observed association in terms of aberrant threat and safety learning, alternatively this aggregate measure may as well reflect general arousal or the reactivity in SCR independent of associative learning processes. Future studies employing experimental paradigms to capture generalization of fear may clarify whether the association reported by Abend et al. could also be interpreted in terms of fear generalization. In another publication, computational modeling was applied to SCR to the CS+ which reveal that learning rate correlates positively with cortical thickness of the ventromedial prefrontal cortex (vmPFC), dACC and anterior insula²². In addition to these findings from cue-conditioning studies, a positive association between total hippocampal but not amygdala volume and differential second interval but not first interval SCR, see²³ was reported during context conditioning¹⁵. No significant associations were observed with SCR during extinction¹⁵. In this study, CS–US contingency awareness showed a relationship with total brain volume, but not hippocampal or amygdala volume. Another study from the same research group reported an association between bilateral hippocampus volume and differential CS–US contingency ratings in late but not early acquisition training in a cue-conditioning paradigm—which seemed to be driven by a negative correlation with ratings to the CS–¹⁴. Differences in results might be attributable to the fact that the studies used a contextual and a cued fear conditioning paradigm respectively.

While the work summarized above focused on acquisition training, some studies have also investigated structural brain-behavior associations during extinction training and extinction retention. During immediate, early but not late extinction training, differential [(CS+) – (CS–)] SCR was correlated with the thickness of three clusters of the right vmPFC¹⁶. Notably, however, earlier studies^{20,24} did not test for any associations between prefrontal thickness and differential SCR during extinction training as they focused on 24 h extinction retention.

During a 24-delayed retention test (also often referred to as ‘extinction recall’, for a discussion on terminology see⁴), a positive correlation between a non-differential (i.e., CS+ specific) ‘extinction retention index’ in SCR and thickness of the medial OFC was observed when tested in the extinction context (i.e. ‘extinction retention’²³) as well as in the mOFC portion of the vmPFC when tested in both the acquisition (i.e., ‘renewal’)

References	N	Segmentation approach	RIR (%)	Extinction	# of Acq trials for CS+/CS-	# of Ext trials for CS+/CS-	Outcome measures		Tested associations with				SCR quantification via	SCR scoring criteria; CS duration	Covariates
							SCR	Fear rating	CSdiff	CS+	CS-	CSavg			
Abend et al., 2019 ¹⁷	250	Freesurfer	80	Immediate	10/10	8/8	✓	✓	✗	✗	✗	✓	TTP	0–5 s post CS onset; 7 s CS	Age, anxiety
Abend et al., 2020 ²²	351	Freesurfer	80	N/A	10/10	8/8	✓	✓	✗	✓ ^a	✗	✗	TTP	1–5 s post CS onset; 7 s CS	Age, anxiety
Cacciaglia et al., 2013 ¹⁴	52	Manual	50	Immediate	36/36	18/18	✓ ✓	✓	✓	✓	✓	✗	TTP	1–9 s post CS onset; 6 s CS	Age, gender, anxiety, education
Ehlers et al. (current study)	107	Freesurfer	100	Delayed	14/14	14/14	✓	✓	✓	✓	✓	✗	TTP	0.9–3.5 s post CS onset; 6 s CS	TIV, sex
Hartley et al., 2011 ²⁰	18	Freesurfer	17	N/A	21/15;	N/A	✓	✗	✓	✗	✗	✗	TTP	0.5–4.5 s post CS onset; 4 s CS	Sex, anxiety
	12	Freesurfer	35	Immediate	23/15	15/15	✓	✗	✓	✗	✗	✗	TTP	0.5–4.5 s post CS onset; 4 s CS	Sex, anxiety
Milad et al., 2005 ²⁴	14	Freesurfer	100	Immediate	5/5	10/10	✓	✗	✓	✓	✓	✗	b.c.	Max (12 s post CS onset)-mean (2 s pre CS onset); 12 s CS	N/A
Milad et al., 2007 ²¹	14	Freesurfer	100	Immediate	5/5	10/10	✓	✗	✓	✗	✗	✗	b.c.	Max (12 s post CS onset)-mean (2 s pre CS onset); 12 s CS	N/A
Rauch et al., 2005 ²⁵	14	Freesurfer	100	Immediate	5/5	10/10	✓	✗	✗	✓	✗	✗	b.c.	Max (12 s post CS onset)-mean (2 s pre CS onset); 12 s CS	Sex, extraversion, neuroticism
Winkelman et al., 2015 ¹⁶	68; 53	Freesurfer	50	Immediate	36/36	18/18	✓	✓	✓	✗	✗	✗	Ledalab	Sum (SCRs 1–7 s post CS onset); 6 s CS	TIV, age, gender

Table 1. Experimental design overview of studies investigating associations between brain morphology and associative processes during fear acquisition training and extinction in human participants. Two studies (Abend et al. 2019, Abend et al. 2020) that did not investigate associative processes during fear acquisition training but average responding to the CS+ and CS– across experimental phases are included for completeness. None of the studies explicitly instructed the participants with regard to the CS/US contingencies, Abend et al. (2019) and Hartley et al. (2011), however, informed participants about the fact that association can be learning during the experiment. *RIR* reinforcement rate, *N/A* information not available, *CSdiff* differential SCR [(CS+) – (CS–)], *CSavg* SCR averaged across the CS+ and CS– as well as across fear acquisition and extinction training, *TTP* trough to peak, *b.c.* baseline correction, *TIV* total intracranial volume. ^aIn Abend et al. (2020) computational modeling of SCR to the CS+ was used to predict SCR over the course of learning and assess learning rate during acquisition and extinction.

and extinction context (i.e. ‘extinction retention’²²). Another study reported a positive correlation between conditioned responding during a retention test and thickness of the vmPFC at a very lenient statistical threshold of $p < 0.003$ uncorrected following extinction training but not following cognitive regulation²⁰. These studies^{20,24,25} quantified (extinction) retention in SCR through versions of the non-differential (i.e., CS+ specific) “Extinction retention index (ERI)”, which has recently been challenged from both a theoretical and empirical perspective for lacking construct validity: More specifically, non-differential ERIs likely measure general arousal or orienting responding rather than associative processes such as the retention of extinction memory²⁶. As these studies did not investigate brain morphological associations during the preceding extinction training phase, the specificity of the findings pertaining to the retention test phase also remains unclear—in particular given the reported associations between volume in ventromedial prefrontal areas and differential SCR responding during extinction training itself—which always precedes a retention test.

To date, only a limited number of studies has linked inter-individual differences in brain morphology in areas known to be generally implicated in fear acquisition, extinction and extinction recall to conditioned autonomic (SCR) and subjective (valence and arousal ratings, fear ratings and CS–US awareness) measures of defensive responding.

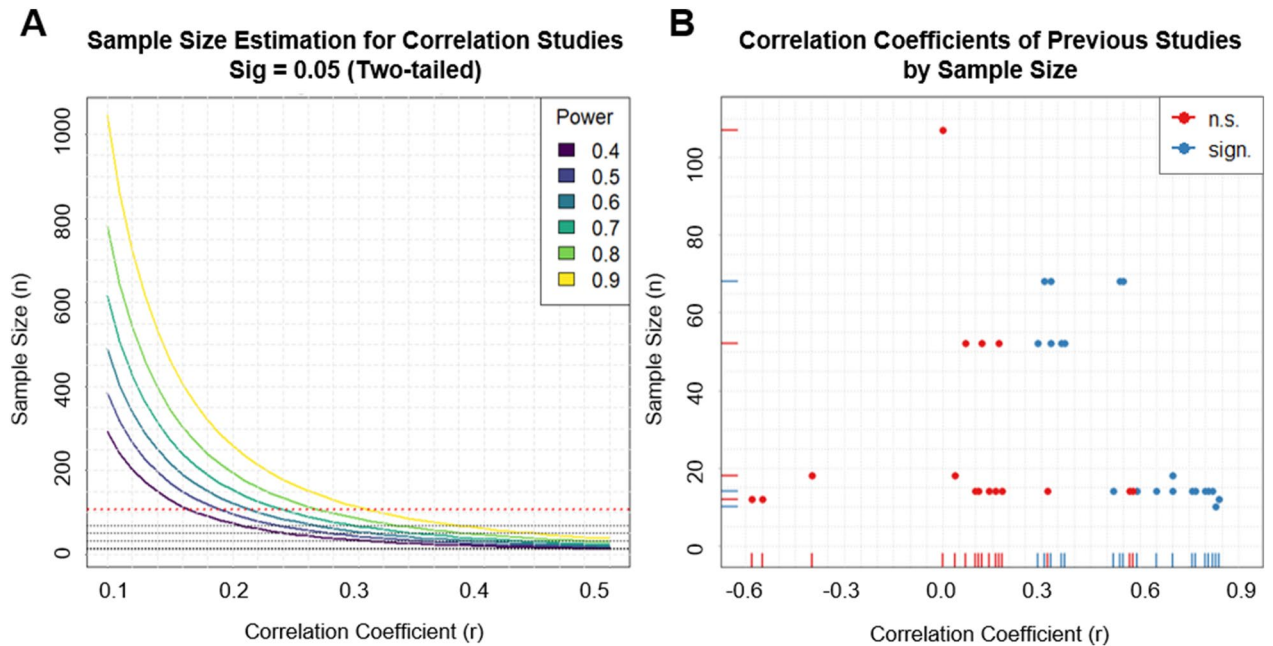


Figure 1. Illustration of (A) estimated power of correlation studies given a certain sample size. Dotted grey lines indicate sample sizes of previous studies investigating differential or CS-specific associations with brain morphological measures. Note that in some previous studies regressions were performed rather than correlations. For comparability all effects size measures were transformed to correlation coefficients. The red line indicates the sample size of the current study. (B) Effect sizes expressed as correlation coefficients obtained in previous studies and the current study plotted by sample size. Note that in some previous studies regressions were performed instead of correlation. For comparability all effects size measures were transformed to correlation coefficients. Red dots indicate non-significant and blue dots indicate significant findings. Note that some studies report more than one association and are hence represented with multiple dots.

From the perspective of the research standards in 2020, particularly the early studies report rather implausibly high correlation coefficients (illustrated in Fig. 1A) and partly employ (very) lenient statistical thresholds originating from what now has to be considered massively underpowered sample sizes^{27,28} (see Fig. 1B for a power curve plot) with only 10–18 participants^{20,21,24,25} (see Table 1).

While structural MRI measures themselves have been shown to have excellent reliability^{28,29} (minimal test–retest reliability 0.82³⁰), the robustness of structural brain-behavior associations in general has been challenged recently^{5,13} and given this, the aim of the current pre-registered study (<https://doi.org/10.17605/osf.io/y73qw>) is to replicate previously reported associations between individual differences in brain morphology and physiological (i.e., SCR) and subjective (i.e., fear ratings) measures of defensive responding during fear acquisition and delayed extinction in a larger sample of healthy adults ($N = 107$). More precisely, we aim to investigate previously reported associations between the cortical thickness of the dACC and insula as well as amygdala volume during acquisition training and the association between amygdala volume and mOFC thickness and extinction.

Results

Main effects of task. Successful acquisition training is reflected in significantly larger average SCR (see Fig. 2A) elicited by the CS+ than those elicited by the CS– ($t(106) = 12.81, p < 0.001, 95\% \text{ CI } [0.11, 0.15]$). Similarly, ratings of fear, anxiety and tension (see Fig. 2B) were significantly higher to the CS+ relative to the CS– after acquisition training ($t(102) = 19.74, p < 0.001, 95\% \text{ CI } [13.08, 16.00]$).

During extinction, the CS+, on average, still elicited larger SCR (see Fig. 2A) prior to extinction as compared to the CS– ($t(106) = 3.94, p < 0.001, 95\% \text{ CI } [0.01, 0.03]$). At the end of extinction, however (last five trials in SCR for both CS types), SCR elicited by the CS+ and CS– did not differ significantly ($t(106) = 1.57, p = 0.12, 95\% \text{ CI } [-0.003, 0.024]$).

For extinction, a two-way ANOVA for fear ratings revealed a main effect of time (pre vs post extinction) ($F(1, 412) = 25.06, p < 0.001$), a main effect of CS type ($F(1, 412) = 108.65, p < 0.001$) as well as a significant interaction ($F(1, 412) = 37.45, p < 0.001$). Pairwise comparisons showed that the CS+ elicited higher ratings relative to the CS– prior to extinction ($ps < 0.001$) as well post extinction ($ps = 0.010$). Extinction success indicated by fear ratings is however supported by the observation that ratings of the CS+ dropped significantly from pre to post extinction ($ps < 0.001$), but not for the CS– ($ps = 0.892$).

Inter-hemispheric differences in cortical thickness and volume. Significant differences between volumina and cortical thickness in left and right hemisphere were observed for most regions: dACC ($t(106) = 4.80$,

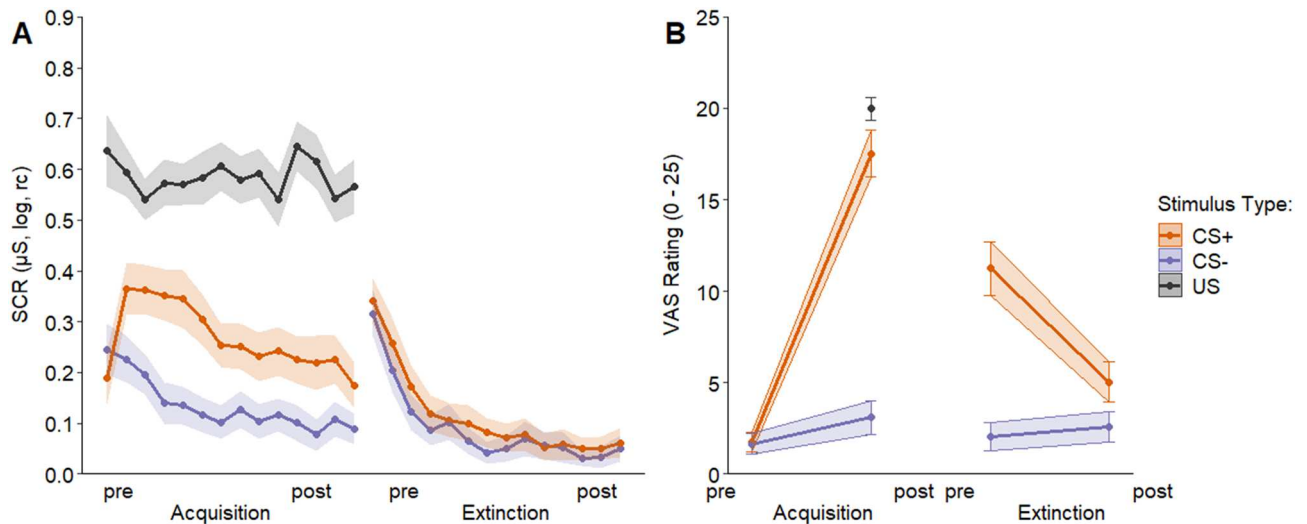


Figure 2. (A) SCR to the CS+ as compared to the CS– as well as the US during acquisition and extinction training (illustrated trial-by-trial) as well as (B) fear ratings in response to CS+ and CS– prior to and after fear acquisition training and extinction as well as aversiveness ratings to the US after acquisition training. 95% confidence intervals are illustrated by coloured bands. Note that a linear learning process is not assumed, the lines are meant to facilitate the visualization of the general trend from pre to post acquisition and extinction.

$p < 0.001$, $d = 0.46$), mOFC ($t(106) = -5.05$, $p < 0.001$, $d = 0.49$) and amygdala ($t(106) = -14.89$, $p < 0.001$, $d = 1.44$) except for the insula ($t(106) = 0.97$, $p = 0.33$, $d = 0.10$) (see Fig. 3). Robustness analyses performing the main analyses reported here (see below) separately for the left and right hemisphere yielded comparable results (see Supplementary Material Section 2.1 for details).

No association between brain morphology and indices of fear learning during acquisition and extinction training.

Our analyses did not replicate previous reports of a significant positive association between the cortical thickness of the dACC and subcortical volume of the amygdala during fear acquisition training as assessed through mean differential SCR and post-acquisition differential fear ratings (see Fig. 4). This was true either when considering the full acquisition phase or the first and second half of acquisition training separately (see Fig. 5A,B; Table 2). Similarly, our additional non-preregistered analyses aiming to replicate previous findings did not provide evidence for a significant association between SCR to the CS+ and CS– separately and thickness of the dACC (see Supplementary Material Section 3.1) or between differential SCR or post-acquisition fear ratings and thickness of the insula (see Supplementary Material Section 3.2). Likewise, no significant association was observed between amygdala volume or mOFC thickness and differential SCR (full phase, first and second half, see Figs. 5C,D, 6A,C) as well as differential ratings during extinction ([pre-post extinction ratings], pre ratings, post ratings, see Figs. 6B,D, 7) see Table 2). For robustness, we checked whether the exclusion of outliers (> 3 SD below or above mean), in fear ratings or SCR affects the results. The analyses were rerun after excluding one participant based on post-acquisition fear ratings, one based on pre-post extinction fear ratings and four based on differential SCR during extinction. The pattern of results remains comparable after excluding outliers for the respective analyses, i.e. all results remained non-significant. For full results see Supplementary Table 4.

In addition to traditional null hypothesis significance testing (NHST), we computed Bayes factors in order to obtain relative evidence against a significant relationship between brain morphology and indices of fear learning. The calculated Bayes factors indicate moderate to strong evidence ($BF_{01} > 3$) for the null or intercept-only model. For one of the tested regression models, only weak evidence for the null model was found ($BF_{01} = 1.99$). Overall, these results demonstrate that there is little reason to believe that morphology in these regions is a significant predictor of conditioned responding during acquisition or extinction training in our study.

Robustness analyses considering data derived from the left and right hemisphere separately (Supplementary Section 2.1, Supplementary Table 1), without the pre-registered covariates (see Supplementary Material Section 2.2, Supplementary Table 2), with raw instead of log-transformed and range corrected SCR scores (see Supplementary Material Section 2.3, Supplementary Table 3) and after removing outliers (see Supplementary Material Section 2.4, Supplementary Table 4) yielded comparable results for both acquisition and extinction training.

Contingency awareness does not moderate a putative association of dACC thickness and fear learning proxies.

Analyses did not confirm our pre-registered exploratory hypothesis of a significant moderation of a putative association between dACC thickness and fear learning proxies (differential SCR and differential subjective fear ratings) during fear acquisition training by contingency awareness (aware, unaware, uncertain). The analysis unsurprisingly revealed that awareness is a significant predictor for differential SCR and fear ratings during fear acquisition training (SCR: $\beta = 0.08$, $p = 0.01$, ratings: $\beta = 4.45$, $p = 0.02$) which, however,

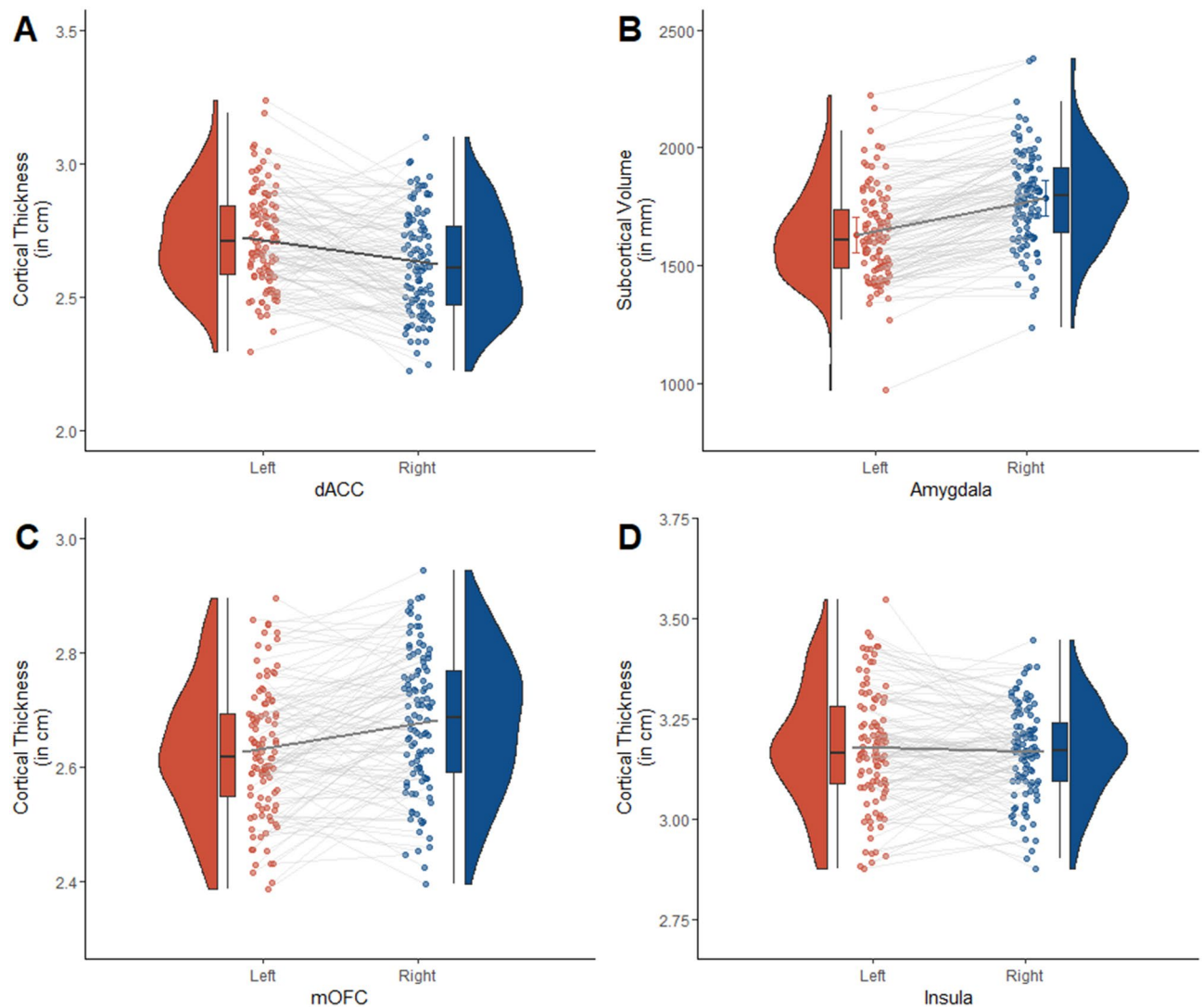


Figure 3. Illustration of (A) cortical thickness of the dACC, (B) volume of the amygdala, (C) cortical thickness of mOFC, and (D) cortical thickness of the insula in left (red) and right (blue) hemisphere. Data are illustrated by smoothed density distributions, individual subject means (dots), medians (boxplots) and interquartile ranges (boxes depict interquartile range and whiskers depict $1.5 \times$ the interquartile range) for each hemisphere with data points derived from a single individual connected through grey lines.

does not moderate the putative relationship between dACC thickness and fear learning proxies (for full results see Supplementary Material Section 1.1).

Aiming to replicate previous associations between amygdala volume and trait anxiety. Non-preregistered analyses in the current sample did not replicate previous reports of a significant association between amygdala volume—considering averaged values as well as left and right hemisphere separately—and trait or state anxiety as measured prior to acquisition training or prior to extinction training (for full results see Supplementary Material Section 3.3).

Discussion

Research regarding a potential association between physiological and subjective measures of conditioned responding during acquisition and/or extinction and its retention and inter-individual differences in brain morphology is sparse to date and most of the few inconsistent results originate from early studies in extremely small samples.

Here, we attempted to (conceptually) replicate these previous findings in a large sample. Our results do not provide support for structural brain-behavior relationships during fear acquisition training and extinction. More precisely, we did not replicate previously reported significant associations between differential SCR or fear ratings and cortical thickness of the dACC, the insula or volume of the amygdala during fear acquisition training or between amygdala volume and mOFC thickness during extinction. Bayes factors provide moderate to strong

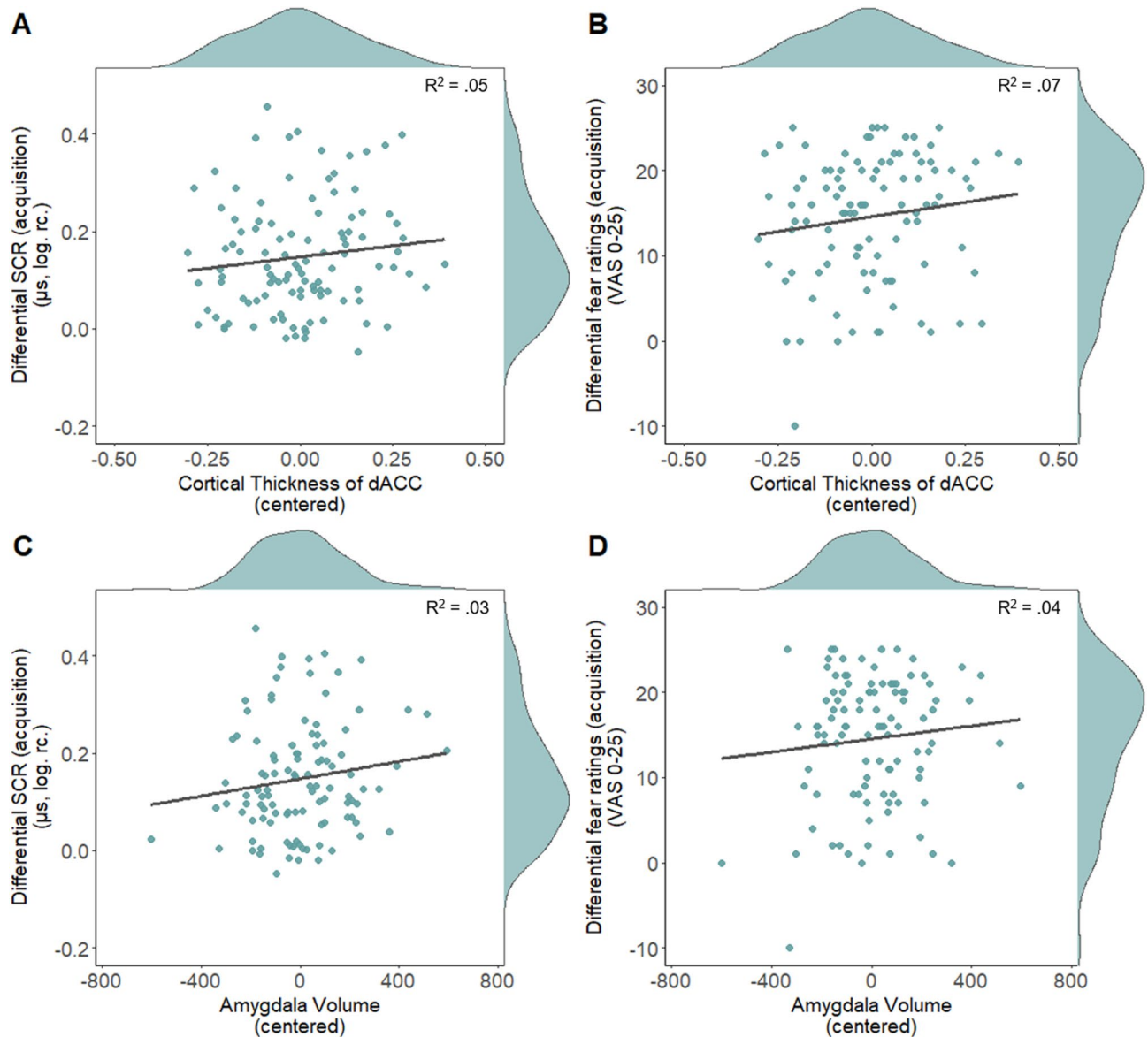


Figure 4. Scatterplots with marginal densities illustrating the (absence of) associations between average differential SCR [(CS+) – (CS–)] during acquisition training and (A) cortical thickness of the dACC and (B) the amygdala as well as between differential post acquisition fear ratings and (C) the dACC and (D) the amygdala.

evidence against a relationship between brain morphology in these regions and physiological or subjective measures of conditioned responding during acquisition or extinction training. Yet, it should be acknowledged that we do not provide a formal close or direct replication as we tested these previously reported associations in a fear conditioning paradigm in which procedural features differ from those in previous work in several ways (i.e., conceptual replication) (see Table 1): the reinforcement ratio, the number of trials, immediate vs. delayed extinction as well as measurement procedures to quantify SCR and estimates of brain morphology.

In more detail, the current study employed a 100% reinforcement rate while many previous studies used partial reinforcement (17–80%)^{14–17,20} and only three studies also employed a 100% reinforcement rate^{20,22,23}. The probability with which the CS+ is coupled with the US during fear acquisition training contributes to the speed of fear acquisition and subsequent extinction learning, with partial reinforcement slowing both the development of conditioned responding and extinction learning^{4,31,32}. It has also been argued that partial reinforcement rate may promote the manifestation of individual differences as opposed to the ‘strong experimental situation’ induced by 100% reinforcement rate³², but this idea still needs to be tested empirically.

Also the number of trials included in the experimental phases differs substantially among previous work (range acquisition: 5–36 trials; range extinction: 8–18 trials per CS type, see Table 1) which renders classifications into ‘first half’ and ‘second half’ inherently ambiguous and difficult to interpret across studies without considering procedural specifications.

In the current study, we aimed to replicate a positive association between amygdala volume and differential SCR observed during the first but not the second half of acquisition training¹⁴. Critically however, Cacciaglia

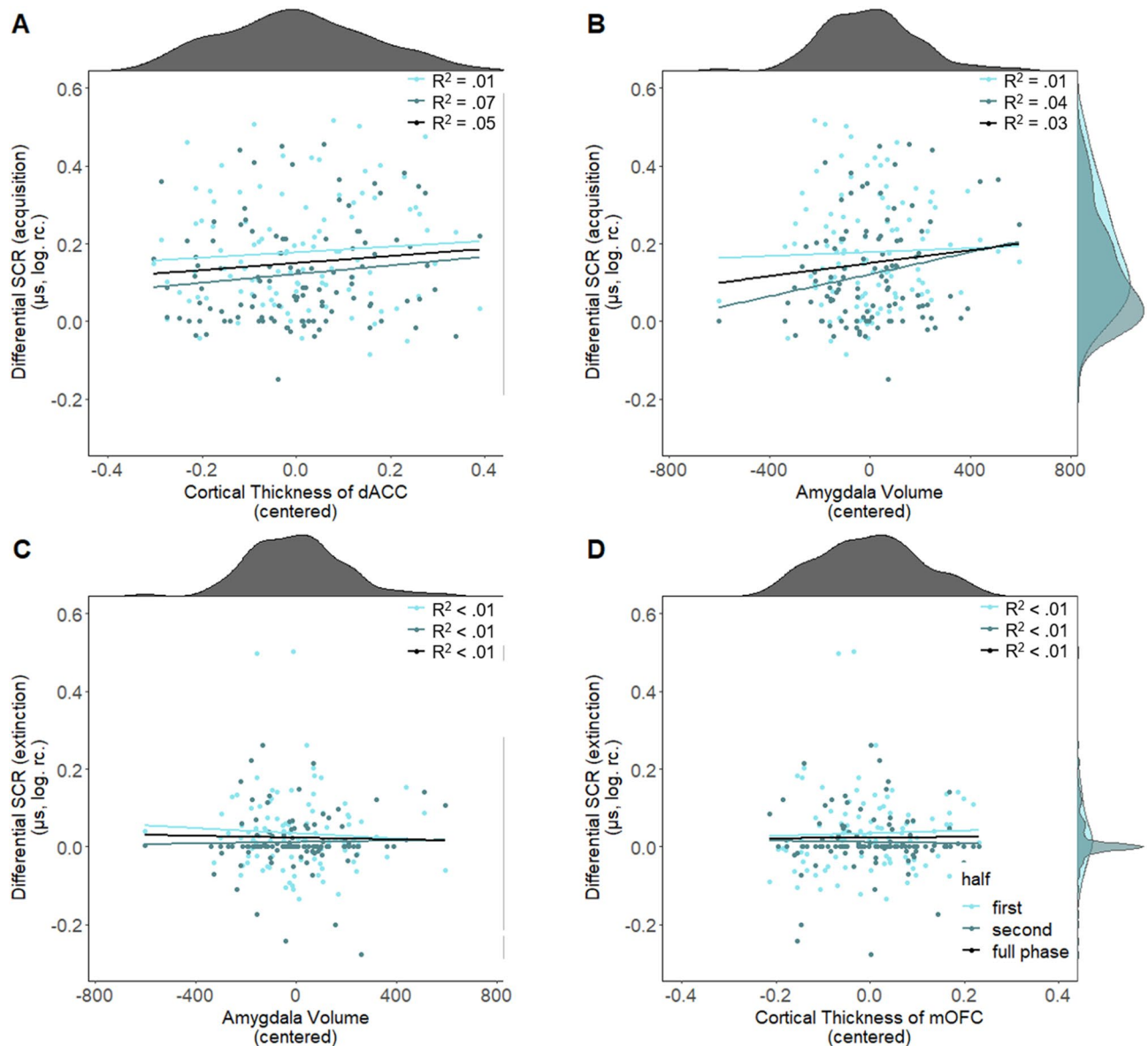


Figure 5. Scatterplots with marginal densities illustrating the (absence of) associations between average differential SCR [(CS+) – (CS–)] during acquisition training (illustrated also for the first and second half of acquisition separately) and (A) cortical thickness of the dACC and (B) amygdala volume as well as between differential SCR [(CS+) – (CS–)] during extinction (illustrated also for the first and second half of acquisition separately) and (C) amygdala volume as well as (D) cortical thickness of the mOFC. Data points are color-coded depending on the first half (light blue), second half (blue).

et al.¹⁴ presented a total of 36 trials per CS type during acquisition training (i.e., 18 CS+ and CS– during both the first and second half of acquisition training), while the current study design included a total of 14 CS+ and CS– trials during acquisition training. Consequently, the total number of trials during fear acquisition training in the present study was shorter than the first half of the previous study. We did, however, employ 100% compared to 50% reinforcement rate¹⁴, which likely led to faster fear acquisition in our study. Yet, we did not observe a significant association between amygdala volume and differential SCR or differential post-acquisition ratings when considering the full acquisition training phase—largely overlapping with the first half of acquisition training in¹⁴—or the first or second half of our acquisition training phase (6 and 7 trials per CS type respectively).

Another important difference that should be acknowledged when interpreting the current results is whether extinction took place immediately after fear acquisition training (i.e., immediate extinction) or after a time delay (i.e., delayed extinction) such as 24 h. Previous studies reporting a relationship between prefrontal thickness and fear learning proxies during extinction learning^{14,16}, extinction recall^{20,25} or renewal²⁴ have all employed an immediate extinction paradigm, while our own data as well as those by Hartley et al.²⁰ (sample 2) are based on a 24 h-delayed extinction procedure.

	dACC		Amygdala		mOFC	
	Regression	BF ₀₁	Regression	BF ₀₁	Regression	BF ₀₁
Fear acquisition training						
Differential SCR: full phase	$F(3,103) = 1.93, p = 0.13, R^2 = 0.05$	5.00	$F(3,103) = 0.93, p = 0.43, R^2 = 0.03$	18.18	–	–
Differential SCR: first half	$F(3,103) = 0.50, p = 0.68, R^2 = 0.01$	32.26	$F(3,103) = 0.22, p = 0.89, R^2 = 0.01$	47.62	–	–
Differential SCR: second half	$F(3,103) = 2.66, p = 0.052, R^2 = 0.07$	1.99	$F(3,103) = 1.61, p = 0.19, R^2 = 0.04$	7.52	–	–
Differential post acquisition fear ratings	$F(3,99) = 2.49, p = 0.06, R^2 = 0.07$	3.01	$F(3,99) = 1.51, p = 0.22, R^2 = 0.04$	11.24	–	–
Extinction training						
Differential SCR: full phase	–		$F(3,103) = 0.05, p = 0.99, R^2 < 0.01$	58.82	$F(3,103) = 0.04, p = 0.99, R^2 < 0.01$	58.82
Differential SCR: first half	–		$F(3,103) = 0.15, p = 0.93, R^2 < 0.001$	52.63	$F(3,103) = 0.08, p = 0.97, R^2 < 0.01$	58.82
Differential SCR: second half	–		$F(3,103) = 0.07, p = 0.98, R^2 < 0.001$	55.56	$F(3,103) = 0.03, p = 0.99, R^2 < 0.01$	62.50
Differential fear ratings [pre–post extinction]	–		$F(3,93) = 0.23, p = 0.88, R^2 = 0.01$	43.48	$F(3,93) = 0.23, p = 0.88, R^2 = 0.01$	45.45
Differential pre extinction fear ratings (fear recall)	–		$F(3,94) = 0.89, p = 0.45, R^2 = 0.03$	20.41	$F(3,94) = 0.89, p = 0.45, R^2 = 0.03$	21.28
Differential post extinction fear ratings	–		$F(3,100) = 0.62, p = 0.60, R^2 = 0.02$	30.30	$F(3,100) = 0.79, p = 0.50, R^2 = 0.02$	24.39

Table 2. Results of regression analyses with cortical thickness/subcortical volume and differential SCR and fear ratings during fear acquisition and extinction training (controlled for sex and TIV) and Bayes factor BF₀₁ providing relative evidence for the intercept-only against the hypothesis based regression model. Bold values indicate pre-registered hypotheses.

None of the studies considered here employed fear learning paradigms explicitly instructing the CS–US relationship^{14–16,20,21,24,25}, but reinforcement rate is another factor known to influence CS–US contingency awareness^{2,4,33}. Thus, we explored whether the previously reported relationship between fear learning proxies and brain morphology is masked by a modulation by contingency awareness. Our results, however, show that awareness affects differential SCR during fear acquisition training, but does not modulate a hypothesized brain behavior relationship. It should be noted though that the group of participants who were unaware was very small (N = 7) and hence this putative null finding needs to be interpreted with caution.

Besides differences in the experimental paradigms across studies, methods for measuring cortical thickness and subcortical volumes as well as SCR quantification differed between previous studies as well as previous work and our work. Assessment of cortical thickness and brain volume was nearly exclusively performed through automated methods as implemented in the software *Freesurfer*^{16,17,20,21,24,25} while only a single study employed manual segmentation of subcortical structures¹⁴. It can be speculated that employing different methods to assess brain morphology might have contributed to the non-identical results obtained in two previous studies which were based on largely overlapping samples^{14,16}. In contrast to the brain morphometry analyses, a plethora of different SCR quantification approaches was employed in previous work, most of which differed from our approach (see Table 1).

In sum, experimental paradigms and methodological procedures differ substantially between the studies in the field including ours. Yet, “It is tempting to explain away nonsignificant results in a line of studies by minor differences in the method, even when random variation is a much more likely explanation.” cf.³⁴—in particular in small, sub optimally powered studies which represent the major share of the work we based our hypotheses on.

While we were not able to (conceptually) replicate any of the previously reported structural brain behavior associations in a large sample of healthy adults, general functional brain activation patterns during fear acquisition and extinction training are relatively well established^{18,19}. Of note, functional activation patterns during fear acquisition, extinction or retention of extinction^{18,19} involve all brain regions that have been reported previously to show structural brain-behavior associations during fear conditioning studies^{14–17,20,21,24,25}. However, it is unclear whether and how inter-individual differences in structural characteristics relate to inter-individual differences in functional activation during different phases of a fear conditioning paradigm (as discussed by¹⁶). Furthermore, it has been suggested that brain structure and function may not be uniformly related but may show high coupling in sensory areas and particularly low coupling in the default mode or salience network³⁵. Critically, the so-called salience network comprises the dACC, orbital frontoinsula as well as limbic regions such as the amygdala³⁶—all regions functionally related to fear processing^{18,19} and serving as regions of interest in this study. In sum, while functional brain activation patterns underlying fear conditioning are well established, associations with brain morphology in the same regions seem questionable. One possible reason could be the low association of brain structure and function especially in neural circuitry underlying fear learning. Moreover, a recent systematic attempt to replicate a number of reported associations between cortical thickness or grey matter volume and psychometric variables and psychological measurements in a large sample of healthy adults showed no significant associations in more than 90% of the performed analyses^{5,13}. This led the authors to conclude that such associations are unlikely to be found and that—even with identical experimental designs—it is highly unlikely to replicate associations between brain morphology and psychometric variables. Importantly, replication rates decreased with decreasing sample size of the replication sample⁵ and associations have been shown to stabilize and become more reproducible in very large samples with N = ~ 2000³⁷. It is well recognized that also in initial studies, small sample sizes are generally linked to low statistical power and inflated effect sizes. Low statistical

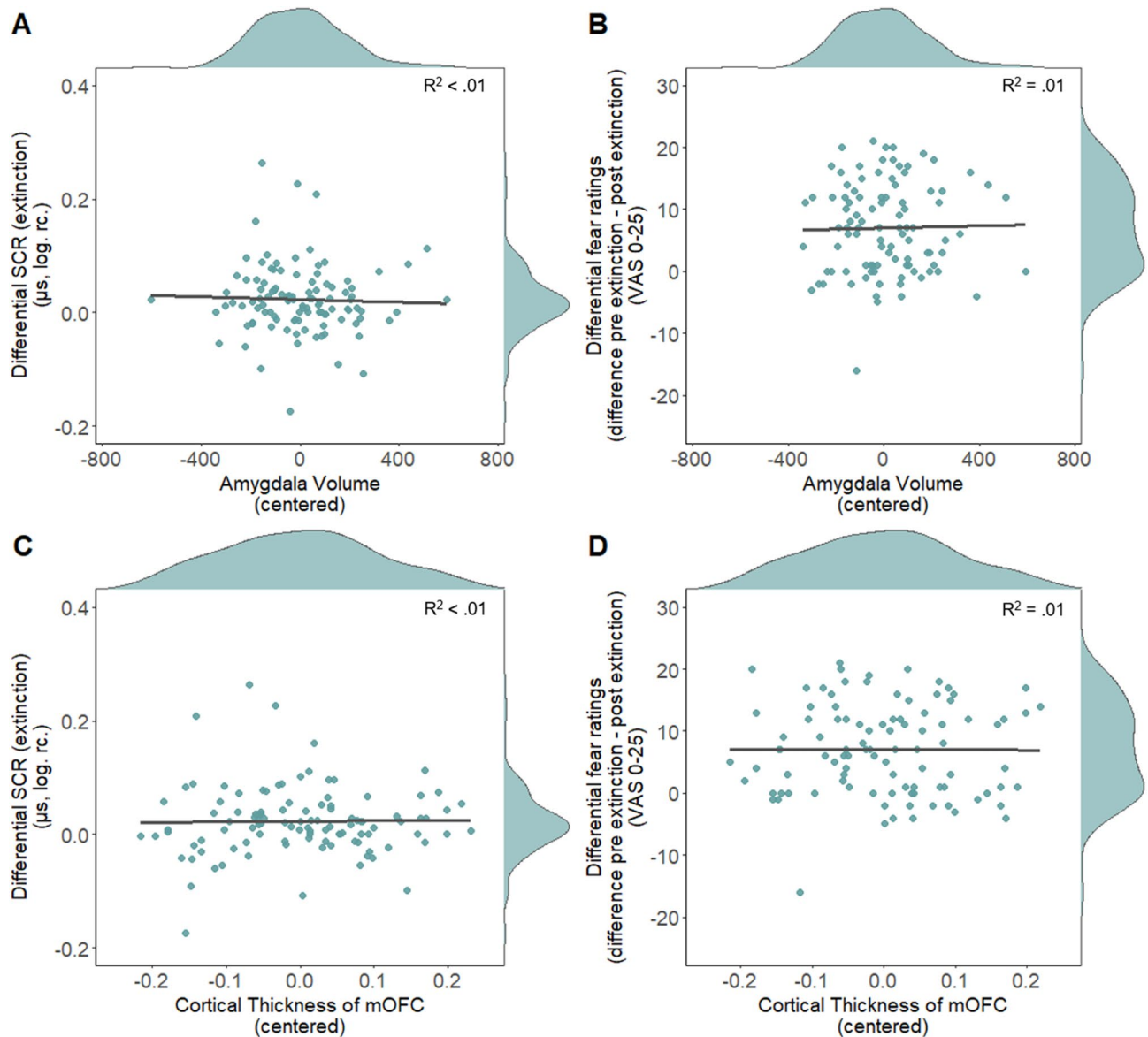


Figure 6. Scatterplots with marginal densities illustrating the (absence of) associations between average differential SCR [(CS+) – (CS–)] during extinction training and (A) amygdala volume and (C) cortical thickness of the mOFC and (B) (illustrated also for the first and second half of acquisition separately in Fig. 5) as well as between differential pre–post extinction fear ratings [(CS+_{pre}) – (CS–_{pre})] – [(CS+_{post}) – (CS–_{post})] and (B) the amygdala and (D) the mOFC thickness.

power does, however, not only reduce the likelihood to detect a true effect but also reduces the likelihood with which a significant finding actually reflects a true population effect. Consequently, small sample sizes are assumed to lead to low replication rates, as discussed for task-based fMRI^{27,38}. In light of this, it is maybe not surprising that we were unable to (conceptually) replicate previous findings which are often derived from extremely small sample sizes with 10–14 participants^{20,21,24,25}. Yet, we were also unable to (conceptually) replicate findings derived from (somewhat) larger samples^{16,17}.

While structural MRI measures themselves have been shown to have excellent test–retest reliability^{28,29}, the reliability of measures of defensive responding, such as SCR and fear ratings during fear acquisition and extinction training remains understudied and underreported. While within-subject reproducibility and test–retest reliability has been established with intermediate reliability coefficients for conditioned SCR across time intervals ranging from 3 weeks to 8 months (8 months³⁹, 3 weeks⁴⁰, 8–12 weeks⁴¹) for maximum CS+ responding, CS– responding and CS+/CS– discrimination in SCR^{39–41}. Reliability of other measures of defensive responding should also be systematically investigated in order to draw conclusions about potential reasons for the lack of associations presented here. This is important, as measurement reliability puts an upper bound to the maximum correlation that can be observed⁴² and it is likely that early reports of correlation coefficients as high as 0.8 (see Fig. 1A) might be inflated and implausibly high.

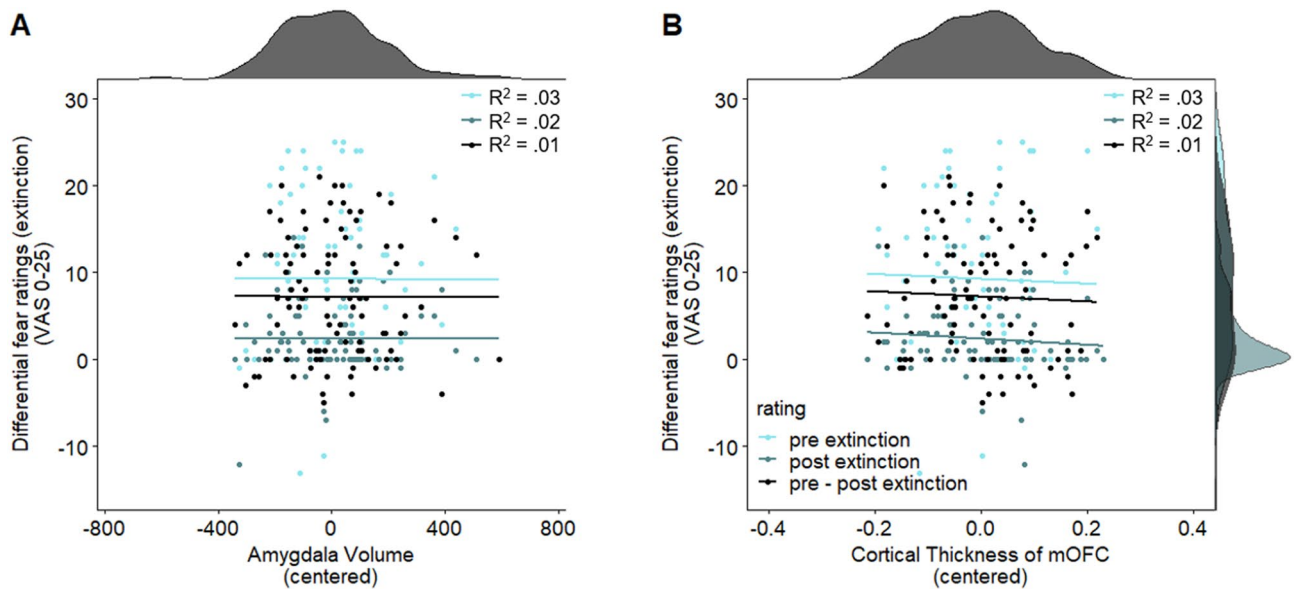


Figure 7. Scatterplots with marginal densities illustrating the (absence of) associations between differential pre, post and pre-post extinction fear ratings for (A) amygdala and (B) mOFC. Data points are color-coded to show fear ratings pre (light blue) and post (blue) extinction as well as the pre-post extinction difference score $[(CS_{pre}^+) - (CS_{pre}^-)] - [(CS_{post}^+) - (CS_{post}^-)]$ (black).

In conclusion, in line with recent studies questioning the existence and robustness of structural brain-behavior associations in healthy adults, we did not observe any associations between cortical thickness or subcortical volume in a number of brain regions and differential SCR and fear ratings as proxies for the acquisition and extinction of conditioned fear. Yet, if a finding cannot be replicated conceptually this may indicate that the association may only be observable under very specific boundary conditions. If true, this hampers the generalizability of the findings substantially. It is important to point out, however, that our work cannot be taken as evidence against the findings reported previously for several reasons: First, we do not provide a close or direct replication of any of these previous studies and second, the absence of a significant p-value in our study and the presence of a significant p-value in a given previous study cannot be taken to infer non-replication of an effect in absence of a formal statistical evaluation of replication (see⁴³ for a formal framework on replicability). Nevertheless, the current results cast some doubt on the idea that differences in brain morphology are likely to contribute to inter-individual differences in fear learning processes.

Future studies should employ longitudinal designs in order to investigate whether changes in brain morphology over time or measures of structural connectivity may have greater predictability for inter-individual differences in defensive responding. Most importantly, however, a focus on measures in general^{26,44} and the reliability of the measures used in studies on inter-individual differences in conditioned responding as well as structural and functional brain imaging are key and need to be scrutinized. In fact, the best research idea and the most transparent reporting methods cannot make up for inappropriate and/or unreliable measures employed. It may be time to take a step back and focus more on our measures because the reliable and reproducible quantification of measurements is fundamental to research in general and individual difference research in particular.

Methods and materials

Participants. The data set is part of the baseline measurement of a longitudinal fear conditioning study. For the current study, fear ratings, SCR and structural neuroimaging data from the first test-timepoint (T0) which consisted of two experimental days (Day 1: habituation, acquisition, Day 2: extinction) were included whereas reinstatement test (Day 2) and fMRI data were not analyzed here. All methods were carried out in accordance with relevant ethical guidelines and regulations. All experimental protocols were approved by the local ethics committee (PV 5157, Ethics Committee of the General Medical Council Hamburg). All participants gave written informed consent before participation. The current data set as well as the analysis code are made publicly available (<https://doi.org/10.17605/osf.io/y2jv9>). The data set has also been used as a case example in our previous publication on a methodological question different from the question addressed here⁴⁵. As pre-registered, several participants had to be excluded from the initial sample (N = 120) due to the following reasons: For Day 1, one participant had to be excluded due to missing data, three participants due to non-responding (no SCR response to the US on more than 9 out of 14 occasions) and an additional participant due to a deviating protocol on both days. Moreover, one participant was excluded due to technical issues on Day 2 in addition to five participants due to non-responding in SCR on Day 2 (see ‘Physiological Measurements—SCR’ for definition). Two participants were, as pre-registered, excluded from the analysis due to assumed technical issues on Day 2. Only after the data analysis did we realize that data for these participants was complete for fear acquisition and extinction training as technical issues only occurred in the subsequent reinstatement phase. Hence, these two participants could have been included but were excluded as pre-registered. After exclusions, structural and

psychophysiological data of $N = 107$ participants (71 females, mean \pm SD age of 24.4 ± 3.7 years, age range 18–34, state-trait anxiety inventory (STAI)⁴⁶ mean \pm SD of 34.6 ± 7.2 , range of 24–55) were included in the analyses. Due to missing data in fear ratings from fear acquisition training, $N = 103$ participants (67 females) were included into the analyses of fear ratings. Fear ratings pre extinction are missing from nine and ratings post extinction are missing from three participants, resulting in a total of $N = 95$ participants (64 females) for the comparison of pre and post extinction ratings.

Stimuli. An electrotactile stimulus administered to the back of the participant's right hand served as the US. The stimulus comprised three 2 ms electrotactile rectangular pulses with an interpulse interval of 50 ms delivered 200 ms before CS+ offset. The pulse was generated by a Digitimer DS7A constant current stimulator (Welwyn Garden City, Hertfordshire, UK) and delivered through a 1 cm diameter platinum pin surface electrode (Specialty Developments, Bexley, UK) placed between the metacarpal bones of the index and middle finger. US intensity was individually calibrated in a step-wise procedure to reach an unpleasant, but not painful level for each participant.

Two light grey fractals served as conditioned stimuli, the allocation of which to CS+ and CS– as well as the order was counterbalanced across participants. All stimuli were presented on grey background.

Experimental design. Participants completed a two-day paradigm consisting of habituation and acquisition training on Day 1 and extinction training, reinstatement administration and reinstatement test on Day 2. In the current study, only data from acquisition and extinction training are presented. For both acquisition and extinction training, CS+ and CS– were each presented 14 times in pseudo-randomized order for a duration of 6–8 s (mean duration: 7 s). Inter-trial intervals (ITIs) consisted of a white fixation cross presented for 10–16 s (mean duration: 13 s). Presentation of all stimuli on a grey background and stimulus timing were controlled by Presentation software (Version 14.8, Neurobehavioral Systems, Inc, Albany California, USA).

Fear ratings and contingency awareness. Fear ratings were completed after habituation and acquisition training on Day 1 as well as before extinction training and after reinstatement on Day 2. Participants were asked how much 'stress, fear and tension' they experienced when they last saw the CS+ and CS–. The ratings after reinstatement test referred to the first CS presentation per CS type after reinstatement administration and the last presentation during the test phase respectively (note that this phase was, however, not analyzed here). Answers were given within a 5 s time window on a visual analog scale (VAS) ranging from zero (answer = none) to 100 (answer = maximum), re-scaled to 0–25. A standardized post-experimental awareness interview adapted from⁴⁷ was conducted after acquisition training in order to assess CS–US contingency awareness. Subsequently, participants were classified as aware, unaware or uncertain of CS–US contingencies by the experimenter.

Physiological measurements—SCR. Physiological data were recorded with a Biopac MP100-amplifier system (BIOPAC Systems Inc, Goleta, California, USA) and AcqKnowledge 3.9.2 software and converted from analog to digital using a CED2502-SA with Spike 2 software (Cambridge Electronic Design, Cambridge, UK). Skin conductance response was measured by placing two self-adhesive, hydrogel Ag/AgCl electrodes on the distal and proximal hypothenar on the palmar side of the left hand. Data was continuously recorded at 1000 Hz with a gain of 5 or 10 $\mu\Omega$ and down-sampled to 10 Hz.

In line with previous recommendations^{48,49}, data were scored semi-manually as a trough to peak (TTP) response between 0.9 and 3.5 s after CS onset using the custom-made program EDA view (developed by Prof. Dr. Matthias Gamer, University of Würzburg). Rise time was set to a maximum of 5 s. Each scored SCR was checked visually, and the scoring suggested by EDA View was corrected if necessary. For example, the algorithm sometimes suggested an SCR outside the scoring window or the foot or trough were misclassified especially when several responses overlapped. Data with recording artifacts or excessive baseline activity (more than half of the response amplitude) were scored as missing values and excluded from the analysis. Response increases smaller than 0.01 μS in the pre-defined time window were set to zero, for a justification see⁴⁵. Raw SCR amplitudes were log transformed for purposes of normalization and range corrected by dividing each SCR by the maximum SCR (to CS or US) for each participant and day.

Physiological 'non-responding' on Day 1 was defined as no SCR response to the US on more than 9 out of 14 occasions. On Day 2, 'non-responding' was defined as no SCR response to any of three USs during reinstatement. A total of eight participants was classified as 'non-responders'⁴⁵.

MRI data acquisition and analysis. T1-weighted structural images ($1 \times 1 \times 1$ mm) were acquired on Day 2 with a 3T PRISMA whole body scanner (Siemens Medical Solutions, Erlangen, Germany) using a 64-channel head coil and magnetization prepared rapid gradient echo (MPRAGE) sequence (TR = 2300 ms, TE = 2.98 ms, field of view = 192×256 mm, 240 slices).

Cortical thickness and volume of subcortical brain regions were reconstructed using the brain imaging software Freesurfer 6.0.1^{7–9}. Thus, the regions of interest used in the current study are defined based on the areas implemented in Freesurfer and visualizations can be found online (<https://surfer.nmr.mgh.harvard.edu/>). The surface-based stream that yields measures of cortical thickness includes an initial Talairach registration, bias field correction, skull stripping, white matter classification, surface generation and gyral labeling⁷. Similarly, the volume-based or subcortical stream involves an initial Talairach registration, initial volumetric labeling, bias field correction, nonlinear volumetric atlas registration and volumetric labeling of subcortical structures⁵⁰. Cortical parcellation is based on the Desikan-Killiany cortical atlas⁵¹ implemented in Freesurfer.

Statistical analysis. The success of fear acquisition and extinction training was assessed by performing *t* tests and ANOVAs comparing averaged SCR elicited by the CS+ and CS− during acquisition and extinction training and fear ratings to the CS+ and CS− after acquisition as well as before and after extinction training. The SCR to the first CS+ and CS− of acquisition training were excluded from all analyses as no learning could possibly have taken place as the first CS+ presentation and the corresponding SCR occur prior to the first US presentation. Paired samples *t* tests were conducted to test for significant differences in cortical thickness and subcortical volume between the left and right hemisphere for the dACC, mOFC, insula and amygdala. For all other analyses, volumina of both hemispheres of a region were averaged and, as pre-registered, sex and total intracranial volume (TIV) were included as covariates.

To test the hypothesis that dACC thickness and amygdala volume predict conditioned responding during acquisition training, separate linear regressions predicting average differential [(CS+) − (CS−)] SCR during acquisition training and differential [(CS+) − (CS−)] post-acquisition fear ratings from dACC thickness and amygdala volume were conducted. Please note that the pre-registration used an ambiguous formulation regarding the ratings. We had used the term “mean differential fear rating” but there was only one rating after the acquisition phase. Additional analyses used average SCR responding during the first half (i.e. trials two to seven for acquisition training and trials one to seven for extinction) and second half (i.e., trials eight to fourteen for acquisition training and extinction) of acquisition and extinction training (pre-registered for amygdala, also performed for dACC for completeness).

For extinction, equivalent analyses were set up with average differential [(CS+) − (CS−)] SCR across all trials during extinction learning and fear ratings as outcome variables and amygdala volume and mOFC thickness as predictors. Regarding the fear ratings, our pre-registration used an ambiguous formulation (“mean differential fear ratings”). As we, in contrast to SCR, did only assess ratings prior to and after but not during extinction training, we specify here that we used the difference in ratings pre and post extinction [pre extinction − post extinction]. For completeness, exploratory analyses were also performed with pre and post extinction ratings instead of the difference score. As pre-registered, mOFC thickness was also tested as a predictor for average differential SCR during first and second half of extinction.

Pre-registered exploratory moderated regression analyses were conducted with dACC as predictor, averaged differential [(CS+) − (CS−)] SCR during fear acquisition training and differential [(CS+) − (CS−)] fear ratings after acquisition training as outcome and contingency awareness as moderator variables (reported in the Supplementary Material Section 1.1, Supplementary Figure 1).

Additionally, some non-preregistered analyses were performed for completeness, as additional robustness checks to the main analyses (because significant differences between volumina/thickness emerged between both hemispheres) and in order to replicate specific findings from individual studies^{16,20,21}. The results of these analyses can be found in the Supplementary Material.

1. The regression analyses testing for the main pre-registered hypotheses were also performed separately for left and right hemisphere. Full results are reported in the Supplementary Material (see Section 2.1 and Supplementary Figures 2 and 3 as well as Supplementary Table 1).
2. Robustness analyses were performed for all main pre-registered analyses including sex as covariate and no covariates in order to ensure that the current results can be generalized to different combinations of covariates⁵². Model fit comparisons were further performed in order to show whether including covariates added predictive power. Full results can be found in the Supplementary Material (see Section 2.2 and Supplementary Table 2).
3. As Milad et al.²¹ reported a correlation of cortical thickness of the dACC with SCR to CS+ and CS− only, we performed correlations with dACC thickness and CS+ and CS− elicited SCR. Additionally, we computed partial correlations controlling for sex and TIV. Results are reported in the Supplementary Material (see Section 3.1, Supplementary Figure 4 and Supplementary Table 5).
4. As Hartley et al.²⁰ reported an association between the right posterior insula and CS+/CS− discrimination during acquisition training, we conducted a correlational analysis for left, right and average insula thickness and differential SCR and fear ratings during acquisition training. Results are reported in the Supplementary Material (see Section 3.2, Supplementary Figure 5 and Supplementary Table 6).
5. As some^{53,54} but not all¹⁶ previous studies reported an association between trait anxiety and amygdala volume, partial correlations were calculated in order to test for a relationship between trait anxiety as well as state anxiety prior to acquisition and extinction training and amygdala volume in addition to amygdala volume and state anxiety one day after acquisition training. Results are reported in the Supplementary Material (see Section 3.3, Supplementary Figure 6 and Supplementary Table 7).

In addition to traditional null hypothesis significance testing (NHST), we computed Bayes factors for all analyses, allowing us to not only to find evidence for our tested hypotheses but to quantify the evidence in favor of the null hypotheses. In the current study, we used the R package “BayesFactor”⁵⁵ in order to calculate Bayes factors to obtain relative evidence for the tested regression (or correlation) model against a null or intercept-only model. Here, we report the Bayes Factor BF_{01} to directly show how much more likely the null hypothesis is relative to the alternative hypothesis given the data. Bayes factors (BF_{01}) > 1 are generally considered as evidence *against* the alternative hypothesis or *for* the null hypothesis⁵⁶. More specifically, weak evidence for the null hypothesis is defined as $BF_{01} = 1-3$, moderate evidence as $BF_{01} = 3-20$ and strong evidence as $BF_{01} = 20-150$ ⁵⁷.

All analyses and data visualizations were performed with the Software package R (Version 1.2.5033) using the following packages: ggpubr⁵⁸, ggplot2⁵⁹, cowplot⁶⁰, writexl⁶¹, car⁶², jtools⁶³, readr⁶⁴, broom⁶⁵, ggfortify⁶⁶, tidyr⁶⁷, scales⁶⁸, plyr⁶⁹, RColorBrewer⁷⁰, reshape2⁷¹, tidyverse⁷², grid⁷³, gridExtra⁷⁴, ggExtra⁷⁵, patchwork⁷⁶, apaTables⁷⁷,

MBESS⁷⁸, egg⁷⁹, ggm⁸⁰, effectsize⁸¹, ppcor⁸², GGally⁸³, psychReport⁸⁴, lsr⁸⁵, ez⁸⁶, lattice⁸⁷, dplyr⁸⁸, rmarkdown⁸⁹, Rmisc⁹⁰, gghalves⁹¹, BayesFactor⁵⁵. Power curves were plotted using open code <https://www.statmethods.net/stats/power.html>. Predictors for all linear regressions were centered in order to be able to investigate interactions and for easier interpretability. All effects are reported at significant level $p < 0.05$ unless indicated otherwise. Effect sizes are reported as Cohen's d . No follow-up analyses were conducted since the pre-registered analyses did not yield any significant results.

Received: 8 July 2020; Accepted: 2 November 2020

Published online: 16 November 2020

References

- Bush, D. E. A., Sotres-Bayon, F. & LeDoux, J. E. Individual differences in fear: isolating fear reactivity and fear recovery phenotypes. *J. Trauma Stress* **20**, 413–422 (2007).
- Lonsdorf, T. B. & Merz, C. J. 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 (2017).
- Myers, K. & Davis, M. Mechanisms of fear extinction. *Mol. Psychiatry* **12**, 120–150 (2007).
- Lonsdorf, T. B. *et al.* Don't fear 'fear conditioning': methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neurosci. Biobehav. Rev.* **77**, 247–285 (2017).
- Masouleh, S. K., Eickhoff, S. B. & Genon, S. Searching for replicable associations between cortical thickness and psychometric variables in healthy adults: empirical facts. *bioRxiv* <https://doi.org/10.1101/2020.01.10.901181> (2020).
- Gaser, C. & Dahnke, R. CAT—a computational anatomy toolbox for the analysis of structural MRI data. *OHBM Conference* (2016).
- Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis I. Segmentation and surface reconstruction. *NeuroImage* **9**, 179–194 (1999).
- Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis. II: inflation, flattening, and a surface-based coordinate system. *NeuroImage* **9**, 195–207 (1999).
- Fischl, B. & Dale, A. M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11050–11055 (2000).
- Walhovd, K. B., Fjell, A. M., Giedd, J., Dale, A. M. & Brown, T. T. Through thick and thin: a need to reconcile contradictory results on trajectories in human cortical development. *Cereb. Cortex N.Y.* **1991**(27), 1472–1481 (2017).
- Boekel, W. *et al.* A purely confirmatory replication study of structural brain-behavior correlations. *Cortex J. Devoted Study Nerv. Syst. Behav.* **66**, 115–133 (2015).
- Genon, S. *et al.* Searching for behavior relating to grey matter volume in a-priori defined right dorsal premotor regions: lessons learned. *NeuroImage* **157**, 144–156 (2017).
- Kharabian Masouleh, S., Eickhoff, S. B., Hoffstaedter, F., Genon, S. & Alzheimer's Disease Neuroimaging Initiative. Empirical examination of the replicability of associations between brain structure and psychological variables. *eLife* **8**, e43464 (2019).
- Cacciaglia, R., Pohlack, S. T., Flor, H. & Nees, F. Dissociable roles for hippocampal and amygdalar volume in human fear conditioning. *Brain Struct. Funct.* **220**, 2575–2586 (2015).
- Pohlack, S. T. *et al.* Hippocampal but not amygdalar volume affects contextual fear conditioning in humans. *Hum. Brain Mapp.* **33**, 478–488 (2012).
- Winkelmann, T. *et al.* Brain morphology correlates of interindividual differences in conditioned fear acquisition and extinction learning. *Brain Struct. Funct.* **221**, 1927–1937 (2015).
- Abend, R. *et al.* Anticipatory threat responding: associations with anxiety, development, and brain structure. *Biol. Psychiatry* **87**, 916–925 (2019).
- Fullana, M. A. *et al.* Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Mol. Psychiatry* **21**, 500–508 (2016).
- Fullana, M. A. *et al.* Fear extinction in the human brain: A meta-analysis of fMRI studies in healthy participants. *Neurosci. Biobehav. Rev.* **88**, 16–25 (2018).
- Hartley, C. A., Fischl, B. & Phelps, E. A. Brain structure correlates of individual differences in the acquisition and inhibition of conditioned fear. *Cereb. Cortex N.Y.* **1991**(21), 1954–1962 (2011).
- Milad, M. R. *et al.* A role for the human dorsal anterior cingulate cortex in fear expression. *Biol. Psychiatry* **62**, 1191–1194 (2007).
- Abend, R. Computational modeling of threat learning: faster conditioning linked to anxiety and thicker prefrontal cortex. *PsyArxiv Preprints* <https://doi.org/10.31234/osf.io/4paqt> (2020).
- Prokasy, W. F. & Ebel, H. C. Three components of the classically conditioned GSR in human subjects. *J. Exp. Psychol.* **73**, 247–256 (1967).
- Milad, M. R. *et al.* Thickness of ventromedial prefrontal cortex in humans is correlated with extinction memory. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 10706–10711 (2005).
- Rauch, S. L. *et al.* Orbitofrontal thickness, retention of fear extinction, and extraversion. *NeuroReport* **16**, 1909–1912 (2005).
- Lonsdorf, T. B., Merz, C. J. & Fullana, M. A. Fear extinction retention: is it what we think it is?. *Biol. Psychiatry* **85**, 1074–1082 (2019).
- Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
- Elliott, M. L. *et al.* What is the test–retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychol. Sci.* **31**, 792–806 (2020).
- Han, X. *et al.* Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *NeuroImage* **32**, 180–194 (2006).
- Iscan, Z. *et al.* Test–retest reliability of freesurfer measurements within and between sites: Effects of visual approval process. *Hum. Brain Mapp.* **36**, 3472–3485 (2015).
- Dunsmoor, J. E., Bandettini, P. A. & Knight, D. C. Impact of continuous versus intermittent CS-UCS pairing on human brain activation during Pavlovian fear conditioning. *Behav. Neurosci.* **121**, 635–642 (2007).
- Lissek, S., Pine, D. S. & Grillon, C. The strong situation: a potential impediment to studying the psychobiology and pharmacology of anxiety disorders. *Biol. Psychol.* **72**, 265–270 (2006).
- Mertens, G. *et al.* Fear expression and return of fear following threat instruction with or without direct contingency experience. *Cogn. Emot.* **30**, 968–984 (2016).
- Lakens, D. & Etz, A. J. Too true to be bad: when sets of studies with significant and nonsignificant findings are probably true. *Soc. Psychol. Pers. Sci.* **8**, 875–881 (2017).

35. Vázquez-Rodríguez, B. *et al.* Gradients of structure–function tethering across neocortex. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21219–21227 (2019).
36. Seeley, W. W. *et al.* Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci. Off. J. Soc. Neurosci.* **27**, 2349–2356 (2007).
37. Marek, S. *et al.* Towards reproducible brain-wide association studies. *bioRxiv* <https://doi.org/10.1101/2020.08.21.257758> (2020).
38. Turner, B. O., Paul, E. J., Miller, M. B. & Barbey, A. K. Small sample sizes reduce the replicability of task-based fMRI studies. *Commun. Biol.* **1**, 62 (2018).
39. Torrents-Rodas, D. *et al.* Testing the temporal stability of individual differences in the acquisition and generalization of fear. *Psychophysiology* **51**, 697–705 (2014).
40. Fredrikson, M., Annas, P., Georgiades, A., Hursti, T. & Tersman, Z. Internal consistency and temporal stability of classically conditioned skin conductance responses. *Biol. Psychol.* **35**, 153–163 (1993).
41. Zeidan, M. A. *et al.* Test–retest reliability during fear acquisition and fear extinction in humans. *CNS Neurosci. Ther.* **18**, 313–317 (2012).
42. Danner, D. Reliability—the precision of a measurement. *GESIS Survey Guidelines*. Mannheim, Germany: GESIS—Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_011 (2016).
43. LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M. & Vanpaemel, W. A unified framework to quantify the credibility of scientific findings. *Adv. Methods Pract. Psychol. Sci.* **1**, 389–402 (2018).
44. Flake, J. K. & Fried, E. I. *Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them*. <https://osf.io/hs7wm>, <https://doi.org/10.31234/osf.io/hs7wm> (2019).
45. Lonsdorf, T. B. *et al.* Navigating the garden of forking paths for data exclusions in fear conditioning research. *eLife* **8**, e52465 (2019).
46. Spielberger, C. D., Gorsuch, R. L. & Lushene, R. E. *Manual for the State-Trait Anxiety Inventory* (Consulting Psychologists Press, Palo Alto, 1983).
47. Bechara, A. *et al.* Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science* **269**, 1115–1118 (1995).
48. Boucsein, W. *et al.* Publication recommendations for electrodermal measurements. *Psychophysiology* **49**, 1017–1034 (2012).
49. Sjouwerman, R. & Lonsdorf, T. B. Latency of skin conductance responses across stimulus modalities. *Psychophysiology* **56**, e13307 (2019).
50. Fischl, B. *et al.* Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341–355 (2002).
51. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980 (2006).
52. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
53. Baur, V., Hänggi, J. & Jäncke, L. Volumetric associations between uncinate fasciculus, amygdala, and trait anxiety. *BMC Neurosci.* **13**, 4 (2012).
54. Blackmon, K. *et al.* Structural evidence for involvement of a left amygdala-orbitofrontal network in subclinical anxiety. *Psychiatry Res.* **194**, 296–303 (2011).
55. Morey, R. D. & Rouder, J. N. *BayesFactor: Computation of Bayes Factors for Common Designs*. R package version 0.9.12-4.2 (2018).
56. Wagenmakers, E.-J. *et al.* Bayesian inference for psychology. Part II: example applications with JASP. *Psychon. Bull. Rev.* **25**, 58–76 (2018).
57. Raftery, A. E. Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–163 (1995).
58. Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.2.3 (2019).
59. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer International Publishing, Cham, 2016). <https://doi.org/10.1007/978-3-319-24277-4>.
60. Wilke, C. O. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.0.0 (2019).
61. Ooms, J. *writexl: Export Data Frames to Excel 'xlsx' Format*. R package version 1.3 (2020).
62. Fox, J. & Weisberg, S. *An R Companion to Applied Regression* 3rd edn. (Sage, Thousand Oaks, 2019).
63. Long, J. A. *jtools: Analysis and Presentation of Social Scientific Data*. R package version 2.0.1 (2019).
64. Wickham, H., Hester, J. & Francois, R. *readr: Read Rectangular Text Data*. R package version 1.3.1 (2018).
65. Robinson, D. & Hayes, A. *broom: Convert Statistical Analysis Objects into Tidy Tibbles*. R package version 0.5.2 (2019).
66. Tang, Y., Horikoshi, M. & Li, W. *ggfortify: unified interface to visualize statistical results of popular R packages*. *R J.* **8**, 474–485 (2016).
67. Wickham, H. & Henry, L. *tidyr: Tidy Messy Data*. R package version 1.0.0 (2019).
68. Wickham, H. *scales: Scale Functions for Visualization*. R package version 1.0.0 (2018).
69. Wickham, H. The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29 (2011).
70. Neuwirth, E. *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2 (2014).
71. Wickham, H. Reshaping data with the reshape package. *J. Stat. Softw.* **21**, 1–20 (2007).
72. Wickham, H. *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1 (2017).
73. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2019).
74. Auguie, B. *gridExtra: Miscellaneous Functions for 'Grid' Graphics*. R package version 2.3 (2017).
75. Attali, D. & Baker, C. *ggExtra: Add Marginal Histograms to 'ggplot2', and More 'ggplot2' Enhancements*. R package version 0.9 (2019).
76. Pedersen, T. L. *patchwork: The Composer of Plots*. R package version 1.0.0 (2009).
77. Stanley, D. *apaTables: Create American Psychological Association (APA) Style Tables*. R package version 2.0.5 (2018).
78. Kelley, K. *MBESS: The MBESS R Package*. R package version 4.7.0 (2020).
79. Auguie, B. *egg: Extensions for 'ggplot2': Custom Geom, Custom Themes, Plot Alignment, Labeled Panels, Symmetric Scales, and Fixed Panel Size*. R package version 0.4.5 (2019).
80. Marchetti, G., Drton, M. & Sadeghi, K. *ggm: Functions for graphical Markov models*. R package version 2.3 (2015).
81. Ben-Shakhar, Makrowski & Lüdecke. *Compute and interpret indices of effect size*. CRAN (2020).
82. Kim, S. *ppcor: Partial and Semi-Partial (Part) Correlation*. R package version 1.1 (2015).
83. Schloerke, B. *et al.* *GGally: Extension to 'ggplot2'*. R package version 1.4.0 (2018).
84. Mackenzie, I. G. *psychReport: Reproducible Reports in Psychology*. R package version 0.7 (2019).
85. Navarro, D. J. *Learning Statistics with R: a tutorial for psychology students and other beginners* (2015).
86. Lawrence, M. A. *ez: Easy Analysis and Visualization of Factorial Experiments*. R package version 4.4-0 (2016).
87. Sarkar, D. *Lattice: Multivariate Data Visualization with R* (Springer, Berlin, 2008). <https://doi.org/10.1007/978-0-387-75969-2>.
88. Wickham, H., Francois, R., Lionel Henry & Müller, K. *dplyr: A Grammar of Data Manipulation*. R package version 0.8.3 (2019).
89. Allaire, J. *et al.* *rmarkdown: Dynamic Documents for R*. R package version 1.16 (2019).
90. Hope, R. M. *Rmisc: Rmisc: Ryan Miscellaneous*. R package version 1.5 (2013).
91. Tiedemann, F. *gghalves: Compose Half-Half Plots Using Your Favourite Geoms*. R package version 0.1.0 (2020).

Acknowledgements

This work was supported by grants awarded by the German Research Foundation to TBL (CRC 58 on “Fear, Anxiety and Anxiety Disorders”, Grant ID INST 211/633-2 and Grant ID LO 1980/4-1).

Author contributions

M.R.E. and J.N. generated hypotheses, pre-registered the study, analyzed, interpreted and visualized data and drafted the initial manuscript. M.K. acquired and prepared the data and helped with hypothesis generation. M.K.-J. was involved in data preparation and visualization. M.K. and M.K.-J. both critically revised the manuscript. T.B.L. acquired funding, generated hypotheses, helped with the pre-registration, data analysis and graphical visualization, interpreted data and drafted the initial manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-76683-1>.

Correspondence and requests for materials should be addressed to M.R.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

Supplementary Material

Revisiting potential associations between brain morphology, fear acquisition and extinction through new data and a literature review

Mana R. Ehlers¹, Janne Nold¹, Manuel Kuhn^{1,2}, Maren Klingelhöfer-Jens, Tina B. Lonsdorf¹

¹ Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Martinistrasse 52, 20246 Hamburg

² Department of Psychiatry, Harvard Medical School, and Center for Depression, Anxiety and Stress Research, McLean Hospital, Belmont, MA 02478 USA.

Table of Contents

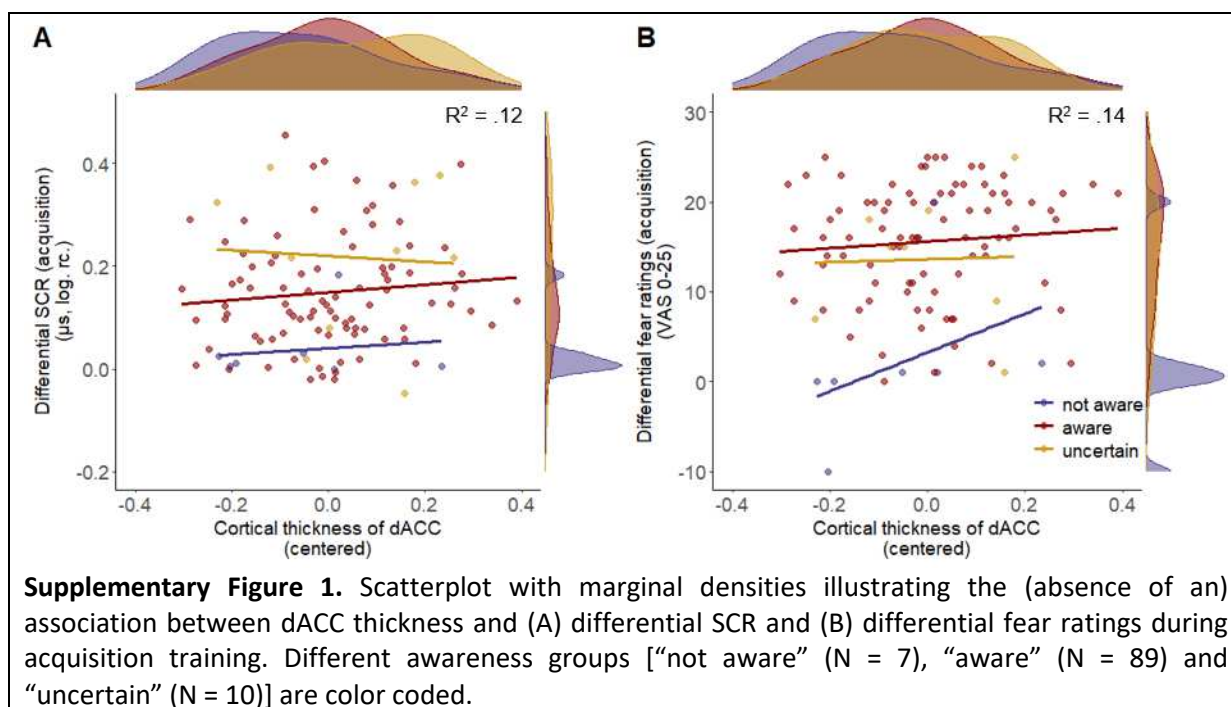
1. PRE-REGISTERED EXPLORATORY ANALYSES	2
1.1 Contingency awareness does not moderate the association between dACC thickness and defensive responding during fear acquisition training	2
2. NON-PRE-REGISTERED ROBUSTNESS ANALYSES	3
2.1 Robustness analyses considering data from right and left hemisphere separately	3
2.2 Robustness analyses including no covariates	7
2.3 Robustness analyses with raw SCR	9
2.4 Robustness analyses – outlier removed	10
3. ADDITIONAL, NON-PRE-REGISTERED ANALYSES AIMING TO (CONCEPTUALLY) REPLICATE PREVIOUSLY REPORTED FINDINGS	11
3.1 No association of dACC cortical thickness and SCR to the CS+ and CS- during acquisition training	11
3.2 No association between thickness of the insula and differential SCR and ratings during fear acquisition and extinction	12
3.3 No association of amygdala volume with trait and state anxiety	14

1. Pre-registered exploratory analyses

1.1 Contingency awareness does not moderate the association between dACC thickness and defensive responding during fear acquisition training

Contingency awareness has been identified as one factor contributing to inter-individual differences in defensive responding during fear acquisition training (Mertens & Engelhard, 2020; Tabbert et al., 2011). Here, we wanted to explore the pre-registered hypothesis that a potential association between dACC thickness and differential SCR and differential fear ratings during acquisition training might be moderated by contingency awareness.

A regression analysis with dACC thickness, contingency awareness as well as the pre-registered covariates sex and total intracranial volume (TIV) as predictors significantly predicted differential SCR ($F(4,101) = 3.52, p = .01, R^2 = .12$) and differential ratings ($F(4,97) = 3.65, p = .01, R^2 = .13$) during acquisition training. Adding the interaction term between contingency awareness and dACC thickness to the analysis still yielded a significant regression (SCR: $F(5,100) = 2.79, p = .02, R^2 = .12$, ratings: $F(5,96) = 3.24, p = .01, R^2 = .14$, see Supplementary Figure 1). The interaction between dACC thickness and contingency awareness was, however, not a significant predictor for differential SCR ($\beta = -.03, p = .86$) or ratings ($\beta = -15.86, p = .22$), rather the significant association was driven by awareness alone (SCR: $\beta = .08, p = .01$, ratings: $\beta = 4.45, p = .02$). These results should, however, be treated with caution since the group sizes differed substantially and were as low as 7 for the unaware group.



2. Non-pre-registered robustness analyses

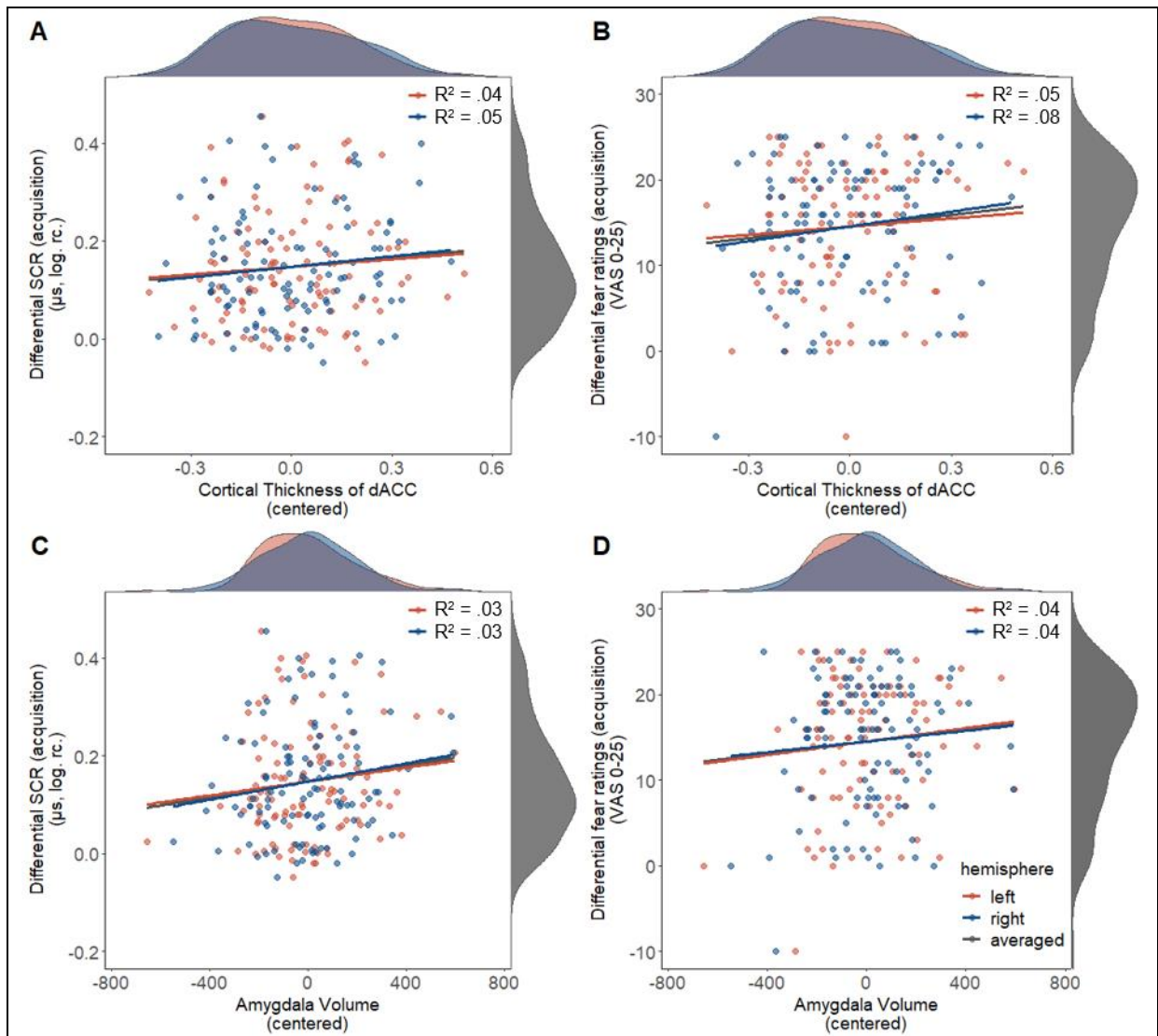
For full transparency, we report robustness analyses to demonstrate that the results presented in the main manuscript (i.e., pre-registered analyses) are not contingent on specific analysis choices, such as using averaged values over both hemispheres (see 2.1), the choice of specific – albeit pre-registered – covariates (see 2.2), transformation of raw scores (see 2.3) or not removing outliers (see 2.4).

2.1 Robustness analyses considering data from right and left hemisphere separately

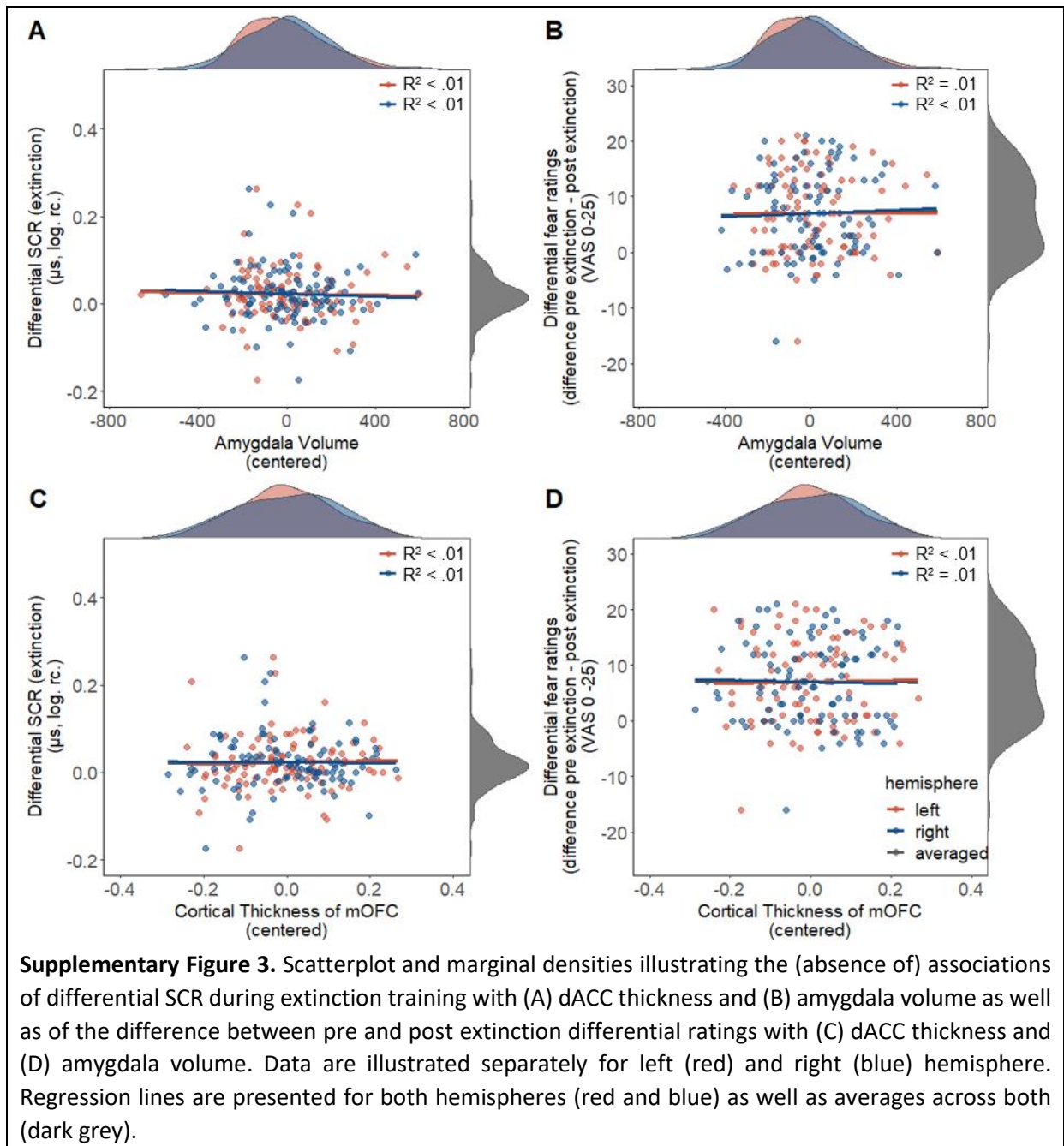
Previous research reported inconsistent lateralization (i.e., left or right lateralization) of the association of volume or cortical thickness and defensive responding during acquisition and extinction training. For instance, despite largely overlapping samples, Cacciaglia and colleagues observed a positive correlation between differential SCR and *left* amygdala volume, while Winkelmann and colleagues reported a positive correlation between differential SCR and *right* amygdala volume (Cacciaglia et al., 2014; Winkelmann et al., 2015). Further, effects were observed for right but not left insula and differential SCR during acquisition training (Hartley et al., 2011) and right but not left vmPFC and differential SCR during extinction training (Winkelmann et al., 2015). Hence, all major, preregistered analyses were also performed separately for left and right hemisphere for full transparency.

For that purpose, separate regression analyses with subcortical volume/cortical thickness derived from left and right hemisphere as predictor and differential SCR or differential fear ratings as outcome variables were performed for acquisition and extinction training for each brain region of interest (see methods, main manuscript). The pre-registered covariates sex and TIV were included as covariates for all analyses.

Similar to the results reported in the main manuscript, no significant association between any of the regions in any hemisphere was observed with either differential SCR or differential fear ratings during acquisition (see Supplementary Figure 2) or extinction training (see Supplementary Figure 3) (for full results see Supplementary Table 1) apart from a significant association between right dACC thickness and post-acquisition fear ratings ($F(3,99) = 2.73$, $p = .048$ $R^2 = .08$). However, the Bayes factor, $BF_{01} = 2.23$, for this association actually provides support for the null hypothesis (i.e. no significant relationship). Similarly, Bayes factors for all other analyses indicate that there is moderate to strong evidence ($BF_{01} > 3$) for the null or intercept-only model (see Supplementary Table 1).



Supplementary Figure 2. Scatterplot and marginal densities illustrating the (absence of) associations between differential SCR during acquisition training with (A) dACC thickness and (B) amygdala volume as well as differential fear ratings post acquisition training with (C) dACC thickness and (D) amygdala volume. Data are illustrated separately for left (red) and right (blue) hemisphere. Regression lines are presented for both hemispheres (red and blue) as well as averages across both (dark grey).



Supplementary Table 1. Results of regression analyses with left and right hemisphere cortical thickness/subcortical volume and differential SCR and fear ratings during fear acquisition and extinction training (controlled for sex and TIV) and Bayes factor BF_{01} providing relative evidence for the intercept-only model against the hypothesis based regression model.

	(A) Fear acquisition training				Amygdala			
	dACC				left		right	
	left	right	left	right	Regression	BF₀₁	Regression	BF₀₁
Differential SCR	$F(1,103) = 1.28,$ $p = .29, R^2 = .04$	11.90	$F(3,103) = 1.92,$ $p = .13, R^2 = .05$	5.15	$F(3,103) = .88,$ $p = .45, R^2 = .03$	19.23	$F(3,103) = .99,$ $p = .40, R^2 = .03$	16.95
Differential post acquisition fear ratings	$F(3,99) = 1.74,$ $p = .16, R^2 = .05$	7.94	$F(3,99) = 2.73,$ $p = .048, R^2 = .08$	2.23	$F(3,99) = 1.50,$ $p = .22, R^2 = .04$	11.11	$F(3,99) = 1.55,$ $p = .21, R^2 = .04$	10.42
	(B) Extinction training				Amygdala			
	mOFC				left		right	
	left	right	left	right	Regression	BF₀₁	Regression	BF₀₁
Differential SCR	$F(3,103) = .08,$ $p = .97, R^2 < .01$	58.82	$F(3,103) = .06,$ $p = .98, R^2 < .01$	58.82	$F(3,103) = .04,$ $p = .99, R^2 < .01$	62.50	$F(3,103) = .09,$ $p = .96, R^2 < .01$	52.63
Differential [pre – post extinction] fear ratings	$F(3,93) = .27,$ $p = .85, R^2 < .01$	43.48	$F(3,93) = .24,$ $p = .87, R^2 = .01$	45.45	$F(3,93) = .25,$ $p = .86, R^2 = .01$	43.48	$F(3,93) = .27,$ $p = .85, R^2 < .01$	41.67

2.2 Robustness analyses including no covariates

Our pre-registered analyses presented in the main manuscript included sex and TIV as covariates. It has been suggested to always include robustness analyses without covariates for full transparency (Simmons et al., 2011) and to ensure that presented results are not contingent on the covariates included.

Consequently, all main pre-registered analyses were also completed with either sex only or no covariates. In addition, for all analyses, the model fit of a regression with sex only or with sex and TIV as covariates was compared to a regression with morphometric estimates as the only predictor and no covariates. This serves the purpose to identify the best fitting model among those included and to determine whether the inclusion of specific covariates significantly alters model fit.

In brief, including only sex as covariate or no covariates yielded comparable results to those reported in the main manuscripts as no significant associations between brain morphology in any of the regions of interest and defensive responding in SCR and fear ratings during fear acquisition or extinction training were observed. Moreover, including covariates did not significantly improve model fit of the regression analyses.

More specifically, for acquisition training (for full results see Supplementary Table 2A), no significant associations between dACC thickness or amygdala volume and differential SCR or differential post acquisition fear ratings were observed with different combinations of covariates – with the exception of a significant association of amygdala volume and differential SCR during the second half of acquisition training when no covariates were included ($F(1,105) = 4.55, p = .04, R^2 = .04$) and Bayes factor of $BF_{01} = 0.65$ indicating moderate support for H_1 . However, it should be noted that applying a simple Bonferroni correction for multiple comparisons would render this result no longer significant (i.e., correcting for 9 tests concerning the amygdala and SCR would result in an alpha level of $\alpha = .006$). Importantly, the regression model of interest only becomes significant when no covariates are included but not with any other combination of covariates further questioning the robustness of this single positive result.

Overall, model fit was not significantly improved by including covariates with the exception of including sex and TIV as covariates in the analysis of the relationship between dACC thickness and fear ratings ($F(1,99) = 5.38, p = .02$).

For extinction training, neither amygdala volume nor mOFC thickness could be significantly predicted from differential SCR or fear ratings regardless of the covariates included. In line with this, model fit was not significantly improved by the addition of covariates (for full results see Supplementary Table 2B).

Supplementary Table 2. Results of robustness analyses for morphology and indices of fear learning including different covariates

(A) Fear acquisition training										
Structure	Outcome measure	Phase	Covariates				Model fit compared to analysis with no covariates			
			Sex		None		Sex		Sex and TIV	
			Regression	BF ₀₁	Regression	BF ₀₁				
dACC	SCR	Full	$F(2,104) = 1.90, p = .15, R^2 = .04$	4.37	$F(1,105) = 1.61, p = .21, R^2 = .02$	2.38	$F(1,104) = 2.19, p = .14$	$F(1,103) = 1.95, p = .17$		
		1 st half	$F(2,104) = .44, p = .65, R^2 = .01$	15.15	$F(1,105) = .62, p = .43, R^2 = .01$	3.70	$F(1,104) = .26, p = .61$	$F(1,103) = .63, p = .43$		
		2 nd half	$F(2,104) = 2.77, p = .07, R^2 = .03$	2.11	$F(1,105) = 1.81, p = .18, R^2 = .02$	2.18	$F(1,104) = 3.73, p = .06$	$F(1,103) = 2.35, p = .13$		
	ratings	Post	$F(2,100) = 1.01, p = .37, R^2 = .02$	8.77	$F(1,101) = 1.99, p = .16, R^2 = .02$	1.98	$F(1,100) = .04, p = .85$	$F(1,99) = 5.38, p = .02$		
Amygdala	SCR	Full	$F(2,104) = 1.19, p = .31, R^2 = .02$	6.85	$F(1,105) = 2.27, p = .13, R^2 = .02$	1.78	$F(1,104) = .13, p = .72$	$F(1,103) = .42, p = .52$		
		1 st half	$F(2,104) = .07, p = .93, R^2 < .01$	18.87	$F(1,105) = .12, p = .73, R^2 < .01$	4.63	$F(1,104) = .02, p = .89$	$F(1,103) = .51, p = .48$		
		2 nd half	$F(2,104) = 2.36, p = .10, R^2 = .04$	2.27	$F(1,105) = 4.55, p = .04, R^2 = .04$	0.65	$F(1,104) = .21, p = .65$	$F(1,103) = .16, p = .69$		
	ratings	Post	$F(2,100) = .80, p = .45, R^2 = .02$	1.03	$F(1,101) = .98, p = .33, R^2 = .01$	3.11	$F(1,100) = .64, p = .43$	$F(1,99) = 2.90, p = .09$		
(B) Extinction training										
Structure	Outcome measure	Phase	Covariates				Model fit compared to analysis with no covariates			
			Sex		None		Sex		Sex and TIV	
			Regression	BF ₀₁	Regression	BF ₀₁				
Amygdala	SCR	Full	$F(2,104) = .07, p = .93, R^2 < .01$	18.18	$F(1,105) = .02, p = .70, R^2 < .01$	4.59	$F(1,104) = .005, p = .95$	$F(1,103) = .003, p = .96$		
		1 st half	$F(2,104) = .23, p = .80, R^2 < .01$	16.39	$F(1,105) = .42, p = .52, R^2 < .01$	4.05	$F(1,104) = .03, p = .85$	$F(1,103) = .004, p = .95$		
		2 nd half	$F(2,104) = .09, p = .91, R^2 < .01$	19.23	$F(1,105) = .06, p = .81, R^2 < .01$	4.76	$F(1,104) = .13, p = .72$	$F(1,103) = .03, p = .86$		
	ratings	pre-post	$F(2,94) = .22, p = .88, R^2 < .01$	16.13	$F(1,95) = .04, p = .84, R^2 < .01$	4.59	$F(1,94) = .40, p = .53$	$F(1,93) = .24, p = .62$		
		Pre	$F(2,95) = .78, p = .46, R^2 = .02$	9.90	$F(1,96) < .01, p = .98, R^2 < .01$	4.69	$F(1,95) = 1.56, p = .22$	$F(1,94) = 1.10, p = .30$		
		Post	$F(2,101) = .40, p = .67, R^2 = .01$	14.08	$F(1,102) < .01, p = .98, R^2 < .01$	4.83	$F(1,101) = .81, p = .37$	$F(1,100) = 1.06, p = .31$		
mOFC	SCR	Full	$F(2,104) = .04, p = .96, R^2 < .01$	21.74	$F(1,105) = .92, p = .88, R^2 < .01$	4.83	$F(1,104) = .07, p = .81$	$F(1,103) = .04, p = .84$		
		1 st half	$F(2,104) = .07, p = .93, R^2 < .01$	20.83	$F(1,105) = .13, p = .72, R^2 < .01$	4.61	$F(1,104) = .02, p = .90$	$F(1,103) = .09, p = .77$		
		2 nd half	$F(2,104) = .05, p = .95, R^2 < .01$	21.28	$F(1,105) = .05, p = .82, R^2 < .01$	4.78	$F(1,104) = .05, p = .83$	$F(1,103) < .01, p = .95$		
	ratings	pre-post	$F(2,94) = .09, p = .91, R^2 < .01$	18.87	$F(1,95) < .01, p = .98, R^2 < .01$	4.67	$F(1,94) = .18, p = .67$	$F(1,93) = .50, p = .48$		
		Pre	$F(2,95) = .70, p = .50, R^2 = .01$	11.49	$F(1,96) = .11, p = .74, R^2 < .01$	4.46	$F(1,95) = 1.28, p = .26$	$F(1,94) = 1.27, p = .26$		
		Post	$F(2,101) = .78, p = .46, R^2 = .02$	11.11	$F(1,102) = .82, p = .37, R^2 = .01$	3.36	$F(1,101) = .74, p = .39$	$F(1,100) = .83, p = .37$		

2.3 Robustness analyses with raw SCR

All main pre-registered hypotheses regarding the association of SCR and brain morphology were also performed with raw SCR scores instead of log-transformed and range corrected SCR scores that were included in the analysis of the main manuscript.

In brief, the analyses reveal a very similar pattern of results to that presented in the main manuscript suggesting no relationship between differential SCR during fear acquisition and extinction training and brain morphology with both traditional NHST and a Bayesian approach.

Supplementary Table 3. Results of regression analyses with cortical thickness/subcortical volume and *raw* differential SCR during fear acquisition and extinction training (controlled for sex and TIV) and Bayes factor BF_{01} providing relative evidence for intercept-only model against the regression model.

	dACC		Amygdala		mOFC	
	Regression	BF_{01}	Regression	BF_{01}	Regression	BF_{01}
(A) Fear acquisition training						
Differential SCR: Full phase	$F(3,103) = 2.55,$ $p = .06 R^2 = .07$	2.31	$F(3,103) = 1.23,$ $p = .3, R^2 = .03$	12.82	---	---
Differential SCR: First half	$F(3,103) = .94,$ $p = .42 R^2 = .03$	18.18	$F(3,103) = .05,$ $p = .98, R^2 < .01$	62.50	---	---
Differential SCR: Second half	$F(3,103) = 2.67,$ $p = .051, R^2 = .07$	1.92	$F(3,103) = 1.68,$ $p = .18, R^2 = .05$	7.04	---	---
(B) Extinction training						
Differential SCR: Full phase	---		$F(3,103) = .16,$ $p = .92, R^2 < .01$	50.00	$F(3,103) = .13,$ $p = .94, R^2 < .01$	55.56
Differential SCR: First half	---		$F(3,103) = .17,$ $p = .92, R^2 < .01$	52.63	$F(3,103), .11,$ $p = .95, R^2 < .01$	55.56
Differential SCR: Second half	---		$F(3,103) = .11,$ $p = .95, R^2 < .01$	58.82	$F(3,103) = .18,$ $p = .91, R^2 = .01$	52.63

2.4 Robustness analyses – outliers removed

We checked the data for outliers (> 3 SD below or above mean, see for example Winkelmann et al., 2015) in fear ratings and SCR. One participant was excluded based on post-acquisition fear ratings, one based on pre-post extinction fear ratings and four based on differential SCR during extinction. The affected analyses were rerun after exclusions and the full results can be found in Supplementary Table 4. In summary, the general pattern of results remained the same with no significant associations.

Supplementary Table 4. Results of regression analyses with cortical thickness/subcortical volume and differential SCR and fear ratings during fear acquisition and extinction training (controlled for sex and TIV) with outliers (> 3 SD below or above mean) removed. Bayes factor BF_{01} provides relative evidence for intercept-only model against the regression model.

	dACC		Amygdala	
(A) Fear acquisition training	Regression	BF_{01}	Regression	BF_{01}
Differential post acquisition fear ratings	$F(3,98) = 1.45,$ $p = .23, R^2 = .04$	1.08	$F(3,98) = 0.99,$ $p = .40, R^2 = .03$	19.61
	Amygdala		OFC	
(B) Extinction training	Regression	BF_{01}	Regression	BF_{01}
Differential SCR: Full phase	$F(3,99) = .16,$ $p = .93, R^2 < .01$	52.63	$F(3,99) = .25,$ $p = .86, R^2 = .01$	45.45
Differential SCRs First half	$F(3,99) = .21,$ $p = .89, R^2 = .01$	47.62	$F(3,99) = .78,$ $p = .51, R^2 = .02$	23.26
Differential SCR: Second half	$F(3,99) = 1.45,$ $p = .20, R^2 = .05$	76.92	$F(3,99) = .14,$ $p = .94, R^2 < .01$	52.63
Differential fear ratings [pre-post extinction]	$F(3,92) = .28,$ $p = .84, R^2 = .01$	41.67	$F(3,92) = .33,$ $p = .81, R^2 = .01$	38.46

3. Additional, non-pre-registered analyses aiming to (conceptually) replicate previously reported findings

3.1 No association of dACC cortical thickness and SCR to the CS+ and CS- during acquisition training

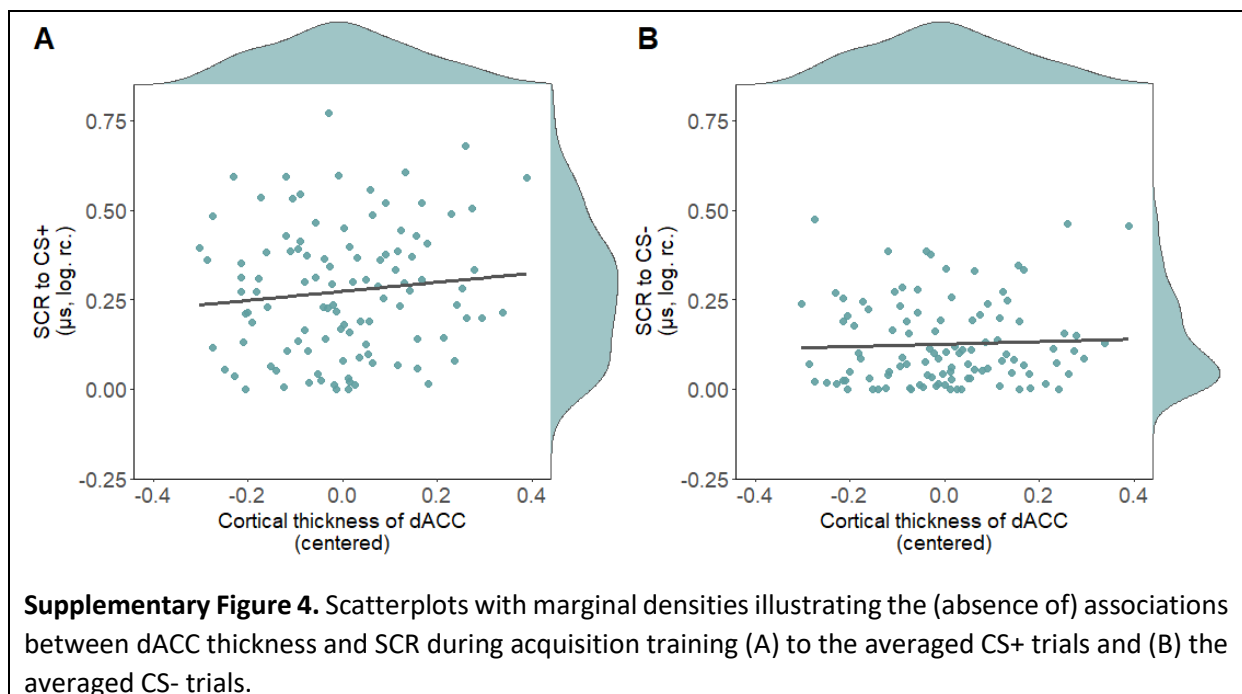
In a non-pre-registered analysis we aimed to replicate the previous finding of a significant correlation between dACC thickness and SCR to the CS+ but not the CS- during fear acquisition training (Milad et al., 2007). To be consistent with our previous analyses, we additionally computed partial correlations with sex as well as sex and TIV.

None of these analyses revealed a significant correlation between dACC thickness and SCR to either the CS+ or the CS- irrespective of covariates included and Bayes factors provide further evidence for the null hypothesis (for full results see Supplementary Table 5 and Supplementary Figure 4).

Supplementary Table 5. (Partial) Correlations between dACC thickness and CS+ and CS- and Bayes factor BF_{01} providing relative evidence for the full correlation against a null model.

	dACC		dACC ^a		dACC ^b		BF_{01}
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	
CS+	.11	.27	.16	.11	.16	.10	2.51
CS-	.04	.65	.08	.39	.08	.43	4.07

Note. ^a corrected for sex, ^b corrected for sex and TIV



3.2 No association between thickness of the insula and differential SCR and ratings during fear acquisition and extinction

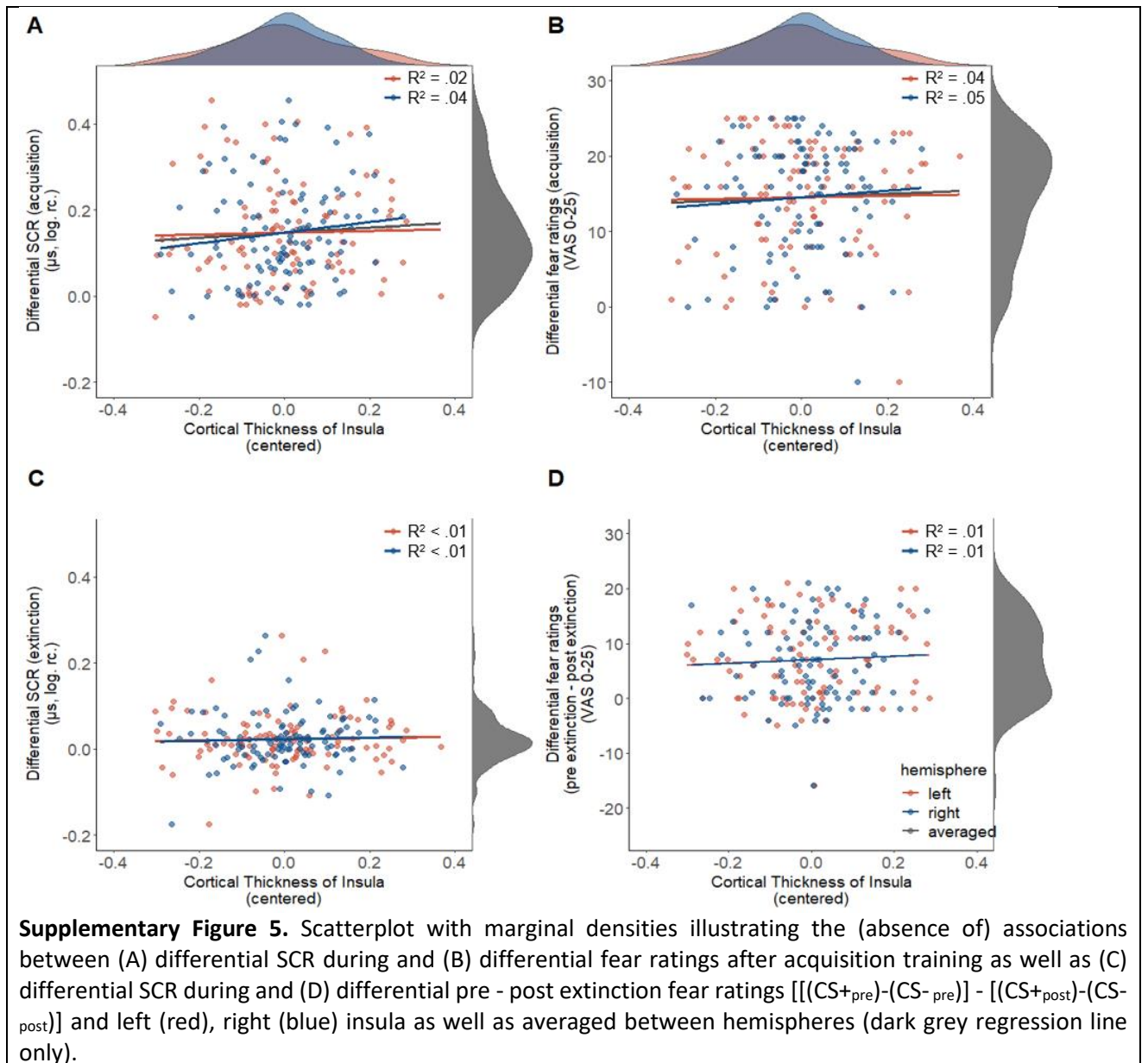
Hartley et al. (2011) reported a positive correlation between right (posterior) insula thickness and differential SCRs during acquisition training – even though in one out of two data sets, the correlation did not survive correction for multiple comparisons. In the current study, we aimed to replicate this finding in a substantially larger sample and for completeness extend them to differential fear ratings as well as extinction training. These analyses were not pre-registered.

We did not observe any significant correlations between differential SCRs or post acquisition ratings during acquisition training (see Supplementary Figure 5A and B) or differential SCRs and ratings (pre – post extinction) during extinction (Supplementary Figure 5C and D) for either right or left hemisphere or averaged insula thickness (for full results see Supplementary Table 6). Bayes factors further provide moderate evidence for the null hypothesis.

Supplementary Table 6.

Correlations between thickness of the insula (averaged over both hemispheres, left and right) with differential SCRs and differential ratings (post acquisition and pre – post extinction) during acquisition and extinction training and Bayes factor BF_{01} providing relative evidence for the null model against the tested correlation.

	Left Insula			Right Insula			Averaged Insula		
	<i>r</i>	<i>p</i>	BF_{01}	<i>r</i>	<i>p</i>	BF_{01}	<i>r</i>	<i>p</i>	BF_{01}
(A) Fear acquisition training									
Differential SCR	.03	.79	4.35	.12	.21	2.14	.08	.43	3.34
Differential post acquisition fear ratings	.02	.84	4.33	.07	.50	3.55	.05	.64	3.98
(B) Extinction training									
Differential SCR	.03	.74	4.27	.04	.72	4.22	.04	.70	4.18
Differential fear ratings [pre-post extinction]	.06	.26	3.65	.05	.63	3.82	.06	.54	3.60
Differential pre extinction fear ratings	.03	.74	4.08	.12	.25	2.29	.08	.43	3.23
Differential post extinction fear ratings	-.04	.72	4.17	.13	.18	1.89	.04	.67	4.07



3.3 No association of amygdala volume with trait and state anxiety

Previously, a *negative* correlation between left amygdala volume and state and trait anxiety (Blackmon et al., 2011), as well as a *positive* correlation between left amygdala volume and trait anxiety (Baur et al., 2012) has been reported while a third study (Winkelmann et al., 2015) did not observe any association between amygdala volume and trait anxiety.

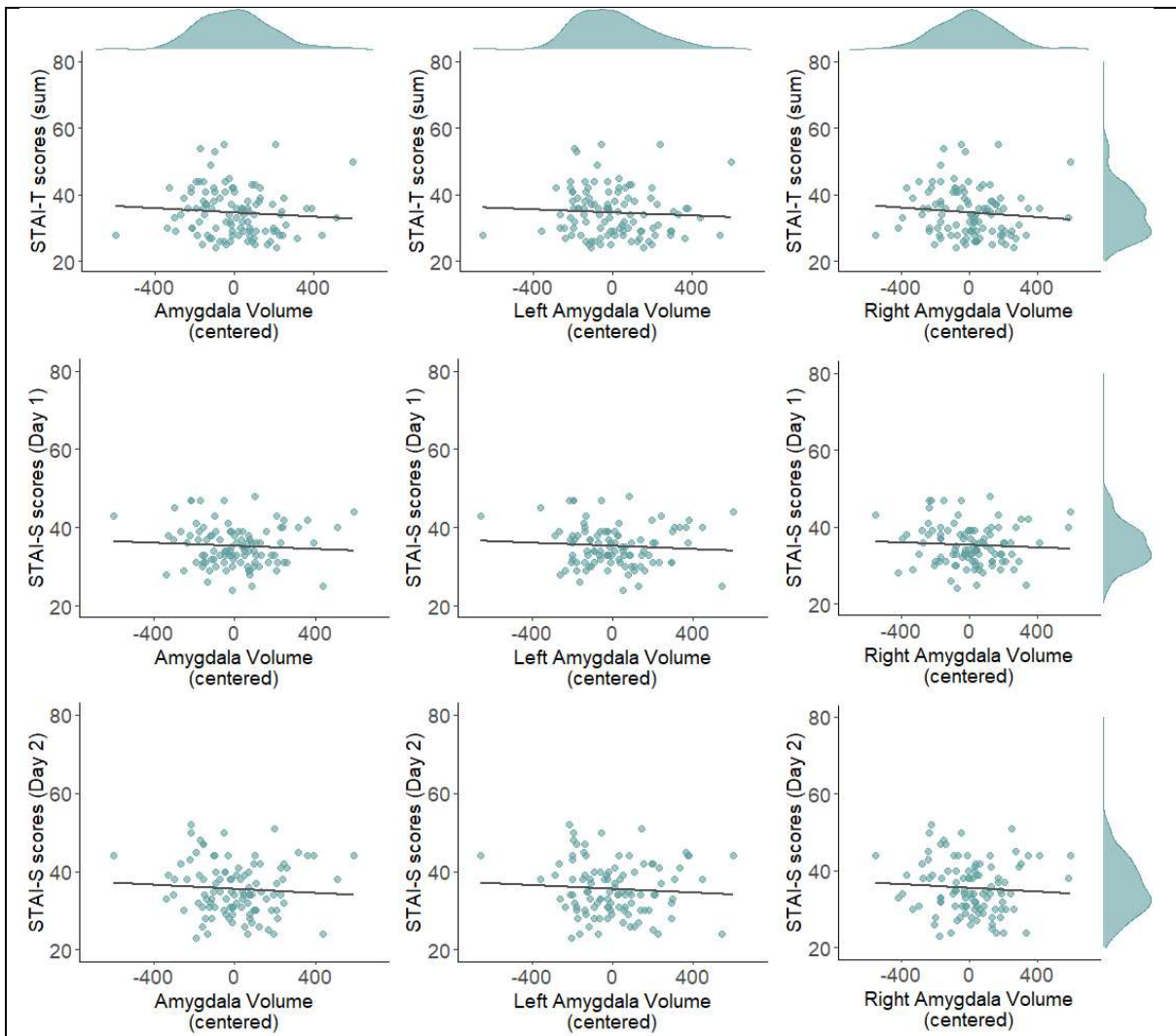
In the current study, we adopted the approach of Winkelmann et al. (2015) and calculated partial correlations between amygdala volume and trait anxiety as well as state anxiety (assessed prior to Day 1 acquisition training and Day 2 extinction training respectively) while controlling for the pre-registered covariates age, sex and TIV. We did not observe any significant associations between state or trait anxiety and averaged amygdala volume or right or left amygdala volume (for full results see Supplementary Table 7 and Supplementary Figure 6), which is further supported by Bayes factors suggesting support for the null hypothesis.

Supplementary Table 7.

Partial correlations of subcortical volume and STAI Trait/STAI State as indicator for anxiety and Bayes factor BF_{01} providing relative evidence for the null model against the full correlation.

	<i>M (SD)</i> [range]	Left Amygdala ^a			Right Amygdala ^a			Averaged Amygdala ^a		
		<i>r</i>	<i>p</i>	BF_{01}	<i>r</i>	<i>p</i>	BF_{01}	<i>r</i>	<i>p</i>	BF_{01}
STAI Trait	34.61 (7.19) [24, 55]	.09	.40	4.10	.02	.83	2.92	.06	.55	3.53
STAI State (Day 1)	35.25 (5.25) [24, 48]	.05	.64	3.36	.10	.35	3.65	.07	.45	3.44
STAI State (Day 2)	35.53 (6.77) [23, 52]	.04	.71	3.79	.06	.58	3.79	.05	.61	3.75

Note. ^a corrected for age, sex and TIV



Supplementary Figure 6. Scatterplots with marginal densities illustrating the lack of an association between trait anxiety (STAI-T) as well as state anxiety (STAI-S) prior to acquisition training (Day 1), and prior to extinction training (Day2) and amygdala volume (centered, for averaged, left and right volume).

8 Study II

This article was published in *eLife*, 11, Klingelhöfer-Jens, M., Ehlers, M. R., Kuhn, M., Keyaniyan, V., & Lonsdorf, T. B., Robust group- but limited individual-level (longitudinal) reliability and insights into cross-phases response prediction of conditioned fear, e78717, License: CC BY 4.0 DEED, <https://creativecommons.org/licenses/by/4.0/>, no changes have been implemented (2022).

Robust group- but limited individual-level (longitudinal) reliability and insights into cross-phases response prediction of conditioned fear

Maren Klingelhöfer-Jens^{1*}, Mana R Ehlers¹, Manuel Kuhn^{1,2}, Vincent Keyaniyan¹, Tina B Lonsdorf¹

¹Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; ²Department of Psychiatry, Harvard Medical School, and Center for Depression, Anxiety and Stress Research, McLean Hospital, Belmont, United States

Abstract Here, we follow the call to target measurement reliability as a key prerequisite for individual-level predictions in translational neuroscience by investigating (1) longitudinal reliability at the individual and (2) group level, (3) internal consistency and (4) response predictability across experimental phases. One hundred and twenty individuals performed a fear conditioning paradigm twice 6 months apart. Analyses of skin conductance responses, fear ratings and blood oxygen level dependent functional magnetic resonance imaging (BOLD fMRI) with different data transformations and included numbers of trials were conducted. While longitudinal reliability was rather limited at the individual level, it was comparatively higher for acquisition but not extinction at the group level. Internal consistency was satisfactory. Higher responding in preceding phases predicted higher responding in subsequent experimental phases at a weak to moderate level depending on data specifications. In sum, the results suggest that while individual-level predictions are meaningful for (very) short time frames, they also call for more attention to measurement properties in the field.

*For correspondence:
m.klingelhoef-jens@uke.de

Competing interest: The authors declare that no competing interests exist.

Funding: See page 28

Preprinted: 18 March 2022

Received: 17 March 2022

Accepted: 12 September 2022

Published: 13 September 2022

Reviewing Editor: Alexander Shackman, University of Maryland, United States

© Copyright Klingelhöfer-Jens et al. This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

Editor's evaluation

The authors assess the psychometric properties of behavioral, psychophysiological, and brain imaging measures of fear conditioning. Six-month retest reliability was generally low, whereas internal-consistency reliability was generally high. At the group level, reliability and criterion validity were generally good. Most measurements proved sensitive to data analytical choices. Results are framed within a larger discussion of the role of measurement properties in individual difference research and clinical translation and have the potential to serve as an important building block towards improvement in both these areas.

Introduction

The increasing incidence (e.g., *Xiong et al., 2022*) and high relapse rates (*Essau et al., 2018; Yonkers et al., 2003*) of anxiety-related disorders call for a better understanding of anxiety- and stress-related processes which might contribute to improving existing treatments or developing more effective interventions. In the laboratory, these processes can be studied using fear conditioning paradigms (*Dunsmoor et al., 2022; Fullana et al., 2020; Milad and Quirk, 2012*).

In differential fear conditioning protocols (see [Lonsdorf et al., 2017a](#)) one stimulus is repetitively paired with an aversive unconditioned stimulus (US; e.g., electrocutaneous stimulation), and as a consequence becomes a conditioned stimulus (CS+) while another stimulus, the CS-, is never paired with the US. After this acquisition training phase, CSs are presented without the US (extinction training) leading to a gradual waning of the conditioned response. Critically, the fear memory (CS+/US association) is not erased, but a competing inhibitory extinction memory (CS+/no US association) is assumed to be formed during extinction training ([Milad and Quirk, 2012](#); [Myers and Davis, 2007](#)). Subsequently, return of fear (RoF) can be induced by procedural manipulations such as a time delay (spontaneous recovery), a contextual change (renewal, [Vervliet et al., 2013a](#)), or a (re-)presentation of an aversive event (reinstatement, [Haaker et al., 2014](#)). Conditioned responding can be subsequently probed in an RoF test phase during which either the absence (i.e., extinction retention) or the return of conditioned responding (i.e., RoF) can be observed ([Bouton, 2004](#); [Lonsdorf et al., 2017a](#)).

Findings from studies employing fear conditioning paradigms hold strong potential for translating neuroscientific findings into clinical applications ([Anderson and Insel, 2006](#); [Cooper et al., 2022a](#); [Fullana et al., 2020](#); [Milad and Quirk, 2012](#)). More precisely, extinction learning is assumed to be the active component of exposure-based treatment ([Graham and Milad, 2011](#); [Milad and Quirk, 2012](#); [Rachman, 1989](#); [Vervliet et al., 2013b](#)) and experimental RoF manipulations have been suggested to serve as a model of clinical relapse ([Scharfenort et al., 2016](#); [Vervliet et al., 2013a](#)). Important findings in the fear conditioning field include the deficient learning of the safety signal (CS-) during acquisition training, impaired extinction learning ([Duits et al., 2015](#)) and the tendency of fear generalization to innocuous stimuli ([Cooper et al., 2022a](#)) in patients suffering from anxiety-related disorders as compared to healthy controls.

To date, both clinical and experimental research using the fear conditioning paradigm have primarily focused on group-level, basic, general mechanisms such as the effect of experimental manipulations – which is important to investigate ([Lonsdorf and Merz, 2017b](#)). Successful clinical translation (e.g., ‘Why do some individuals develop pathological anxiety while others do not?’) and particularly treatment outcome prediction (e.g., ‘Why do some patients benefit from treatment while others relapse?’), however, requires that both the experimental paradigm and the measures employed allow for individual-level predictions over and above prediction of group averages ([Fröhner et al., 2019](#); [Hedge et al., 2018](#); [Lonsdorf and Merz, 2017b](#)). A prerequisite for this is that the measures show stability within and reliable differences between individuals over time. Hence, tackling clinical questions regarding individual-level predictions of symptom development or treatment outcome requires a shift toward and a validation of research methods tailored to individual differences – such as a focus on measurement reliability ([Zuo et al., 2019](#)). This is a necessary prerequisite for the long-term goal of developing individualized intervention and prevention programs. This further relates to the pronounced heterogeneity in symptom manifestation among individuals diagnosed with the same disorders (e.g., post-traumatic stress disorder, PTSD, [Galatzer-Levy and Bryant, 2013b](#)) which cannot be captured in binary clinical diagnoses as two patients with for example a PTSD diagnosis may not share a single symptom ([Galatzer-Levy and Bryant, 2013b](#)).

Measurement reliability has only recently gained momentum in experimental cognitive research ([Fröhner et al., 2019](#); [Hedge et al., 2018](#); [Zuo et al., 2019](#)) and can be assessed through test–retest and longitudinal reliability (i.e., test–retest reliability over longer time intervals, typically assessed through e.g., intraclass correlation coefficients, ICCs, see [Table 1](#)). Importantly, longitudinal reliability (for definitions and terminology, see [Table 1](#)) also has implications for the precision with which associations of one variable (e.g., conditioned responding) with another (individual difference) variable can be measured because the correlation between those two variables cannot exceed the correlations within, that is the reliability, of these two variables ([Spearman, 1910](#)).

Yet, in fear conditioning research, surprisingly little is known about longitudinal reliability at the individual level with time intervals ranging from 9 days to 8 months in prior work ([Supplementary file 1](#); [Cooper et al., 2022b](#); [Fredrikson et al., 1993](#); [Ridderbusch et al., 2021](#); [Torrents-Rodas et al., 2014](#); [Zeidan et al., 2012](#)). Generally (details in [Supplementary file 1](#)), individual-level longitudinal reliability of risk ratings, skin conductance responses (SCRs), and fear potentiated startle (FPS) was within the same range ([Cooper et al., 2022b](#); [Torrents-Rodas et al., 2014](#)) whereas it was numerically somewhat lower for the BOLD response as compared to different rating types ([Ridderbusch et al., 2021](#)). Longitudinal reliability at the individual level appeared higher for acquisition training

Table 1. Definitions of key terms (A) and data specifications applied across analyses (B).

(A)	
Term	Definition
Internal consistency	In our study, internal consistency refers to the reliability of conditioned responding within experimental phases at both time points, respectively. It provides information on the extent to which items – or in our case – trials measure the same construct (e.g., fear acquisition). Odd and even trials were splitted (i.e., split-half method), averaged per subject and correlated across the sample.
Longitudinal reliability at the individual level	Longitudinal reliability at the individual level indicates to which extent responses within the same individuals are stable over time . It takes into account the individual responses of participants, which are then related across time points. Longitudinal reliability at the individual level inherently includes the group level, as it is calculated for the sample as a whole, but the individual responses are central to the calculation.
<ul style="list-style-type: none"> Intraclass correlation coefficients (ICCs) 	‘ICC coefficients quantify the extent to which multiple measurements for each individual (within individuals) are statistically similar enough to discriminate between individuals’ (Aldridge et al., 2017). Here, we calculated two types of ICCs, namely absolute agreement and consistency . To illustrate the difference between absolute agreement and consistency in a short example (Koo and Li, 2016), consider an interrater reliability study with two raters: Consistency indicates the extent to which the score of one rater (y) is equal to the score of another rater (x) plus a systematic error (c) (i.e., $y = x + c$). In contrast, absolute agreement indicates to which degree y equals x. As ‘two raters’ can be replaced by ‘two time points’ and individual responses were taken into account here, we used ICCs to determine longitudinal reliability at the individual level.
<ul style="list-style-type: none"> Within- and between-subject similarity 	<p>Similarity analyses provide information on the extent to which trial-by-trial responses of one individual at one time point are comparable to responses of</p> <ul style="list-style-type: none"> the same individual at a later time point (i.e., within-subject similarity) and all other individuals at a later time point (i.e., between-subject similarity). <p>Comparisons of within- and between-subject similarity were used here to determine longitudinal reliability at the individual level.</p>
<ul style="list-style-type: none"> Overlap at the individual level (applied for BOLD fMRI only) 	Overlap at the individual level reflects the degree of overlap of significant voxels between both time points for single subject-level activation patterns .
Longitudinal reliability at the group level	Longitudinal reliability at the group level indicates to which degree responses within the group as a whole are stable over time . More precisely, longitudinal reliability at the group level relies on first averaging all individuals responses for each trial (for SCR) or voxel (for fMRI) yielding a group average for each trial/voxel. These are then related across time points, that is the calculation is carried out using the trial-wise (for SCR) or voxel-wise (for fMRI) group averages.
<ul style="list-style-type: none"> Overlap at the group level (applied for BOLD fMRI only) 	Overlap at the group level reflects the degree of overlap of significant voxels between both time points for aggregated group-level activations .

Table 1 continued on next page

Table 1 continued

	Measure	Internal consistency	Longitudinal reliability at the individual level			Longitudinal reliability at the group level	Cross-phases predictability
			ICCs	Within- and between-subject similarity	Overlap	Overlap (BOLD fMRI) or R squared (SCR)	
Included time points	All	T0 and T1 separately	T0 and T1	T0 and T1	T0 and T1	T0 and T1	T0
	SCR	CS+, CS-, CS discrimination, US	CS+, CS-, CS discrimination, US*	CS+, CS-, CS discrimination, US	–	CS+, CS-, CS discrimination, US	CS+, CS-, CS discrimination
Included stimuli	Fear ratings	–	CS+, CS-, CS discrimination, US*	–	–	–	CS+, CS-, CS discrimination
	BOLD fMRI	–	CS discrimination†	CS discrimination†	CS discrimination†	CS discrimination†	CS+, CS-, CS discrimination
	SCR	Entire phases (ACQ, EXT, RI-Test; except first trials of ACQ and EXT)	CS+, CS-, and CS discrimination: average ACQ, last two trials ACQ‡, first trial EXT§, average EXT, last two trials EXT¶, first trial RI-Test§ US: average RI	Average ACQ**, average EXT	–	Average ACQ, average EXT	Average ACQ, last two trials ACQ‡, first trial EXT§, average EXT, last two trials EXT¶, first trial RI-Test§
Phase operationalizations	Fear ratings	–	CS+, CS-, and CS discrimination: post-pre ACQ, post ACQ, pre EXT, pre-post EXT, post EXT, first trial RI-Test US: post RI-Test	–	–	–	post-pre ACQ, post ACQ, pre EXT, pre-post EXT, post EXT, first trial RI-Test
	BOLD fMRI††	–	Average ACQ, average EXT	Average ACQ, average EXT	Average ACQ, average EXT	Average ACQ, average EXT	Average ACQ, average EXT
Transformations ††	SCR	None, log-transformation§§, log-transformation and range correction¶¶	None, log-transformation§§, log-transformation and range correction¶¶	None***	–	None, log-transformation§§, log-transformation and range correction¶¶	None, log-transformation§§, log-transformation and range correction¶¶
	Fear ratings	–	None	–	–	–	None
	BOLD fMRI	–	None	None	None	None	None

Table 1 continued on next page

Table 1 continued

(B)

	Measure	Internal consistency	Longitudinal reliability at the individual level			Longitudinal reliability at the group level	Cross-phases predictability
			ICCs	Within- and between-subject similarity	Overlap	Overlap (BOLD fMRI) or R squared (SCR)	
Ordinal ranking ^{†††}	SCR	No ranking	No ranking ^{†††}	No ranking	–	No ranking	No ranking and ordinal ranking ^{§§§}
	Fear ratings	–	No ranking ^{†††}	–	–	–	No ranking and ordinal ranking
	BOLD fMRI	–	No ranking	No ranking	No ranking	No ranking	No ranking

The specifications we used here are exemplary and are not intended to cover all possible data specifications. Note that internal consistency, within- and between-subject similarity and reliability at the group level could not be calculated for fear ratings due to the limited number of trials. ACQ = acquisition training, EXT = extinction training, RI = reinstatement, RI-Test = reinstatement-test.

*Non-pre-registered ICCs for SCRs to the USs and US aversiveness ratings were calculated as we considered these informative.

[†]For BOLD fMRI, ICCs were calculated only for CS discrimination and not for CS+ and CS– given the fact that the calculations are based on first-level T contrast maps and contrasts against baseline are not optimal.

^{††}In addition to the averaged acquisition and extinction training performance, the last two SCR trials of acquisition (pre-registered) and extinction training (not pre-registered) were separated from the previous trials and averaged as equivalent to the post-acquisition/-extinction ratings. The first extinction trial was taken into account separately as fear recall.

[§]Fear recall and reinstatement-test were operationalized as the first extinction training trial and the first reinstatement-test trial (since the reinstatement effect fades away relatively quickly, [Haaker et al., 2014](#)), respectively.

^{†††}The operationalization of extinction training as the last two trials was not pre-registered and included for completeness. In cases where phase operationalizations included more than one SCR trial, trials were averaged.

**Note that reliability at a group level for SCRs during reinstatement-test was not calculated as correlations between two SCR data points are not meaningful.

^{††††}fMRI data for the reinstatement-test were not analyzed in the current study since data from a single trial do not provide sufficient power.

^{†††††}The pre-registered transformation types were identified to be typically employed data transformations in the literature by for example [Sjouwerman et al., 2022](#) who also pre-registered these transformation types.

^{§§}Raw SCR amplitudes were log-transformed by taking the natural logarithm to normalize the distribution ([Levine and Dunlap, 1982](#)).

^{††††††}Log-transformed SCR amplitudes were range corrected by dividing each individual SCR trial by the maximum SCR trial across all CS and US trials. Due to potentially different response ranges, the maximum SCR trial was determined separately for experimental days as recommended by [Lonsdorf et al., 2017a](#). Range correction was recommended to control for interindividual variability ([Lykken, 1972; Lykken and Venables, 1971](#)).

^{†††††††}We also carried out similarity analyses for log-transformed as well as for log-transformed and range corrected data. However, results were almost identical to the results from the raw data. For reasons of space, we only report results for raw data.

^{††††††††}Ranking of the data was included to investigate to which degree individuals occupy the same ranks at both time points as pre-registered or put differently, whether the quality of predictions changes when the predictions were not based on the absolute values but on a coarser scale.

^{†††††††††}As opposed to what was pre-registered, in ICC analyses, we included non-ranked data only as closer inspection of the conceptualization of ICC_{con} revealed that it would be redundant to calculate both ICC_{abs} and ICC_{con} with ranked and non-ranked data as ICC_{con} itself ranks the data.

^{§§§§}Ranks of SCRs were built upon raw, log-transformed as well as log-transformed and range corrected values.

than for extinction training (SCRs: *Fredrikson et al., 1993; Zeidan et al., 2012*), but comparable to generalization (*Cooper et al., 2022b; Torrents-Rodas et al., 2014*). Moreover, it appeared higher for extinction training than for reinstatement-test (for BOLD fMRI but not ratings: *Ridderbusch et al., 2021*) and higher for CS+ than CS− responses (SCRs: *Fredrikson et al., 1993*) and CS discrimination (ratings and BOLD fMRI: *Ridderbusch et al., 2021*; SCR: *Zeidan et al., 2012*).

However, it is difficult to extract a comprehensive picture from these five studies as they differ substantially in sample size ($N = 18$ – 100), paradigm specifications, experimental phases reported, outcome measures, time intervals, and employed reliability measures (see *Supplementary file 1*).

Given that the predominance of research on group-level generic mechanisms in fear conditioning research, it is even more surprising that, to our knowledge, no study to date has investigated longitudinal reliability at the group level and only few studies have (*Fredrikson et al., 1993*) targeted internal consistency (i.e., the degree to which all test items capture the same construct, see *Table 1*). More precisely, longitudinal reliability at the group level indicates the extent to which responses averaged across the group as a whole are stable over time, which is important to establish when investigating basic, generic principles such as the impact of experimental manipulations. Even though it has to be acknowledged that the group average is not necessarily representative of any individual in the group and the same group average may arise from different and even opposite individual responses at both time points in the same group, group-level reliability is important to establish in addition to individual-level reliability. Group-level reliability is relevant not only to work focusing on the understanding of general, generic processes but also for questions about differences between two groups of individuals such as patients vs. controls (e.g., see meta-analyses of *Cooper et al., 2022a; Duits et al., 2015*). Of note, many fear conditioning paradigms were initially developed to study general group-level processes and to elicit robust group effects (*Lonsdorf and Merz, 2017b*). Hence, it is important to investigate both group- and individual-level reliability given the challenges of attempts to employ cognitive tasks that were originally designed to produce robust group effects in individual difference research (*Elliott et al., 2020; Hedge et al., 2018; Parsons, 2020; Parsons et al., 2019*).

As pointed out above, individual-level reliability is a prerequisite for individual-level predictions such as treatment outcomes. Since the different experimental phases of fear conditioning paradigms serve as experimental models for the development, treatment, and relapse of anxiety- and stress-related disorders, it is also an important question whether responding across phases can be reliably predicted at the individual level. Interestingly, it is often implicitly assumed that responding in one experimental phase reliably predicts responding in a subsequent phase (e.g., see *Milad et al., 2009*; critically discussed in *Lonsdorf et al., 2019a*) even though empirical evidence is lacking. As a result it has been suggested to routinely ‘correct for responding’ during fear acquisition training when studying performance in later experimental phases such as extinction training or retention/RoF test (critically discussed in *Lonsdorf et al., 2019a*). However, empirical evidence on this cross-phases predictability (for definition and terminology, see *Table 1*) is scarce to date.

Evidence from experimental work on cross-phase predictability in rodents and humans is mixed. In rodents, freezing during acquisition training and 24-hrs-delayed extinction training were uncorrelated (*Plendl and Wotjak, 2010*) and responding during extinction training did not predict extinction retention (i.e., lever-pressing suppression: *Bouton et al., 2006*; or freezing behavior: *Shumake et al., 2014*). Similarly, in humans, extinction performance (FPS, SCR, and US expectancy ratings) did not predict performance at 24-hrs-retention test (*Prenoveau et al., 2013*). Yet, a computational modeling approach suggests that the mechanism of extinction learning (i.e., the formation of a new extinction memory trace in comparison to an update of the original fear memory trace) predicts the extent of spontaneous recovery in SCR (*Gershman and Hartley, 2015*).

Also evidence from work in patient samples is mixed (for a review, see *Craske et al., 2008*). The extent of fear reduction within therapeutic sessions was unrelated to overall treatment outcome in some studies (*Kozak et al., 1988; Pitman et al., 1996; Riley et al., 1995*), while others observed an association (*Foa et al., 1983*). Similarly, significant correlations of fear reduction between therapeutic sessions with treatment outcome were observed for reported distress (*Rauch et al., 2004*) and heart rate, but not for SCR (*Kozak et al., 1988; Lang et al., 1970*) and for self-reported fear post treatment, but not at follow-up (*Foa et al., 1983*). In addition, evidence that responding in different phases is related comes from pharmacological manipulations with the cognitive enhancer D-cycloserine which facilitates learning and/or consolidation. D-cycloserine promoted long-term extinction retention

(Rothbaum et al., 2014; Smits et al., 2013a; Smits et al., 2013b) only if within-session learning was achieved.

With this pre-registered study, we follow the call for a stronger appreciation and more systematic investigations of measurement reliability (Zuo et al., 2019). We address longitudinal reliability and internal consistency as well as predictability of cross-phase responding in SCRs, fear ratings, and the BOLD response. For this purpose, we reanalyzed data from 120 participants that underwent a differential fear conditioning paradigm twice (at time points T0 and T1, 6 months apart) – with habituation and acquisition training on day 1 and extinction, reinstatement and reinstatement-test on day 2 to allow for fear memory consolidation prior to extinction. Part of the data have been used previously in method focused work (Kuhn et al., 2022; Lonsdorf et al., 2022; Lonsdorf et al., 2019a; Sjouwerman et al., 2022) and work investigating the association of conditioned responding with brain morphological measures (Ehlers et al., 2020).

Specifically, we (1) estimated internal consistency of SCRs at both time points and (2) systematically assessed longitudinal reliability of SCRs, fear ratings and BOLD fMRI at the individual level by calculating ICCs. This was complemented by investigations of response similarity (SCR and BOLD fMRI) and the degree of overlap of activated voxels at both time points (BOLD fMRI) as additional measurements of longitudinal reliability at the individual level that allow for a more detailed picture than the coarser ICCs (see **Table 1** for terminology and definitions). We also (3) assessed whether SCR and BOLD fMRI show longitudinal reliability at the group level. Finally, we (4) investigated if individual level responding during an experimental phase is predictive of individual-level responding during subsequent experimental phases. All hypotheses are tested across different pre-registered data specifications to account for procedural heterogeneity in the literature (see **Supplementary file 1**): More precisely, we follow a pre-registered multiverse-inspired approach and include (1) responses to the CS+, CS–, US, and CS discrimination, (2) different phase operationalizations, (3) different data transformations none, log-transformed, log-transformed and range-corrected, and (4) ordinally ranked vs. non-ranked data (for justification of these choices, see **Table 1**). We acknowledge that the specifications used here are not intended to cover all potentially meaningful combinations as in a full multiverse study (Lonsdorf et al., 2022; Sjouwerman et al., 2022; Steegen et al., 2016) but can be viewed as a manyverse (Kuhn et al., 2022) in which we a priori pre-registered a number of meaningful combinations.

Results

For a comprehensive overview of the different reliability measures used here and of the analyses conducted, see **Table 1**.

Satisfactory internal consistency

To assess internal consistency of SCRs, trials were split into odd and even trials (i.e., odd–even approach), averaged for each individual subject and then correlated (Pearson's correlation coefficient). This was done separately for each time point and experimental phase. Internal consistency at T0 (see **Figure 1A**) and T1 (see **Figure 1B**) of raw SCRs to the CS+ and CS– ranged from 0.54 to 0.85 and for raw SCRs to the US from 0.91 to 0.94 for all phases. In comparison, internal consistency was lower for CS discrimination with values ranging from –0.01 to 0.60. Log-transformation did not impact internal consistency but log-transformation in combination with range correction largely resulted in reduced reliability (see **Figure 1—figure supplement 1**).

Longitudinal reliability at the individual level

Longitudinal reliability at the individual level refers to the time stability of individual responses which we assessed through several measures (see **Table 1**).

As a first measure, absolute agreement ICCs (ICC_{abs}) and consistency ICCs (ICC_{con}) were calculated across both time points (T0, T1) for all data specifications (see **Figure 1**) while for BOLD fMRI these were only calculated for CS discrimination (see Materials and methods for justification). While ICC_{abs} refers to the extent to which measurements at T0 correspond with measurements at T1 in absolute terms, ICC_{con} allows for deviations at T1 due to systematic error (Koo and Li, 2016).

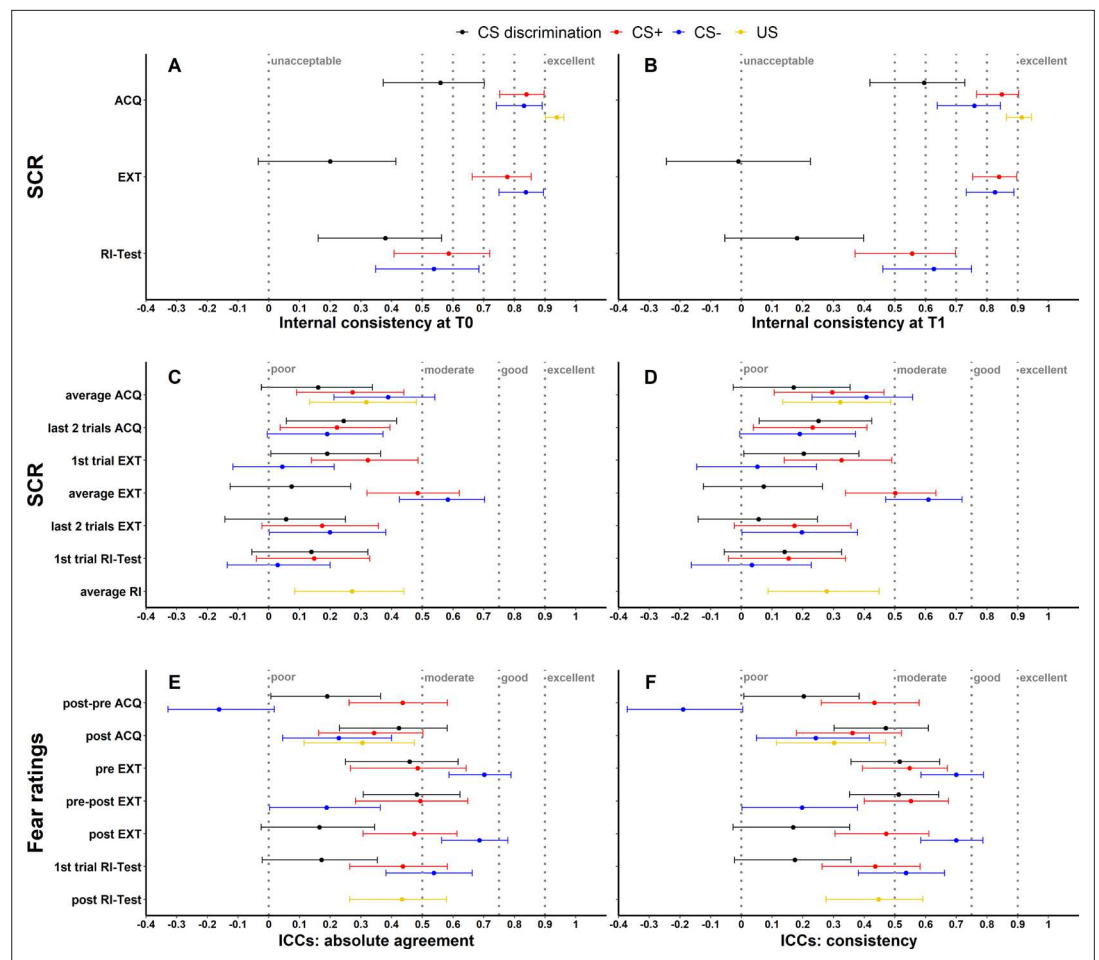


Figure 1. Illustration of internal consistency for skin conductance responses (SCRs) at T0 (A) and T1 (B) as well as ICC_{abs} and ICC_{con} for SCR (C, D) and fear ratings (E, F) color coded for stimulus type. Internal consistency indicates the reliability of responses within each time point, while intraclass correlation coefficients (ICCs) indicate the reliability across both time points. Note that assessment of internal consistency was not possible for fear ratings as only two ratings (pre, post) were available. Error bars represent 95% confidence intervals and indicate significance, when zero is not included in the interval. The y-axis comprises the different phases or phase operationalizations. In the literature, internal consistency is often interpreted using benchmarks (Kline, 2013) for unacceptable (<0.5), poor (>0.5 but <0.6), questionable (>0.6 but <0.7), acceptable (>0.7 but <0.8), good (>0.8 but <0.9), and excellent (≥ 0.9). Common benchmarks in the literature for ICCs are poor (<0.5), moderate (>0.5 but <0.75), good (>0.75 but <0.9), and excellent (≥ 0.9) (Koo and Li, 2016). These benchmarks are included here to provide a frame of reference but we point out that these benchmarks are arbitrary and most importantly derived from psychometric work on trait self-report measures and should hence not be overinterpreted in the context of responding in experimental paradigms which bear more sources of noise (Parsons, 2020). ACQ = acquisition training, EXT = extinction training, RI = reinstatement, RI-Test = reinstatement-test, pre = prior to the experimental phase, post = subsequent to the experimental phase.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Illustration of (A, B) internal consistency for log-transformed (log) as well as (C, D) log-transformed and range corrected (log rc) skin conductance responses (SCRs) at T0 and T1 color coded for stimulus type.

Figure supplement 2. Illustration of (A, B) intraclass correlation coefficients (ICCs) of log-transformed (log) as well as (C, D) log-transformed and range corrected (log, rc) skin conductance responses (SCRs) color coded for stimulus type.

Figure supplement 3. Illustration of ICC_{abs} of trial-by-trial raw skin conductance responses (SCRs) for phases (A–D: Acquisition, E–G: Extinction, H–J: Reinstatement-Test, K: Reinstatement) and stimulus types separately.

Figure 1 continued on next page

Figure 1 continued

Figure supplement 4. Illustration of ICC_{con} of trial-by-trial raw skin conductance responses (SCRs) for phases (A–D: Acquisition, E–G: Extinction, H–J: Reinstatement-Test, K: Reinstatement) and stimulus types separately.

Figure supplement 5. Illustration of ICC_{abs} of trial-by-trial log-transformed skin conductance responses (SCRs) for phases (A–D: Acquisition, E–G: Extinction, H–J: Reinstatement-Test, K: Reinstatement) and stimulus types separately.

Figure supplement 6. Illustration of ICC_{con} of trial-by-trial log-transformed skin conductance responses (SCRs) for phases (A–D: Acquisition, E–G: Extinction, H–J: Reinstatement-Test, K: Reinstatement) and stimulus types separately.

Figure supplement 7. Illustration of ICC_{abs} of trial-by-trial log-transformed and range corrected skin conductance responses (SCRs) for phases (A–D: Acquisition, E–G: Extinction, H–J: Reinstatement-Test, K: Reinstatement) and stimulus types separately.

Figure supplement 8. Illustration of ICC_{con} of trial-by-trial log-transformed and range corrected skin conductance responses (SCRs) for phases (A–D: Acquisition, E–G: Extinction, H–J: Reinstatement-Test, K: Reinstatement) and stimulus types separately.

Note that internal consistency and ICCs for SCRs are shown for raw data only. Results of log-transformed as well as log-transformed and range corrected data are presented in **Figure 1—figure supplement 1** and **Figure 1—figure supplement 2** for completeness.

SCR and fear ratings

Across data specifications, ICC_{abs} and ICC_{con} ranged from 0.03 to 0.58 and 0.03 to 0.61 for SCRs and from -0.16 to 0.70 as well as from -0.19 to 0.70 for fear ratings respectively (see **Figure 1**, for detailed results see also **Supplementary file 3** and **Supplementary file 4**). ICCs for log-transformed and raw SCRs were similar (see **Figure 1—figure supplement 2A-B**) while log-transformation and range correction resulted in increased reliability for some data specifications (e.g., CS+ and CS- responses averaged across acquisition training, see **Figure 1—figure supplement 2C-D**) but in reduced reliability for others (e.g., CS- responses during fear recall, i.e., the first extinction trial).

Exploratory, non-pre-registered analyses of trial-by-trial SCRs revealed, overall, only minor changes in ICCs upon stepwise inclusion of additional SCR trials (see **Figure 1—figure supplements 3–8**) with few exceptions: Including more trials resulted in an increase of ICC point estimates for SCRs to the CS+ and CS- during acquisition (log-transformed and range corrected data) and extinction training (all transformation types). Note, however, that this was – at large – only statistically significant when comparing ICCs based on the first (i.e., single trial at T0 and T1) and the maximum number of trials (as indicated by non-overlapping 95% confidence interval [CI] error bars). Interestingly, ICC point estimates for reinstatement-test (all transformation types) were numerically lower with an increasing number of trials, likely because of the transitory nature of the reinstatement effect (**Haaker et al., 2014**).

BOLD fMRI

For BOLD fMRI, both ICC types suggest rather limited reliability for CS discrimination during acquisition (both ICC_{abs} and ICC_{con} = 0.17) and extinction training (both ICC_{abs} and ICC_{con} = 0.01). For individual regions of interest (ROIs: anterior insula, amygdala, hippocampus, caudate nucleus, putamen, pallidum, nucleus accumbens [NAcc], thalamus, dorsal anterior cingulate cortex [dACC], dorsolateral prefrontal cortex [dlPFC], and ventromedial prefrontal cortex [vmPFC]), ICCs were even lower (all ICCs ≤ 0.001 ; for full results see **Supplementary file 5**).

Higher within- than between-subject similarity in BOLD fMRI but not SCRs

While ICCs provide information on the absolute quantity of longitudinal reliability at the individual level, comparison of within- and between-subject similarity as a complementary measure of longitudinal reliability at the individual level (see **Table 1**) reflects the extent to which responses in SCR and BOLD activation of one individual at T0 were more similar to themselves at T1 than to other individuals at T1 (see **Figures 2 and 3**).

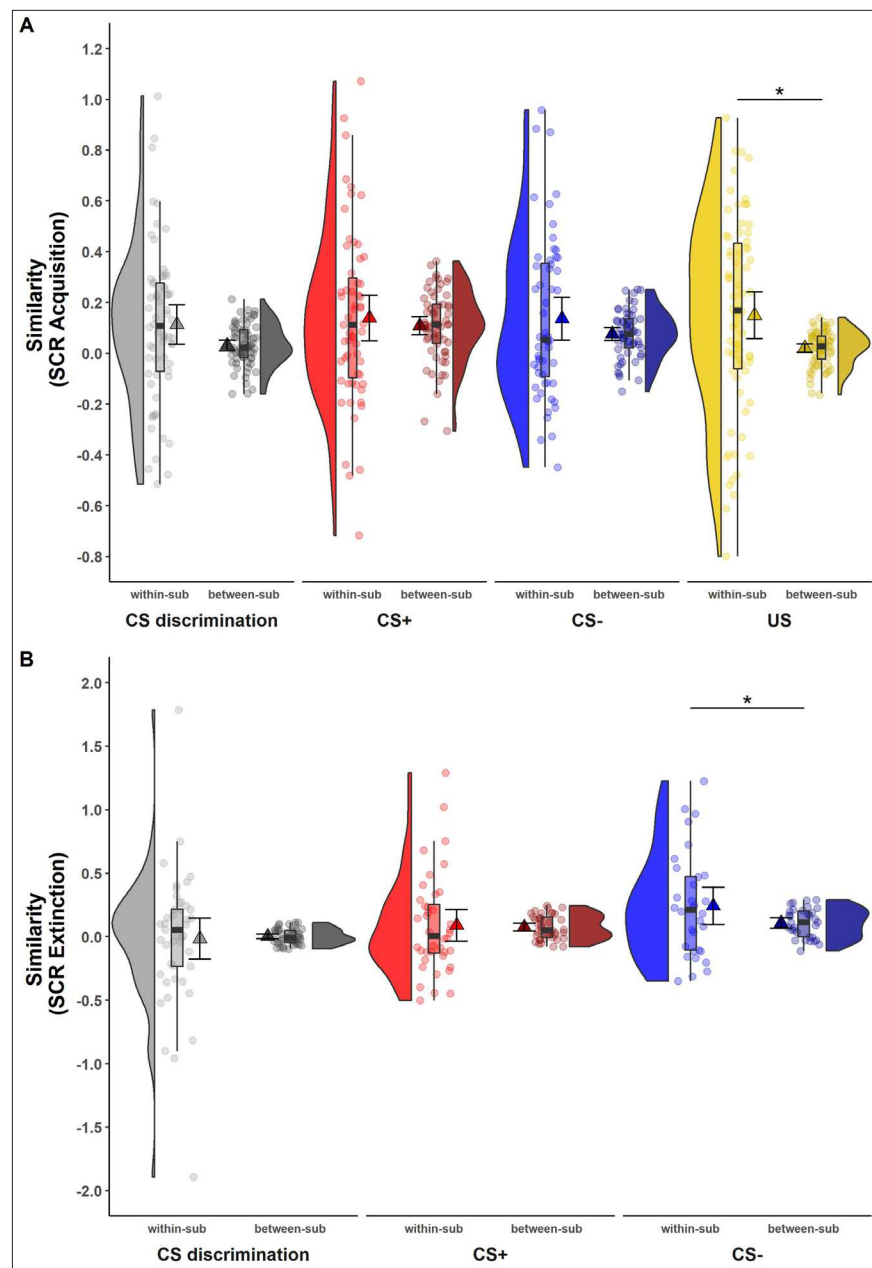


Figure 2. Illustration of within- and between-subject similarity for raw skin conductance responses (SCRs) during (A) acquisition and (B) extinction training separately for CS discrimination (gray), CS+ (red), CS- (blue), and unconditioned stimulus (US) responses (yellow). Results for log-transformed as well as log-transformed and range corrected SCR were almost identical to the results from raw data and are hence not reported here. Single data points represent Fisher r -to- z transformed correlations between single trial SCR of each subject at T0 and T1 (within-subject similarity) or averaged r -to- z transformed correlations between single trial SCR of one subject at T0 and all other subjects at T1 (between-subject similarity). Triangles represent mean correlations, corresponding error bars represent 95% confidence intervals. Boxes of boxplots represent the interquartile range (IQR) crossed by the median as bold line, ends of whiskers represent the minimum/maximum value in the data within the range of 25th/75th percentiles ± 1.5 IQR. Distributions of the data are illustrated by densities next to the boxplots. One data point had a similarity above 3.5 (within-subject similarity of SCR to the CS+) and is not shown in the figure. * $p < 0.05$. Note that the variances differ strongly between within- and between-subject similarity because between-subject similarity is based on correlations averaged across subjects, whereas within-subject similarity is based on non-averaged correlations calculated for each subject. Note also that similarity calculations were based on different sample sizes for acquisition and extinction training and CS discrimination as well as SCR to the CS+, CS-, and US, respectively (for details, see Materials and methods). within-sub = within-subject; between-sub = between-subject.

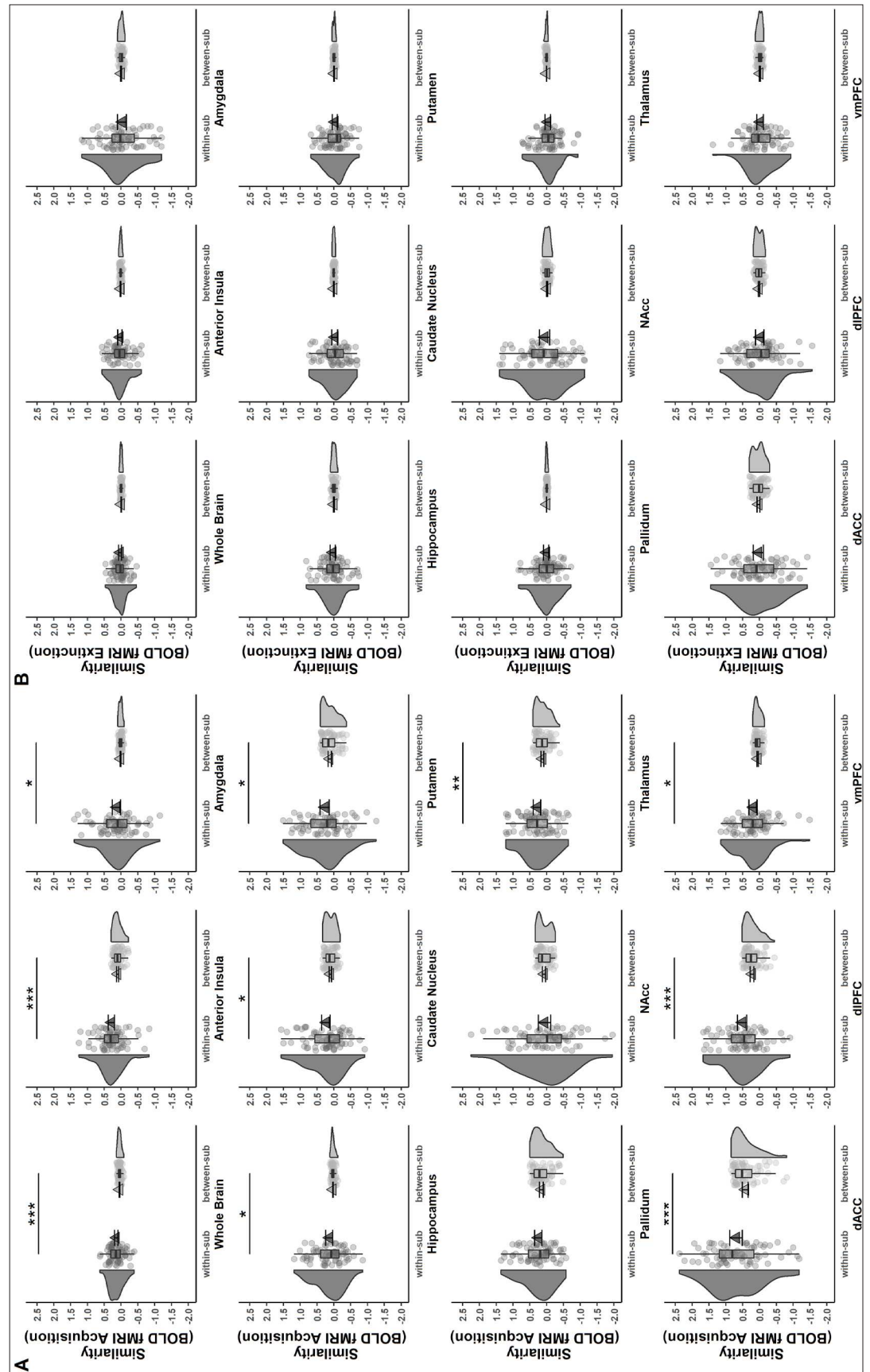


Figure 3. Acquisition (A) and extinction (B) training within- and between-subject similarities (Fisher *r*-to-*z* transformed) of voxel-wise brain activation patterns (based on beta maps) for CS discrimination at T0 and T1 for the whole brain and different regions of interest (ROIs). Triangles represent mean correlations, corresponding error bars represent 95% confidence intervals. Single data points represent Fisher *r*-to-*z* transformed correlations between the first-level response patterns of brain activation of each subject at T0 and T1 (within-subject similarity) or averaged *r*-to-*z* transformed correlations between the first-level response patterns of brain activation of one subject at T0 and all other subjects at T1 (between-subject similarity). Boxes of boxplots represent the interquartile range (IQR) crossed by the median as bold line, ends of whiskers represent the minimum/maximum value in the data within the range of 25th/75th percentiles ± 1.5 IQR. Distributions of the data are illustrated with densities next to the boxplots. fMRI data for the reinstatement-test were not analyzed in the current study since data from a single trial do not provide sufficient power. **p* < 0.05, ***p* < 0.01, ****p* < 0.001. NAcc = nucleus accumbens; dACC = dorsal anterior cingulate cortex; dlPFC = dorsolateral prefrontal cortex; vmPFC = ventromedial prefrontal cortex; within-sub = within-subject; between-sub = between-subject.

SCR

For SCRs, within-subject similarity (i.e., within-subject correlation of trial-by-trial SCR across time points) and between-subject similarity (i.e., correlation of trial-by-trial SCR between one individual at T0 and all other individuals at T1; see **Figure 2**) did not differ significantly for most data specifications. This was true for CS discrimination ($t(64) = 1.78, p = 0.079, d = 0.22$) as well as for SCRs to the CS+ ($t(61) = 0.84, p = 0.407, d = 0.11$) and CS- ($t(55) = 1.50, p = 0.138, d = 0.20$) during acquisition training and for CS discrimination ($t(44) = -0.23, p = 0.823, d = -0.03$) and SCRs to the CS+ ($t(39) = 0.25, p = 0.801, d = 0.04$) during extinction training. This indicates that SCRs of one particular individual at T0 were mostly not more similar to their own SCRs than to those of other individuals at T1. The only exceptions where within-subject similarities were significantly higher than between-subject similarity were SCRs to the US during acquisition training ($t(70) = 2.54, p = 0.013, d = 0.30$) and to the CS- during extinction training ($t(31) = 2.05, p = 0.049, d = 0.36$). Note, however, that within-subject similarity had a very wide spread pointing to substantial individual differences (while this variance is removed in calculations of between-subject similarity).

fMRI data

In contrast to what was observed for SCRs, within-subject similarity was significantly higher than between-subject similarity in the whole brain ($p < 0.001$) and most of the ROIs for fear acquisition training (see **Figure 3A** and **Supplementary file 6**). This suggests that while absolute values for similarity might be low, individual brain activation patterns during fear acquisition training at T0 were – at large – still more similar to the same subject’s activation pattern at T1 than to any others at T1. For extinction training, however, no significant differences between within- and between-subject similarity were found for any ROI or the whole brain (all *p*’s > 0.306; see **Figure 3B** and **Supplementary file 6**).

Table 2. Overlap in significantly activated voxels at the individual and group level across both time points for CS discrimination.

Level	Phase	Coeff.	ROI												
			Whole brain	Insula	Amygdala	Hippocampus	Caudate	Putamen	Pallidum	Accumbens	Thalamus	dACC	dlPFC	vmPFC	
(A) Individual	Acq	Jaccard	0.076	0.075	0.011	0.012	0.039	0.037	0.018	0.017	0.033	0.132	0.080	0.039	
		Dice	0.131	0.121	0.018	0.021	0.057	0.058	0.029	0.024	0.055	0.189	0.118	0.061	
	Ext	Jaccard	0.007	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.003	0.001	0.005	
		Dice	0.014	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.001	0.006	0.002	0.009
	(B) Group	Acq	Jaccard	0.620	0.595	0.294	0.323	0.613	0.740	0.747	0.441	0.834	0.898	0.895	0.045
			Dice	0.765	0.745	0.448	0.472	0.760	0.847	0.855	0.595	0.910	0.946	0.944	0.086
(B) Group	Ext	Jaccard	0.057	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.044	0.014	0.000	
		Dice	0.108	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.085	0.028	0.000	

Note. Results are shown for the whole brain as well as for selected regions of interest (ROIs) for fear acquisition training and extinction training. Both coefficients range from 0 (no overlap) to 1 (perfect overlap). Note that the Jaccard can be interpreted as % (Maitra, 2010). NAcc = nucleus accumbens; dACC = dorsal anterior cingulate cortex; dlPFC = dorsolateral prefrontal cortex; vmPFC = ventromedial prefrontal cortex.

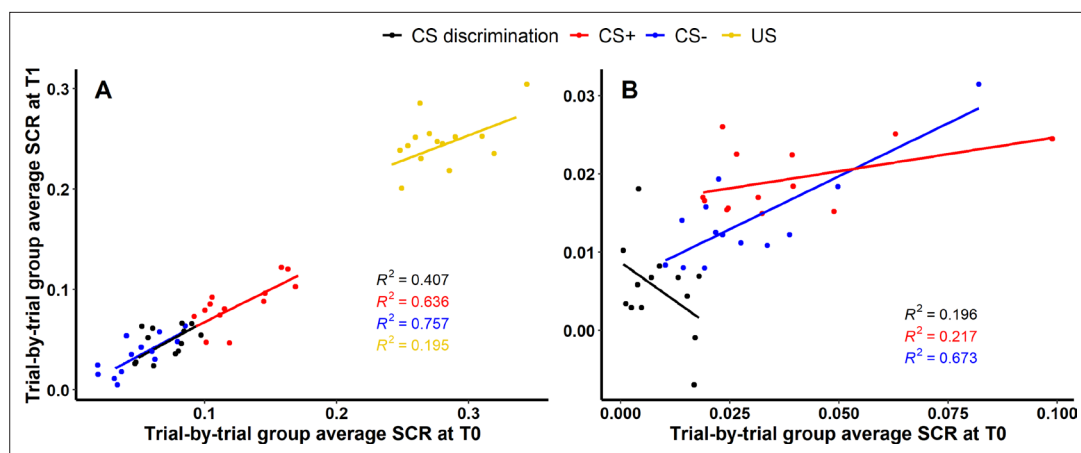


Figure 4. Scatter plots illustrating longitudinal reliability at the group level during (A) acquisition and (B) extinction training for raw skin conductance responses (SCRs) (in μS). Results for log-transformed as well as log-transformed and range corrected data are presented in **Figure 4—figure supplement 1**. Longitudinal reliability at the group level refers to the extent of explained variance in linear regressions comprising SCRs at T0 as independent and SCRs at T1 as dependent variable. Results are shown for trial-by-trial group average SCRs to the CS+ (red), CS- (blue), the unconditioned stimulus (US; yellow), and CS discrimination (black). Single data points represent pairs of single trials at T0 and T1 averaged across participants. Note that no US was presented during extinction training and hence, no reliability of the US is shown in (B).

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Scatter plots illustrating longitudinal reliability at the group level during (A, C) acquisition and (B, D) extinction training for log-transformed (A, B) as well as log-transformed and range corrected (C, D) skin conductance responses (SCRs).

Low overlap at the individual level between both time points

As opposed to similarity measures (see above) which reflect the correlation of activated voxels between time points, overlap at the individual level denotes the degree of overlap of significantly activated voxels.

The overlap at the individual level was low with the Jaccard coefficient indicating 7.60% and 0.70% whole brain overlap for acquisition and extinction training, respectively (see **Table 2A**). Of note, individual values ranged from 0% to 39.65% overlap during acquisition, suggesting large interindividual differences in overlap.

While overlap during acquisition for individual ROIs was comparable to the whole brain, Jaccard and Dice coefficients indicate close to 0 overlap at extinction (see **Table 2A**).

Robust longitudinal reliability at the group level

While longitudinal reliability at the individual level relies on (mean) individual subject responding at both time points, longitudinal reliability at the group level relies on the percentage of explained variance of group averaged trials at T1 by group averaged trials at T0 (i.e., R squared for SCR) or the degree of group level overlap of significant voxels expressed as Dice and Jaccard indices (i.e., BOLD fMRI).

SCR

For acquisition training (see **Figure 4A**), 40.66% ($F(1, 11) = 7.54$, $p = 0.019$), 63.59% ($F(1, 11) = 19.21$, $p = 0.001$) and 75.67% ($F(1, 11) = 34.20$, $p < 0.001$) of the variance of SCRs at T1 could be explained by SCRs at T0 for CS discrimination, CS+ and CS-, respectively, indicating robust longitudinal reliability of SCRs at the group level for CS responding during acquisition. Interestingly, only 19.53% ($F(1, 12) = 2.91$, $p = 0.114$) of the variance of SCRs to the US could be explained. For extinction training, in contrast, only 19.58% ($F(1, 11) = 2.68$, $p = 0.130$) and 21.70% ($F(1, 11) = 3.05$, $p = 0.109$) of the SCR variance at T1 could be explained by SCRs at T0 for CS discrimination and CS+, respectively, indicating only limited longitudinal reliability at the group level. However, with 67.35% ($F(1, 11) = 22.69$,

$p = 0.001$) explained variance at T1, longitudinal reliability of SCRs to the CS- appeared to be more robust as compared to CS discrimination and responses to the CS+ (see **Figure 4B**).

BOLD fMRI

In stark contrast to the low overlap of individual-level activation (see **Table 2A**), the overlap at the group level was rather high with 62.00% for the whole brain and up to 89.80% for individual ROIs (i.e., dACC and dlPFC; Jaccard) for CS discrimination during acquisition training (see **Table 2B**). Similar to what was observed for overlap at the individual level, a much lower overlap for extinction training as compared to acquisition training was observed for the whole brain (5.70% overlap) and all ROIs (all close to zero).

Cross-phases predictability of conditioned responding

Finally, we investigated if responding in any given experimental phase predicted responding in subsequent experimental phases. To this end, simple linear regressions with robust standard errors were computed for both SCRs and fear ratings and all data specifications (see **Figure 5** and **Supplementary file 7, Supplementary file 8**). To approximate these analyses, correlations of patterns of BOLD brain activation between experimental phases were calculated (see **Figure 6**).

SCR

Stronger CS discrimination in SCRs during (delayed) fear recall (i.e., first trial of extinction training) was significantly predicted by both average and end-point performance (i.e., last two trials) during acquisition training for most data specifications (**Figure 5A**, columns 1 and 2). In contrast, average CS discrimination during extinction training was significantly predicted by acquisition training performance only if data were ordinally ranked (columns 3 and 4). Strikingly, all predictions of extinction end-point performance (columns 5 and 6) as well as performance at reinstatement-test (columns 7–11) were non-significant – irrespective of phase operationalizations and data transformation.

The majority of predictions of SCRs to the CS+ and CS- were significant with few exceptions (see white cells in **Figure 5A**) – irrespective of experimental phases, their operationalization and data transformation. Most non-significant regressions included log-transformed and range corrected data. Strikingly, extinction end-point performance never predicted performance at reinstatement-test – irrespective of data transformation (column 11).

Fear ratings

Higher ratings for the CS+ as well as higher CS discrimination during acquisition training predicted higher CS+ ratings and CS discrimination at fear recall (**Figure 5B**, columns 1 and 2), extinction training (columns 3 and 4), and at reinstatement-test (columns 7 and 8). Higher responding to the CS+ and higher CS discrimination at fear recall predicted higher responding at reinstatement-test (column 9) – irrespective of data transformations. In contrast, predictions of CS discrimination and CS+ ratings after extinction training were mostly non-significant (columns 5 and 6). Higher CS+ ratings during extinction training significantly predicted higher ratings at reinstatement-test which was not true for CS discrimination (columns 10 and 11).

Higher CS- ratings after acquisition training predicted higher CS- ratings at fear recall as well as after extinction training and CS- ratings after extinction training predicted the performance at reinstatement-test – irrespective of ranking of the data (columns 2, 6, and 11). Furthermore, when based on ordinally ranked data, the difference between ratings prior to and after acquisition predicted CS- ratings at fear recall and CS- ratings after acquisition training predicted the difference between CS- ratings prior to and after extinction training (columns 1 and 4). All other predictions were non-significant.

In sum, all significant predictions observed were positive with weak to moderate associations and indicate that higher responding in preceding phases predicted higher responding in subsequent phases for both SCRs and fear ratings.

BOLD fMRI

In short, all but one association (CS discrimination in the NAcc) was positive, showing that higher BOLD response during acquisition was associated with higher BOLD responding during extinction

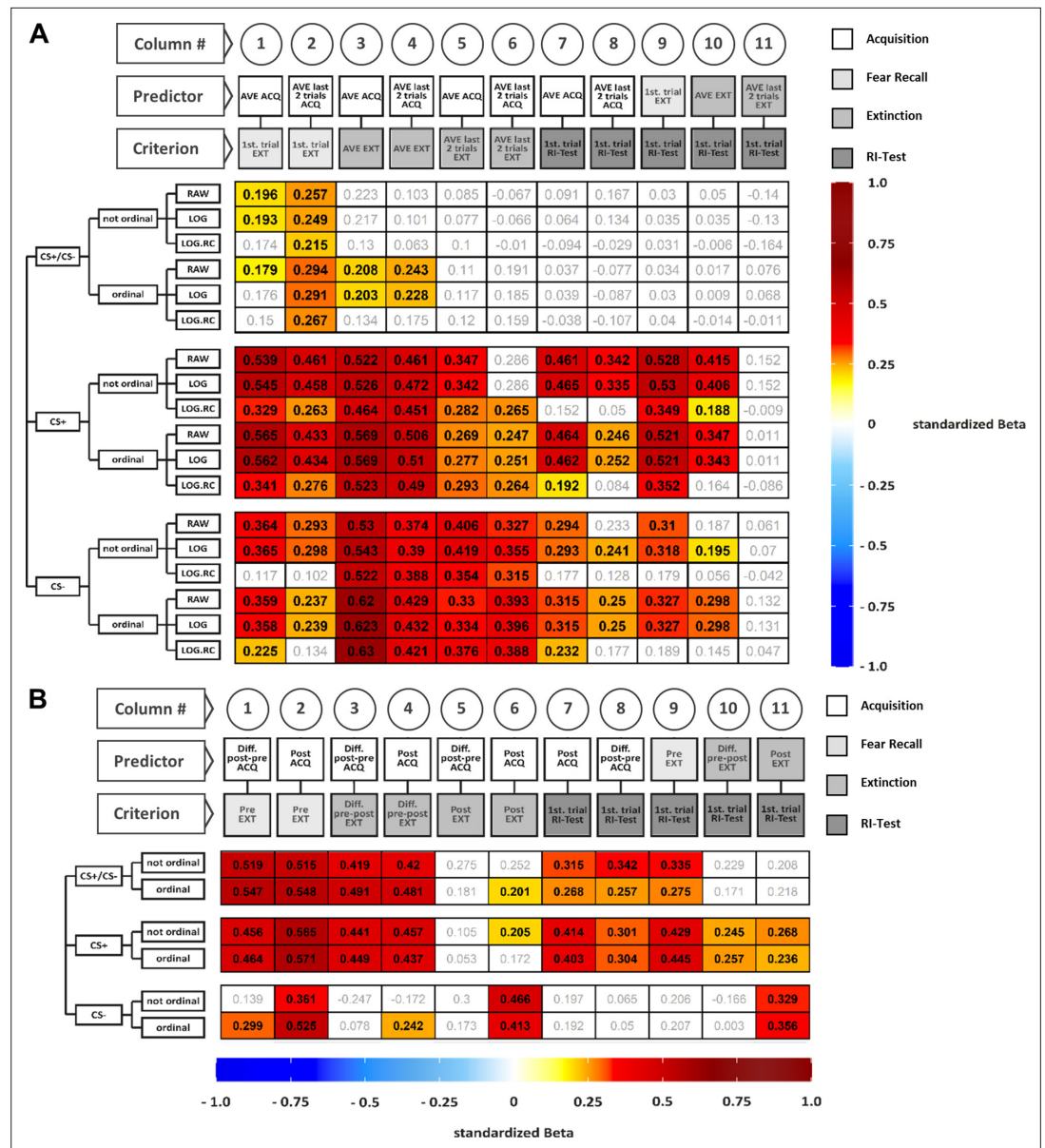


Figure 5. Illustration of standardized betas derived from regressions including skin conductance responses (SCRs) (A) and fear ratings (B) for all data specifications. Colored cells indicate statistical significance of standardized betas, non-colored cells indicate non-significance. Standardized betas are color coded for their direction and magnitude showing positive values from yellow to red and negative values from light blue to dark blue. Darker colors indicate higher betas. On the y-axis, the following data specifications are plotted from left to right: CS type, ranking of the data and transformation of the data. On the x-axis, the following information is plotted: Number of the columns for better orientation, predictor, and criterion included in the regression. For example, the beta value at the top left in (A) (i.e., 0.196) is the standardized beta as retrieved from the linear regression including CS discrimination in non-ranked and raw SCR during average acquisition as predictor and the first extinction trial as criterion. For exploratory non-preregistered regressions including a small manyverse of approximations of SCR extinction training learning rates, see **Figure 5—figure supplement 1**. Tables containing regression parameters beyond the standardized betas depicted in panels A and B are presented in **Supplementary file 7** and **Supplementary file 8**. AVE = average, LOG = log-transformed data, LOG.RC = log-transformed and range corrected data, not ordinal = not ordinally ranked data, ordinal = ordinally ranked data.

The online version of this article includes the following figure supplement(s) for figure 5:

Figure supplement 1. As per reviewer’s request, we illustrate standardized betas derived from non-pre-registered regressions including skin conductance response (SCR) extinction training learning rates (LR EXT).

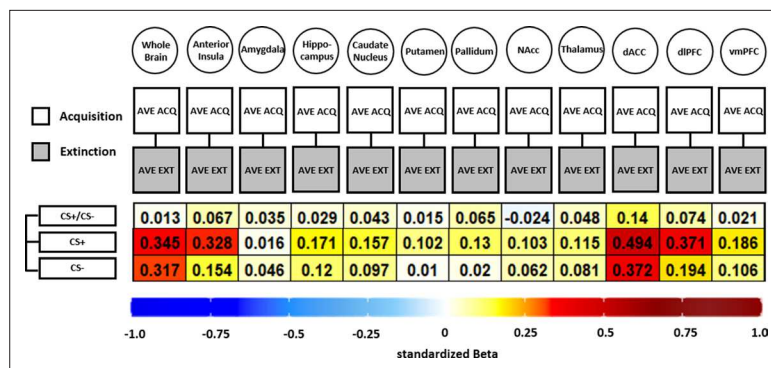


Figure 6. Illustration of standardized betas derived from correlation analyses between brain activation patterns during acquisition and extinction training in different regions of interest (ROIs) and different data specifications. Standardized betas are color coded for their direction and magnitude showing positive values from yellow to red and negative values from light blue to dark blue. Darker colors indicate higher betas. NAcc = nucleus accumbens; dACC = dorsal anterior cingulate cortex; dlPFC = dorsolateral prefrontal cortex; vmPFC = ventromedial prefrontal cortex.

training (see **Figure 6**). However, the standardized beta coefficients are mostly below or around 0.3 except for CS+ associations in the dACC, indicating non-substantial associations for all ROIs and CS specifications that were near absent for CS discrimination. Analysis of CS+ and CS- data was included here as the analysis is based on beta maps and not T-maps (as in previous analyses) where a contrast against baseline is not optimal.

Cross-phases predictability depends on data specifications

Pooled across all other data specifications, some interesting patterns can be extracted: First, standardized betas were significantly lower for raw ($t(65) = 8.08$, $p < 0.001$, $d = 0.99$) and log-transformed ($t(65) = 8.26$, $p < 0.001$, $d = 1.02$) as compared to log-transformed and range corrected SCRs while standardized betas derived from the former did not differ significantly ($t(65) = -0.26$, $p = 0.794$, $d = -0.03$). Second, standardized betas derived from ranked and non-ranked analyses were comparable for fear ratings ($t(32) = 1.26$, $p = 0.218$, $d = 0.22$) but not for SCRs with significantly higher betas for non-ranked as opposed to ranked SCRs ($t(98) = 2.37$, $p = 0.020$, $d = 0.24$). Third, standardized betas for CS discrimination were significantly lower than for CS+ and CS- for both SCRs (CS+: $t(65) = -15.31$, $p < 0.001$, $d = -1.88$ and CS-: $t(65) = -12.34$, $p < 0.001$, $d = -1.52$) and BOLD fMRI (CS+: $t(11) = -4.65$, $p < 0.001$, $d = -1.34$ and CS-: $t(11) = -3.05$, $p = 0.011$, $d = -0.88$), while for ratings, standardized betas for CS discrimination were higher than for the CS- ($t(21) = 3.11$, $p = 0.005$, $d = 0.66$) and comparable to those for the CS+ ($t(21) = -0.57$, $p = 0.572$, $d = -0.12$). Furthermore standardized betas were larger for the CS+ than for the CS- for SCRs ($t(65) = 3.79$, $p < 0.001$, $d = 0.47$), ratings ($t(21) = 3.12$, $p = 0.005$, $d = 0.67$) and BOLD fMRI ($t(11) = 4.34$, $p = 0.001$, $d = 1.25$). Fourth, standardized betas derived from regressions predicting fear recall were significantly higher than for reinstatement-test for both SCRs ($t(124) = 4.35$, $p < 0.001$, $d = 0.86$) and fear ratings ($t(40) = 5.15$, $p < 0.001$, $d = 1.76$).

Discussion

In fear conditioning research, little is known about longitudinal reliability (in the literature often referred to as test-retest reliability) for common outcome measures and almost nothing is known about their internal consistency and to what extent predictability across experimental phases is possible.

Here, we aimed to fill this gap and complement traditionally used approaches focusing on ICCs (summarized in **Supplementary file 1**) with (1) analyses of response similarity, (2) the degree of overlap of individual-level brain activation patterns as well as (3) by exploring longitudinal reliability at the group level in addition to (4) internal consistency across outcome measures.

Moreover, we also directly investigated predictability of responding from one experimental phase to subsequent experimental phases. For all analyses, we followed a multiverse-inspired approach (Parsons, 2020) by taking into account different data specifications.

Overall, longitudinal group-level reliability was robust for SCRs (see **Figure 4**) and the BOLD response (see **Table 2B**) while longitudinal individual-level reliability as assessed by ICCs (see **Figure 1C–F**), and individual-level BOLD activation overlap (see **Table 2A**) was more limited across outcome measures and data specifications – particularly during extinction training. This is in line with previous work in fear conditioning (Cooper et al., 2022b; Fredrikson et al., 1993; Ridderbusch et al., 2021; Torrents-Rodas et al., 2014; Zeidan et al., 2012) reporting figures for longitudinal individual-level reliability comparable to ours across outcome measures (SCRs, fear ratings, BOLD fMRI) and experimental phases. Importantly, however, it remains a challenge to interpret the results as benchmarks for ICCs are derived from psychometric work on trait self-report measures and it is plausible that what is interpreted as ‘low’ and ‘high’ reliability in experimental work should be substantially lower (Parsons et al., 2019).

Our complementary analyses beyond traditional ICCs indicate that SCRs of one individual at T0 were not more similar to responses of the same individual at T1 than compared to others at T1 (see **Figure 2**). For BOLD fMRI, however, acquisition-related individual BOLD activation patterns at T0 were more similar to their own activation patterns at T1 than to other individuals’ activation patterns (see **Figure 3**). This was, however, not the case for extinction. Hence, this may suggest that BOLD fMRI might be more sensitive to detect similarity at individual-level responses within participants than SCRs in our data – maybe due to the dependence on spatial (i.e., voxel-by-voxel) rather than temporal (i.e., trial-by-trial) patterns.

Furthermore, we observed a few differences in longitudinal reliability at the individual level depending on data processing specifications (see also Parsons, 2020). For most data specifications, reliability was slightly higher for log-transformed and range-corrected SCRs (as opposed to raw and only log-transformed data) while – in contrast to what has been shown for other paradigms and outcome measures (Baker et al., 2021; see also <https://shiny.york.ac.uk/powercontours/>) – an increasing number of trials included in the calculation of ICCs did not generally improve reliability (see **Figure 1—figure supplements 3–8**). Together, this suggests that longitudinal reliability at the individual level is relatively stable across different data transformations and paradigm specifications (e.g., number of trials within the range used here, i.e., 1 to maximum 14) which is important information facilitating the integration of previous work using different time intervals, reliability indices, and paradigms (see **Supplementary file 1**; Cooper et al., 2022b; Fredrikson et al., 1993; Ridderbusch et al., 2021; Torrents-Rodas et al., 2014; Zeidan et al., 2012).

In contrast, we observed quite robust longitudinal reliability at the group level for both SCRs (see **Figure 4**) and BOLD fMRI (see **Table 2B**) between both time points with substantial (i.e., up to 90%) overlap in group-level BOLD fMRI activation patterns (whole brain and ROI based) as well as substantial (i.e., up to 76%) explained variance at T1 by variance at T0 for SCRs. However, this was generally only true for acquisition but not extinction training. This pattern of higher reliability during acquisition compared to extinction training has been described in the literature (SCRs: Fredrikson et al., 1993; Zeidan et al., 2012) and was also evident in the similarity analyses of BOLD fMRI and the group-level reliability of SCRs. While this pattern did not emerge across all analyses, it appears to be particularly present when examining reliability of CS discrimination as it was the case for BOLD fMRI and as it also emerged in individual-level reliability analyses of CS discrimination in SCRs (internal consistency and ICCs) and fear ratings (ICCs). Since CS discrimination is typically lower during extinction as compared to acquisition training, this restriction of variance potentially resulted in a floor effect which might have lowered the internal consistency and longitudinal reliability of CS discrimination during extinction training.

Reports regarding this discrepancy between group- and individual-level longitudinal reliability were recently highlighted for a number of (classic) experimental paradigms (Fröhner et al., 2019; Hedger et al., 2018; Herting et al., 2018; Plichta et al., 2012; Schumann et al., 2020). Our results add fear conditioning and extinction as assessed by SCRs and BOLD fMRI to this list and have important implications for translational questions aiming for individual-level predictions – particularly since findings obtained at the group level are not necessarily representative for any individual within the group (Fisher et al., 2018).

In addition to these methods-focused insights, we observed significant associations between responding in different experimental phases for SCR (see **Figure 5A**), fear ratings (see **Figure 5B**) and

BOLD fMRI (see **Figure 6**) revealing that higher responses in previous phases were generally modestly associated with higher responses in subsequent phases in all outcome measures. However, a remarkable amount of predictions were non-significant – which was particularly true for CS discrimination in SCRs and BOLD fMRI. This may be explained by difference scores (i.e., CS+ minus CS–) being generally less reliable (*Infantino et al., 2018; Lynam et al., 2006*) due to a subtraction of meaningful variance (*Moriarty and Alloy, 2021*) particularly in highly correlated predictors (*Thomas and Zumbo, 2012*). Especially at the end of the extinction, CS discrimination is low and hence, variance limited. Therefore, floor effects may contribute to the non-significant effects for extinction end-point performance.

Mixed findings in the literature support both the independence of conditioned responding in different experimental phases (*Bouton et al., 2006; Plendl and Wotjak, 2010; Prenoveau et al., 2013; Shumake et al., 2014*) but also their dependence – particularly in clinical samples (*Foa et al., 1983; Rauch et al., 2004; Rothbaum et al., 2014; Smits et al., 2013a; Smits et al., 2013b*). These diverging findings in experimental and clinical studies might point toward a translational gap. However, our work may suggest that the strengths of associations between responding in different phases depended on the specific outcome measure and its specifications (e.g., responses specified as CS discrimination, CS+, or CS–). Yet another explanation – in particular for predictions spanning a 24 hrs delay in experimental phases – might be that individual differences in consolidation efficacy (e.g., how efficiently the fear and extinction memories are consolidated after performing acquisition and extinction training, respectively) may underlie differences in predictability. For example, the performance during a retention or RoF test phase is considered to be determined by the strength of the fear and extinction memory, respectively. Memory strength, however, is not only determined by the strength of the initially acquired memory but also by its consolidation (discussed in *Lonsdorf et al., 2019b*). Thus, as acquisition training preceded the extinction training and reinstatement-test by 24 hrs, it is highly likely that individual differences in consolidation efficacy also impact on performance at test. This has also implications for the common practice of correcting responses during one experimental phase for responding during preceding experimental phases (discussed in *Lonsdorf et al., 2019b*).

Importantly, together with our observation of robust internal consistency (see **Figure 1** and also *Fredrikson et al., 1993*), this pattern of findings suggests that individual-level predictions at short intervals are plausible but might be more problematic for longer time periods as suggested by the limited stability over time in our data.

Yet, we would like to point out that the values we report may in fact point toward good and not limited longitudinal individual-level reliability as our interpretation is guided by benchmarks that were not developed for experimental data but from psychometric work on trait self-report measures. We acknowledge that the upper bound of maximally observable reliability may differ between both cases of application as empirical neuroscientific research inherently comes with more noise. The problem remains that predictions in fear conditioning paradigms appear to not be meaningful for longer periods of time (~6 months). Thus, a key contribution of our work is that it highlights the need to pay more attention to measurement properties in translational research in general and fear conditioning research specifically (e.g., implement reliability calculations routinely in future studies). To date, it remains an open question what ‘good reliability’ in experimental neuroscientific work actually means (*Parsons et al., 2019*).

Yet, before discussing implications of our results in detail, some reflections on potential (methodological) reasons for (1) limited individual-level but robust group-level reliability and (2) on the role of time interval lengths deserve attention:

First, the limited longitudinal individual-level reliability might indicate that the fear conditioning paradigm employed here – which is a rather strong paradigm with 100% reinforcement rate – may be better suited for investigations of group effects and to a lesser extent for individual difference questions – potentially due to limited variance between individuals (*Hedge et al., 2018; Parsons, 2020; Parsons et al., 2019*). However, high reliability appears to be possible in principle, as we can conclude from the robust internal consistency of SCRs that we observed. This speaks against a limited between-subject variance and a general impracticability of the paradigm for individual difference research. Hence, we call for caution and warn against concluding from our report that fear conditioning and our outcome measures (SCRs, BOLD fMRI) are unreliable at the individual level.

Second, limited individual-level but robust group-level longitudinal reliability might be (in part) due to different averaging procedures which impacts error variance (**Kennedy et al., 2021**). More precisely, compared to individual-level data, group-level data are based on highly aggregated data resulting in generally reduced error variance which increases group-level reliability.

Third, different operationalizations of the same measurement might have different reliabilities (**Kragel et al., 2021**). For instance, amygdala habituation has been shown to be a more reliable measure than average amygdala activation (**Plichta et al., 2014**) and more advanced analytical approaches such as intraindividual neural response variability (**Månsson et al., 2021**) and multivariate imaging techniques (**Kragel et al., 2021; Marek et al., 2020; Noble et al., 2021; Visser et al., 2021**) have been suggested to have better (longitudinal) reliability than more traditional analyses approaches. Similarly, methodological advances (e.g., techniques to adjust the functional organization of the brain across participants, **Kong et al., 2021**; or hyperalignment, **Feilong et al., 2021**) in measurement quality and tools may ultimately result in better reliability estimates (**DeYoung et al., 2022**).

Fourth, as discussed above, caution is warranted as traditional benchmarks for ‘good’ reliability were not developed for experimental work but mainly from psychometric work on trait self-report measures (see above).

Finally, longitudinal reliability refers to measurements obtained under the same conditions and hence it is both plausible and well established that higher reliability is observed at short test–retest intervals (see also **Noble et al., 2021; Werner et al., 2022**). Longer intervals are more susceptible to true changes of the measurand – for instance due to environmental influences such as seasonality, temperature, hormonal status, or life events (see **Specht et al., 2011; Vaidya et al., 2002**). Indeed most longitudinal reliability studies in the fMRI field used shorter intervals (<6 months, see **Elliott et al., 2020; Noble et al., 2021**) than our 6-month interval and hence our results should be conceptualized as longitudinal stability rather than a genuine test–retest reliability. The satisfactory internal consistency speaks against excessive noisiness inherent to our measures as a strong noisiness would also be evident in measurements within one time point and not only emerge across our retest interval. Thus, we rather suggest a true change of the measurand during our retest interval and hence a potentially stronger state than trait dependency.

What do our findings imply? Fear conditioning research has been highlighted as a particularly promising paradigm for the translation of neuroscientific findings into the clinics (**Anderson and Insel, 2006; Cooper et al., 2022a; Fullana et al., 2020; Milad and Quirk, 2012**) and some of the most pressing translational questions are based on individual-level predictions such as predicting treatment success. Our results, however, suggest that measurement reliability may allow for individual-level predictions for (very) short but potentially less so for longer time intervals (such as our 6 months retest interval). Importantly, however, robust group-level reliability appears to allow for group-level predictions over longer time intervals. This applies to SCRs and BOLD fMRI in our data but note that the latter was not investigated for fear ratings. A potential solution and promising future avenue to make use of both good group-level reliability and individual-level predictions might be the use of homogenous (latent) subgroups characterized by similar response profiles (e.g., rapid, slow or no extinction, **Galatzer-Levy et al., 2013a**) – to exploit the fact that reliability appears to be higher for more homogenous samples (**Gulliksen, 1950**).

While general recommendations and helpful discussions on the link between reliability and number of trials (**Baker et al., 2021**), statistical power (**Parsons, 2020**), maximally observable correlations (**Parsons, 2020**), sample and effect size (**Hedge et al., 2018; Parsons, 2020**) considerations exist, our results highlight the need for field and subdiscipline specific considerations. Our work allows for some initial recommendations and insights. First, we highlight the value of using multiple, more nuanced measures of reliability beyond traditional ICCs (i.e., similarity, overlap, **Fröhner et al., 2019**) and second, the relation between number of trials and reliability in an experiment with a learning component (i.e., no increase in reliability with an increasing number of trials). Importantly, our work can also be understood as an empirically based call for action, since more work is needed to allow for clear-cut recommendations, and as a starting point to develop and refine comprehensive guidelines in the future. We also echo the cautionary note of Parsons that ‘estimates of reliability refer to the measurement obtained – in a specific sample and under particular circumstances, including the task parameters’ (cf. **Parsons, 2020**). Hence, it is important to remember that reliability is a property of a measure that is not fixed and may vary depending on

task specifications and samples. In other words, reliability is not a fixed property of the task itself, here fear conditioning.

We argue that we may need to take a (number of) step(s) back and develop paradigms and data processing pipelines explicitly tailored to individual difference research (i.e., correlation) or experimental (i.e., group level) research questions (e.g., [Parsons, 2020](#)) and focus more strongly on measurement reliability in experimental work – which has major consequences on effect sizes and statistical power ([Elliott et al., 2020](#)). More precisely, multiverse-type investigations ([Parsons, 2020](#); [Steege et al., 2016](#)) that systematically scrutinize the impact of several alternative and equally justifiable processing and analytical decisions in a single dataset ([Kuhn et al., 2022](#); [Lonsdorf et al., 2022](#); [Sjouwerman et al., 2022](#)) – as also done here for transformations and number of trials – may be helpful to ultimately achieve this overarching aim. This could be complemented by systematically varying design specifications ([Harder, 2020](#)) which are extensively heterogeneous in fear conditioning research ([Lonsdorf et al., 2017a](#)). Calibration approaches, as recently suggested [Bach et al., 2020](#) follow a similar aim.

Such work on measurement questions should be included in cognitive-experimental work as a standard practice ([Parsons, 2020](#)) and can (often) be explored in a cost and resource effective way in existing data which in the best case are openly available – which, however, requires cross-lab data sharing and data management homogenization plans. Devoting resources and funds to measurement optimization is a valuable investment into the prospect of this field contributing to improved mental health ([Moriarty and Alloy, 2021](#)) and to resume the path to successful translation from neuroscience discoveries into clinical applications.

Materials and methods

Pre-registration

This project has been pre-registered on the Open Science Framework (OSF) (August 03, 2020; retrieved from <https://doi.org/10.17605/OSF.IO/NH24G>). Deviations from the pre-registered protocol are made explicit in brief in the methods section and reasons are specified in [Supplementary file 2](#) as recommended by [Nosek et al., 2018](#), who note that such deviations are common and occur even in the most predictable analysis plans.

Participants

Participants were selected from a large cohort providing participants for subsequent studies as part of the Collaborative Research Center CRC 58. Participants from this sample were recruited for this study through a phone interview. Only healthy individuals between 18 and 50 years of age without a history of childhood trauma according to the Childhood Trauma Questionnaire (CTQ, critical cutoffs as identified by [Bernstein et al., 2003](#); [Häuser et al., 2011](#)). Additional exclusion criteria were claustrophobia, cardiac pacemaker, non-MR-compatible metal implants, brain surgery, left handedness, participation in pharmacological studies within the past 2 weeks, medication except for oral contraceptives, internal medical disorders, chronic pain, neurological disorders, psychiatric disorders, metabolic disorders, acute infections, complications with anesthesia in the past and pregnancy. Participants were right handed and had normal or corrected to normal vision. All participants gave written informed consent to the protocol which was approved by the local ethics committee (PV 5157, Ethics Committee of the General Medical Council Hamburg). The study was conducted in accordance with the Declaration of Helsinki. All participants were naive to the experimental setup and received a financial compensation of 170€ for completion of experiments at both time points (T0 and T1).

The total sample consisted of 120 participants (female_N = 79, male_N = 41, age_M = 24.46, age_{SD} = 3.73, age_{range} = 18–34). At T0 on days 1 and 2, in total 13 participants were excluded due to technical issues (day 1: N = 0; day 2: N = 3), deviating protocols (day 1: N = 2; day 2: N = 0) and SCR non-responding (day 1: N = 3; day 2: N = 5, see below for definition of 'non-responding'). Accordingly, the final dataset for the cross-sectional analysis of T0 data consists of 107 subjects (female_N = 70, male_N = 37, age_M = 24.30, age_{SD} = 3.68, age_{range} = 18–34). 84.11% of these participants were aware and 6.54% were unaware of CS–US contingencies. The remaining 9.35% subjects uncertain of the CS–US contingencies were classified as semi-aware. CS–US contingency awareness of participants was assessed with a standardized post-experimental awareness interview (adapted from [Bechara et al., 1995](#)). On

average, the US aversiveness was rated on day 1 with a value of 19.82 (SD = 3.28) and on day 2 with a value of 16.46 (SD = 4.75) on a visual analog scale (VAS) ranging from 0 to 25. The US intensity was 8.04 mA (SD = 8.28) on average. Averaged STAI-S (Strait-Trait Anxiety Inventory – State; *Spielberger, 1983*) scores were 35.38 (SD = 5.26) on day 1 and 35.57 (SD = 6.69) on day 2.

At T1, 16 subjects were excluded due to technical issues (day 1: $N = 1$; day 2: $N = 1$), deviating protocols (day 1: $N = 3$; day 2: $N = 0$) and SCR non-responding (day 1: $N = 5$; day 2: $N = 6$; see below for definition of ‘non-responding’). Additionally, 20 participants dropped out between T0 and T1 leaving 71 subjects for longitudinal analyses (female_N = 41, male_N = 30, age_M = 24.63, age_{SD} = 3.77, age_{range} = 18–32). 88.73% of the participants were aware and 1.41% were unaware of CS–US contingencies. The remaining 9.86% were classified as semi-aware. US aversiveness was rated with $M = 19.96$ (SD = 2.99) on day 1 and with $M = 17.73$ (SD = 3.90) on day 2 (VAS = 0–25). On average, the US intensity amounted to 9.76 mA (SD = 13.18). Averaged STAI-S scores were 36.33 (SD = 6.09) on day 1 and 35.83 (SD = 7.10) on day 2.

Experimental design

Here, we reanalyzed pre-existing data that are part of a larger longitudinal study that spanned six time points. In the current study, we included data from a 2-day fear conditioning experiment which were collected at two time points (T0 and T1) 6 months apart. The 2-day experimental procedure and the stimuli were identical at both time points. Measures acquired during the full longitudinal study that are not relevant for the current work such as questionnaires, hair, and salivary cortisol are not described in detail here. For an illustration of the experimental design, see also *Figure 7*.

Experimental protocol and stimuli

The protocol consisted of a habituation and a fear acquisition training phase on day 1 and an extinction training, reinstatement, and reinstatement-test phase on day 2. Acquisition and extinction training included 28 trials each (14 CS+/14 CS–), habituation and the reinstatement-test phase 14 trials each (7 CS+/7 CS–). Acquisition training was designed as delay conditioning with the US being presented 0.2 s before CS+ offset with 100% reinforcement rate (i.e., all CS+ presentations followed by the US). CSs were two light gray fractals (RGB [230, 230, 230]), 492*492 pixels presented in a pseudo-randomized order, with no more than two identical stimuli in a row, for 6–8 s (mean: 7 s). During the intertrial interval (ITI), a white fixation cross was shown for 10–16 s (mean: 13 s). Reinstatement consisted of three trials with a duration of 5 s each presented after a 10 s ITI. Reinstatement USs were delivered 4.8 s after each trial onset. The reinstatement phase was followed by a 13 s ITI before the next CS was presented during reinstatement-test. All stimuli were presented on a gray background (RGB [100, 100, 100]) using *Presentation software, 2010* (Version 14.8, Neurobehavioral Systems, Inc, Albany, CA USA) keeping the context constant to avoid renewal effects (*Haaker et al., 2014*). Visual stimuli were identical for all participants, but allocation to CS+/CS– and CS type of the first trial of each phase were counterbalanced across participants.

The electrotactile US consisted of a train of three 2 ms electrotactile rectangular pulses with an interpulse interval of 50 ms generated by a Digitimer DS7A constant current stimulator (Welwyn Garden City, Hertfordshire, UK) and was administered to the back of the right hand of the participants through a 1-cm diameter platinum pin surface electrode. The electrode was attached between the metacarpal bones of the index and middle finger. The US was individually calibrated in a standardized stepwise procedure controlled by the experimenter aiming at an unpleasant, but still tolerable level rated by the participants between 7 and 8 on scale from zero (=stimulus was not unpleasant at all) to 10 (=stimulus was the worst one could imagine within the study context). Participants were, however, not informed that we aimed at a score of 7–8.

Outcome measures

Skin conductance responses

SCRs were acquired continuously during each phase of conditioning using a BIOPAC MP 100 amplifier (BIOPAC Systems, Inc, Goleta, CA, USA) and Spike 2 software (Cambridge Electronic Design, Cambridge, UK). For analog to digital conversion, a CED2502-SA was used. Two self-adhesive hydrogel Ag/AgCl-sensor recording SCR electrodes (diameter = 55 mm) were attached on the palm of the left hand on the distal and proximal hypothenar. A 10 Hz lowpass filter and a gain of 5 Ω were

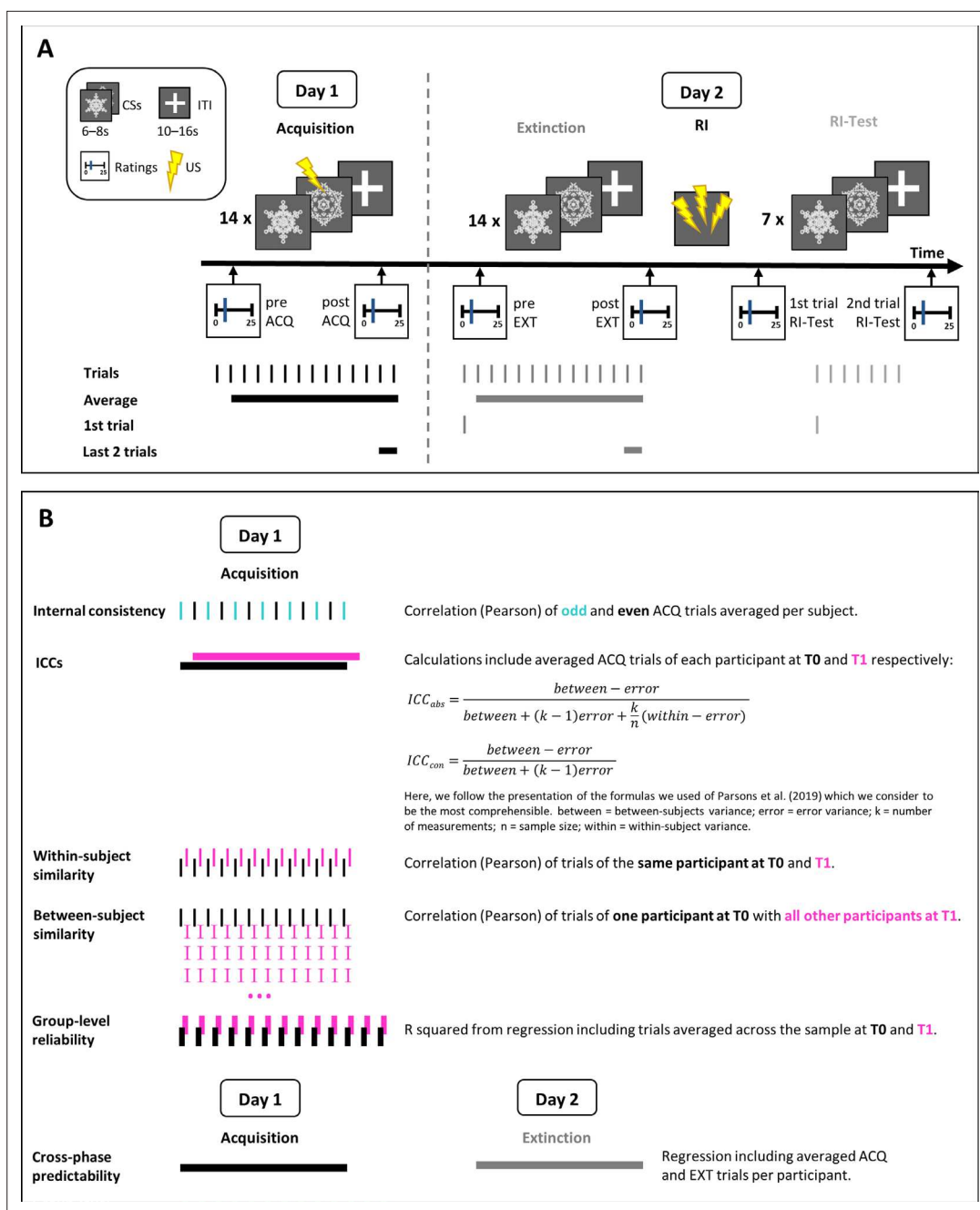


Figure 7. Illustration of the experimental design (A) and of the calculations of different measures for skin conductance responses (SCRs) including (averaged) acquisition trials (B). Note that the habituation phase is not shown in the figure, but described in the text.

applied. Data were recorded at 1000 Hz and later downsampled to 10 Hz. Subsequently, SCRs were scored semi-manually using the custom-made computer program EDA View (developed by Prof. Dr. Matthias Gamer, University of Würzburg). The program is used to quantify the SCR amplitude based on the trough-to-peak method with the trough occurring at 0.9–3.5 s after CS onset and 0.9–2.5 s after US onset (Boucsein et al., 2012; Sjouwerman and Lonsdorf, 2019). The maximum rise time was set to maximally 5 s (Boucsein et al., 2012) unless the US occurred earlier. SCRs confounded by recording artifacts due to technical reasons, such as electrode detachment or responses moving beyond the sampling window, were discarded and scored as missing values. SCRs smaller than 0.01

μ S within the defined time window were defined as zero responses. Participants with zero responses to the US in more than two-thirds (i.e., more than 9 out of 14) of US acquisition trials were classified as non-responders on day 1. On day 2, non-responding was defined as no response to any of the three reinstatement USs.

SCR data were prepared for response quantification by using **MATLAB, 2016** version R2016b. No learning could have possibly taken place during the first CS presentations as the US occurred only after the CS presentation. Consequently, the first CS+ and CS- trials during acquisition training were excluded from analyses. Hence, a total of 26 trials (13 differential SCRs) for the acquisition training phase were included in the analyses. For US analyses, all 14 trials were entered into the analyses.

Similarly, responses to the first CS+ and CS- during extinction training have to be considered a 24 hrs delayed test of fear recall as no extinction learning could have taken place. Hence, the first trial and the remaining trials of the extinction were analyzed separately. CS discrimination was computed by subtracting (averaged) CS- responses from (averaged) CS+ responses.

Fear ratings

Fear ratings to the CSs were collected prior to and after acquisition and extinction training as well as after the reinstatement-test. Participants were asked 'how much stress, fear and tension' they experienced when they last saw the CS+ and CS-. After reinstatement-test, ratings referred to (1) the first CS presentation per CS type directly after reinstatement as well as (2) the last CS presentation during reinstatement-test. After acquisition training and the reinstatement-test, subjects were also asked how uncomfortable they experienced the US itself. All ratings were given on a VAS ranging from zero (answer = none) to 100 (answer = maximum). For analyses, the rating scale was reduced to 0–25. Participants had to confirm the ratings via button press. A lack of confirmation resulted in exclusion of the trial from analyses. CS discrimination was computed by subtracting CS- from CS+ ratings.

BOLD fMRI: data acquisition, preprocessing, and first-level analysis

The inclusion of BOLD fMRI data was not initially planned and is included here as an additional non-pre-registered outcome measure.

Data acquisition

Functional data were acquired with a 3 Tesla PRISMA whole body scanner (Siemens Medical Solutions, Erlangen, Germany) using a 64-channel head coil and an echo planar imaging sequence (repetition time: 1980 ms, echo time: 30 ms, number of slices: 54, slice thickness: 1.7 mm [1 mm gap], field of view = 132 × 132 mm). T1-weighted structural images were acquired using a magnetization prepared rapid gradient echo (MPRAGE) sequence (TR: 2300 ms, TE: 2.98 ms, number of slices: 240, slice thickness: 1 mm, field of view = 192 × 256 mm).

Preprocessing

fMRI data analysis was performed using SPM12 (Wellcome Department of Neuroimaging, London, UK) and **MATLAB, 2019**. Preprocessing included realignment, coregistration, normalization to a group-specific DARTEL template and smoothing (6 mm full width at half maximum, FWHM).

First-level analysis

Regressors for the first-level analysis of acquisition training data included separate regressors for the first CS+ and CS- trials and the remaining CS+ and CS- trials because no learning could have occurred at the first presentation of the CSs. Nuisance regressors included habituation trials, US presentation, fear ratings and motion parameters. Likewise, separate regressors for the first CS+ and CS- trials of extinction (because no extinction has taken place yet) as well as the remaining CS+ and CS- trials were included as regressors of interest in the first-level analysis of extinction data acquired on day 2, while US, rating onset and motion parameters were included as regressors of no interest. No second-level analysis was completed in the current study, instead different analyses were carried out based on first-level models as further detailed in the statistical analysis section.

Regions of interest

A total of 11 ROIs (i.e., bilateral anterior insula, amygdala, hippocampus, caudate nucleus, putamen, pallidum, NAcc, thalamus, dACC, dlPFC, and vmPFC) were included in the current study. Amygdala, hippocampus, caudate nucleus, putamen, pallidum, ventral striatum (i.e., NAcc), and thalamus anatomical masks were extracted from the Harvard-Oxford atlas (*Desikan et al., 2006*) at a maximum probability threshold of 0.5. The anterior insula was defined as the overlap between the thresholded anatomical mask from the Harvard-Oxford atlas (threshold: 0.5) and a box of size 60 × 30 × 60 mm centered around MNIxyz = 0, 30, 0 based on anatomical subdivisions (*Nieuwenhuys, 2012*). The cortical ROI dlPFC and dACC were created by building a box of size 20 × 16 × 16 mm around peak voxels obtained in a meta-analysis (with the x coordinate set to 0 for the dACC) (left dlPFC: MNIxyz = -36, 44, 22, right dlPFC: MNIxyz = 34, 44, 32, dACC: MNIxyz = 0, 18, 42, *Fullana et al., 2016*). As previously reported (*Lonsdorf et al., 2014*), the cortical ROI vmPFC was created by using a box of size 20 × 16 × 16 mm centered on peak coordinates identified in prior studies of fear learning (vmPFC: MNIxyz = 0, 40, -12, e.g., *Kalisch et al., 2006, Milad et al., 2007*) with the x coordinate set to 0 to obtain masks symmetric around the midline.

All analyses of BOLD fMRI as described below were conducted separately not only for the whole brain but also for these 11 selected ROIs.

Statistical analyses

For a comprehensive overview of which analysis was carried out for which outcome measures, stimuli, phases and data transformations (see *Table 1*). For an illustration of which data were included in the different analyses, see also *Figure 7B*.

Internal consistency

We assessed the internal consistency of SCRs for both time points and experimental phases separately (for details, see *Table 1*): trials of the respective time point and phase were split into odd and even trials (i.e., odd–even approach) and averaged for each individual subject. Averaged odd and even trials were then correlated by using Pearson’s correlation coefficient. To obtain a rather conservative result, we refrained from applying the Spearman–Brown prophecy formula. We considered the odd–even approach as the most appropriate since our paradigm constitutes a learning experiment and we suggest that adjacent trials measure a more similar construct compared to other possible splits of trials such as a split into halves or a large number of random splits as implemented in the permutation-based approach recommended by *Parsons et al., 2019*. Calculations of internal consistency were not possible for fear ratings and BOLD fMRI due to the limited number of data points for fear ratings and an experimental design that did not allow for a trial-by-trial analysis of BOLD fMRI data. Internal consistency was interpreted using benchmarks for unacceptable (<0.5), poor (>0.5 but <0.6), questionable (>0.6 but <0.7), acceptable (>0.7 but <0.8), good (>0.8 but <0.9), and excellent (≥0.9) (*Kline, 2013*).

Longitudinal reliability at the individual and group level

While internal consistency indicates the extent to which all items of a test or – here, trials of an experimental phase – measure the same construct (*Revelle, 1979*), longitudinal reliability reflects the variability across two or more measurements of the same individual under the same conditions and is therefore indicative of the degree of correlation and agreement between measurements (*Koo and Li, 2016*). For calculations of longitudinal reliability, we included data from both time points T0 and T1 from the same experimental phase. To capture different aspects of longitudinal reliability, we chose a dual approach of calculating longitudinal reliability at both (1) the individual level and (2) at the group level (for details see also *Table 1*). To this end, longitudinal reliability at the individual and group level indicates to which extent responses within the same individual and within the group as a whole are stable over time. More precisely, whereas longitudinal reliability at the individual level takes into account the individual responses of participants, which are then related across time points, reliability at the group level first averages the individual responses across the group and then relates them across time points. Reliability at the individual level inherently includes the group level, as it is calculated for the sample as whole, but the individual responses are central to the calculation. Contrarily, for reliability at the group level, the calculation is carried out using group averages.

Reliability at the individual level was investigated as (1) ICCs encompassing both time points, (2) within- and between-subject similarity of individual trial-by-trial responding (i.e., SCRs) or BOLD fMRI activation patterns between time points, and (3) as the degree of overlap of significant voxels between time points within an individual (for methodological details see below). Reliability at the group level was investigated as (1) trial-by-trial group average SCRs and (2) the degree of overlap of significant voxels between time points within the group as a whole (for methodological details see below).

Assessments of internal consistency, within- and between-subject similarity, overlap at the individual and group level as well as longitudinal reliability of SCRs at the group level were not pre-registered but are included as they provide valuable additional and complementary information. Overlap and similarity analyses follow the methodological approach of *Fröhner et al., 2019*.

Longitudinal reliability at the individual level

Intraclass correlation coefficients

ICCs were determined separately for each experimental phase by including data from both time points T0 and T1. Generally, larger ICCs indicate higher congruency of within-subject responding between time points and increased distinction of subjects from each other (*Noble et al., 2021*). *Parsons et al., 2019* recommend the calculations of ICCs in cognitive-behavioral tasks through a two-way mixed-effects model of single rater type labeled ICC(2,1) (absolute agreement, in the following referred to as ICC_{abs}) and ICC(3,1) (consistency, in the following referred to as ICC_{con}) according to *Shrout and Fleiss, 1979* convention and to report their 95% CIs. Due to their slightly different calculations, ICC_{abs} tends to be lower than ICC_{con} (see *Table 1*).

However, as the pre-registered mixed-effects approach resulted in non-convergence of some models for SCRs and ratings, we implemented an analysis of variance (ANOVA) instead of the mixed-effects approach to calculate ICC_{abs} and ICC_{con} (*Shrout and Fleiss, 1979*). To calculate ICCs for BOLD fMRI (additional not pre-registered analyses), the SPM-based toolbox fmrel (Fröhner et al., 2019) was used. BOLD fMRI ICCs were determined for each voxel and averaged across the whole brain and for selected ROIs.

Furthermore, we investigated whether or to what extent ICCs change when ICC calculations were based on different numbers of trials. To this end, we included (additional non-pre-registered) analyses of trial-by-trial ICCs for SCRs in the supplementary material: First, ICCs were only computed for the first trial. Then, all subsequent trials of the respective phase were added stepwise to this first trial. After each step, trials were averaged and ICCs were calculated (see *Figure 1—figure supplements 3–8*).

Within the figures, values less than 0.5 are classified as poor reliability, values between 0.5 and 0.75 as indicative of moderate reliability, values between 0.75 and 0.9 are classified as good reliability and values greater than 0.9 as excellent reliability, as suggested by *Koo and Li, 2016*. These benchmarks are included here to provide a frame of reference but we point out that these benchmarks are arbitrary and should hence not be overinterpreted in particular in the context of responding in experimental paradigms as these benchmarks have been developed in different contexts (i.e., trait self-report measures).

Within- and between-subject similarity

Both ICCs and within-subject similarity indicate to which extent responses of an individual at one time point are comparable to responses of the same individual at a later time point. Both were calculated separately for each experimental phase by including data from both time points. There are, however, two main differences: First, ICCs were calculated by decomposition of variances as applied for ANOVA, whereas similarity was calculated as correlation of responses between both time points (1) within one individual (within-subject similarity) and (2) between this individual and all other individuals (between-subject similarity). Second, while ICCs are interpreted in terms of absolute values using cutoffs that provide information on the quantity of longitudinal reliability, within-subject similarity was compared to between-subject similarity showing if responses of one subject at T0 were more similar to themselves at T1 than to responses of all others at T1. The approach to the assessment of similarity was derived from the idea of representational similarity analysis (RSA) introduced by *Kriegeskorte et al., 2008* and previously used by *Fröhner et al., 2019* for the comparison of fMRI BOLD activation patterns between different sessions.

Here, within-subject similarity was calculated by correlating (Pearson's correlation coefficient) (1) individual trial-by-trial SCRs and (2) the first-level response patterns of brain activation for CS discrimination (i.e., CS+ > CS-) of each individual subject between T0 and T1 resulting in one value of within-subject similarity per subject (e.g., SCR acquisition trials of subject 1 at T0 were correlated with SCR acquisition trials of subject 1 at T1). Between-subject similarity was calculated by correlating trial-by-trial SCRs or the first-level response patterns of brain activation of each individual subject at T0 with those of all other individuals at T1 (e.g., SCR acquisition trials of subject 1 at T0 were correlated with SCR acquisition trials of subject 2–71 at T1). This resulted in 70 correlation coefficients for each subject. These correlation coefficients were then averaged to yield one correlation coefficient per subject as an indicator of between-subject similarity.

For comparisons of within- and between-subject similarity in SCR and BOLD fMRI, similarities were Fisher *r*-to-*z* transformed and compared using paired *t*-tests or Welsh tests in cases where the assumption of equal variances was not met. Cohen's *d* is reported as effect size.

Note that within-subject similarities of SCRs could not be calculated for participants with a single non-zero response at the same trial (e.g., trial 1) at both time points or only zero responses to the CS+ or CS- in one particular phase. This is because arrays that include only zeros can not be correlated and correlations of 1 (e.g., resulting from non-zero responses at the same trial at both time points) result in infinite Fisher *r*-to-*z* transformed correlations. Thus, different numbers of participants had to be included in the analyses of SCRs during acquisition ($N_{\text{CS discrimination}} = 65$, $N_{\text{CS+}} = 62$, $N_{\text{CS-}} = 56$, $N_{\text{US}} = 71$) and extinction training ($N_{\text{CS discrimination}} = 45$, $N_{\text{CS+}} = 40$, $N_{\text{CS-}} = 32$).

Overlap at the individual level

For BOLD fMRI, overlap in individual subject activation patterns across both time points was calculated as a third indicator of reliability at the individual level. Thus, overlap was determined separately for experimental phases by including data from both time points T0 and T1. To this end, activation maps from first-level contrasts (here CS+ > CS or CS discrimination) were compared such that the degree of overlap of significant voxels at a liberal threshold of $p_{\text{uncorrected}} < 0.01$ between T0 and T1 was determined and expressed as the Dice and Jaccard coefficients (Fröhner et al., 2019). Both coefficients range from 0 (no overlap) to 1 (perfect overlap), with the Jaccard index being easily interpretable as percent overlap (Fröhner et al., 2019). While overlap reflects the degree of voxels activated at both time points, similarity measures (see above) are based on the correlation of activated voxels between time points and can be considered a continuous approach based on CS+ > CS- contrast specific beta values and not thresholded T-maps.

Longitudinal reliability at the group level

As opposed to longitudinal reliability at the individual level which indicates the stability of individual responses across time points, longitudinal reliability at the group level refers to how stable group average responding is over time. Longitudinal reliability at the group level was calculated separately for experimental phases by including data from both time points T0 and T1.

We define longitudinal reliability at the group level (1) for SCRs as the percentage of explained variance of group averaged trials at T1 by group averaged trials at T0 (i.e., *R* squared) and (2) for BOLD fMRI as the degree of overlap of group averaged activated voxels between both time points. Different analysis approaches were chosen as SCR and BOLD fMRI data are inherently different measures: trial-by-trial analyses in fMRI require slow-event related designs with long ITIs as well as fixed trial orders and ideally partial reinforcement rate to not confound CS and US responses (Visser et al., 2016). Hence, trial-by-trial analyses were not possible given our design and thus overlap at a group level was defined as overlap at voxel rather than at trial level.

For SCRs, simple linear regressions were computed with group averaged SCR trials at T0 as independent and group averaged SCR trials at T1 as dependent variable and *R* squared was extracted. This was done separately for experimental phases. Although the Pearson's correlation coefficient is often calculated to determine longitudinal reliability, *R* squared, which like overlap can also be expressed as a percentage, appears closest to the concept of overlap of significant voxels at T0 and T1 as applied to BOLD fMRI data.

For overlap in BOLD fMRI at the group level, the degree of overlap of significant voxels between both time points was determined for aggregated group-level activations instead of single subject-level

activation patterns (see ‘Overlap at the individual level’) and expressed using the Dice and Jaccard indices as described above.

Cross-phases predictability of conditioned responding

Simple linear regressions were calculated to assess the predictability of SCRs and fear ratings across experimental phases at T0. During data analysis, inspection of the data revealed heteroscedasticity. Therefore and deviating from the pre-registration, regressions with robust standard errors were calculated by using the HC3 estimator (*Hayes and Cai, 2007*). Two consecutive phases represent the independent and the dependent variable, respectively, with the preceding phase as the independent variable and the following phase as the dependent variable. For SCR and fear ratings, standardized betas as derived from linear regressions are reported. In simple linear regression, as implemented here, standardized betas can be also interpreted as Pearson’s correlation coefficients.

For fMRI data, we adopted the cross-phases predictability analysis of SCR and fear ratings by calculating Pearson’s correlation coefficients between patterns of voxel activation (i.e., first-level beta maps). Correlations were first calculated at the individual subject level and subsequently averaged.

Standardized betas (resulting from SCR and fear rating regressions) and correlation coefficients (resulting from BOLD fMRI correlational tests) were interpreted as demonstrating weak, moderate, or strong associations between variables with values of <0.4 , ≥ 0.4 , and ≥ 0.7 , respectively (*Dancey and Reidy, 2007*). Tables containing regression parameters beyond the standardized betas depicted in **Figure 5A, B** are presented in the Supplement (see **Supplementary file 7, Supplementary file 8**).

For SCR and fear rating predictions, we assessed if predictions differ in their strength or direction when they are summarized across certain data specifications (see **Table 1**). For BOLD fMRI, correlation coefficients were pooled across ROIs. *T*-tests or Welch tests in cases where the assumption of equal variances was not met were performed on individual Fisher *r*-to-*z* transformed standardized betas (SCR and fear ratings) or correlation coefficients (BOLD fMRI). We highlight that these analyses can be interpreted as an example for predictive validity (i.e., the extent to which a score on a test predicts a score on a criterion measure). As our aim here is, however, not validation, we use the term cross-phase prediction throughout. (More precisely, we believe that ‘cross-phase predictions’ in our study cannot be used interchangeably with ‘criterion or predictive validity’ since our aim was not to validate one experimental phase against the other. Predictive validity in psychometrics is defined as ‘the extent to which a score on a scale (or test) predicts scores on some criterion measure’ (cf. *Cronbach and Meehl, 1955*). For instance, a cognitive test for job performance would have predictive validity if the observed correlation between the test score and the performance rating by the company were statistically significant. Rather, we investigated whether responses in earlier experimental phases could predict responses in later experimental phases – both of which cannot be expected to ‘measure the same thing’.)

For all statistical analyses described above, a level of $p < 0.05$ (two-sided) was considered significant. Since we were more interested in patterns of results and less in the result of one specific test, it was not necessary to correct for multiple comparisons. Moreover, multiverse approaches, as approximated in our study, are assumed to be insensitive to multiple comparisons (*Lonsdorf et al., 2022*).

For data analyses and visualizations as well as for the creation of the manuscript, we used R (Version 4.1.3; *R Development Core Team, 2020*) and the R-packages *apa* (*Aust and Barth, 2020*; Version 0.3.3; *Gromer, 2020*), *car* (Version 3.0.10; *Fox and Weisberg, 2019*; *Fox et al., 2020*), *carData* (Version 3.0.4; *Fox et al., 2020*), *cowplot* (Version 1.1.1; *Wilke, 2020*), *DescTools* (Version 0.99.42; *Andri mult, 2021*), *dplyr* (Version 1.0.8; *Wickham et al., 2021*), *effsize* (*Torchiano, 2020*), *flectable* (Version 0.6.10; *Gohel, 2021*), *gghalves* (Version 0.1.1; *Tiedemann, 2020*), *ggplot2* (Version 3.3.5; *Wickham, 2016*), *ggpubr* (Version 0.4.0; *Kassambara, 2020*), *ggsignif* (Version 0.6.3; *Constantin and Patil, 2021*), *gridExtra* (Version 2.3; *Auguie, 2017*), *here* (Version 1.0.1; *Müller, 2020*), *kableExtra* (Version 1.3.1; *Zhu, 2020*), *knitr* (Version 1.37; *Xie, 2015*), *lm.beta* (Version 1.5.1; *Behrendt, 2014*), *lmtest* (Version 0.9.38; *Zeileis and Hothorn, 2002*), *officedown* (Version 0.2.4; *Gohel and Ross, 2022*), *papaja* (Version 0.1.0.9997; *Aust and Barth, 2020*), *patchwork* (Version 1.1.0; *Pedersen, 2020*), *psych* (Version 2.0.9; *Revelle, 2020*), *renv* (Version 0.13.2; *Ushey, 2020*), *reshape2* (Version 1.4.4; *Wickham, 2007*), *sandwich* (*Zeileis, 2004*; *Zeileis, 2006*; Version 3.0.1; *Zeileis et al., 2020*), *stringr* (Version 1.4.0; *Wickham, 2019*), *tidyr* (Version 1.2.0; *Wickham, 2020*), *tinylab* (Version 0.2.3; *Barth, 2022*), and *zoo* (Version 1.8.8; *Zeileis and Grothendieck, 2005*).

Acknowledgements

The authors would like to thank Claudia Immisch, Janne Nold, Kevin Rozario, and Habiba Schiller for help with data collection and Karoline Rosenkranz for help with data preprocessing, Mario Reutter for methodological discussions and comments on an earlier draft as well as Juliane Tkotz for support with reproducible manuscript writing.

Additional information

Funding

Funder	Grant reference number	Author
Deutsche Forschungsgemeinschaft	INST 211/633-2	Tina B Lonsdorf
Deutsche Forschungsgemeinschaft	LO 1980/4-1	Tina B Lonsdorf
Deutsche Forschungsgemeinschaft	LO 1980/7-1	Tina B Lonsdorf

The funders had no role in study design, data collection, and interpretation, or the decision to submit the work for publication.

Author contributions

Maren Klingelhöfer-Jens, Conceptualization, Data curation, Software, Formal analysis, Visualization, Methodology, Writing - original draft, Pre-registration of the study; Mana R Ehlers, Conceptualization, Formal analysis, Visualization, Methodology, Writing - original draft; Manuel Kuhn, Data curation, Software, Investigation, Writing – review and editing; Vincent Keyaniyan, Formal analysis, Visualization, Methodology, Writing – review and editing, Pre-registration of the study; Tina B Lonsdorf, Conceptualization, Resources, Supervision, Funding acquisition, Methodology, Writing - original draft, Pre-registration of the study


Author ORCIDs

Maren Klingelhöfer-Jens  <http://orcid.org/0000-0002-5393-7871>

Mana R Ehlers  <http://orcid.org/0000-0002-1316-3787>

Manuel Kuhn  <http://orcid.org/0000-0003-2210-9130>

Vincent Keyaniyan  <http://orcid.org/0000-0002-5674-5197>

Tina B Lonsdorf  <http://orcid.org/0000-0003-1501-4846>

Ethics

All participants gave written informed consent to the protocol which was approved by the local ethics committee (PV 5157, Ethics Committee of the General Medical Council Hamburg). The study was conducted in accordance with the Declaration of Helsinki.

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.78717.sa1>

Author response <https://doi.org/10.7554/eLife.78717.sa2>

Additional files

Supplementary files

- Supplementary file 1. Overview of experimental specifications and results of five previous studies reporting test–retest reliabilities in human fear conditioning research.
- Supplementary file 2. Deviations from pre-registration.
- Supplementary file 3. ICC_{abs} and ICC_{con} for all data specifications of SCRs.
- Supplementary file 4. ICC_{abs} and ICC_{con} for all data specifications of fear ratings.
- Supplementary file 5. ICC_{abs} and ICC_{con} for CS discrimination during fear acquisition (Acq) and extinction training (Ext).

- Supplementary file 6. Paired sample *t*-tests comparing between- and within-subject similarity for whole brain activation pattern as well as activation pattern in the ROIs for acquisition training (Acq) and extinction training (Ext).
- Supplementary file 7. Detailed results of linear regressions: SCR.
- Supplementary file 8. Detailed results of linear regressions: fear ratings.
- Transparent reporting form
- MDAR checklist

Data availability

The data that support the findings of this study and the R Markdown files that generate this manuscript are openly available in Zenodo at <https://doi.org/10.5281/zenodo.7323547>.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Klingelhöfer-Jens M, Ehlers MR, Kuhn M, Keyaniyan V, Lonsdorf TB	2022	Robust group- but limited individual-level (longitudinal) reliability and insights into cross-phases response prediction of conditioned fear	https://doi.org/10.5281/zenodo.7323547	Zenodo, 10.5281/zenodo.7323547

References

- Aldridge VK**, Dovey TM, Wade A. 2017. Assessing test-retest reliability of psychological measures. *European Psychologist* **22**:207–218. DOI: <https://doi.org/10.1027/1016-9040/a000298>
- Anderson KC**, Insel TR. 2006. The promise of extinction research for the prevention and treatment of anxiety disorders. *Biological Psychiatry* **60**:319–321. DOI: <https://doi.org/10.1016/j.biopsych.2006.06.022>, PMID: [16919521](https://pubmed.ncbi.nlm.nih.gov/16919521/)
- Andri mult S**. 2021. DescTools: tools for descriptive statistics. R-Project. <https://cran.r-project.org/package=DescTools>
- Augue B**. 2017. GridExtra: miscellaneous functions for “ grid ” graphics. R-Project. <https://CRAN.R-project.org/package=gridExtra>
- Aust F**, Barth M. 2020. Papaja: create APA manuscripts with R markdown. 4.0.3. Github. <https://github.com/crsh/papaja>
- Bach DR**, Melinščak F, Fleming SM, Voelkle MC. 2020. Calibrating the experimental measurement of psychological attributes. *Nature Human Behaviour* **4**:1229–1235. DOI: <https://doi.org/10.1038/s41562-020-00976-8>, PMID: [33199857](https://pubmed.ncbi.nlm.nih.gov/33199857/)
- Baker DH**, Vilidaite G, Lygo FA, Smith AK, Flack TR, Gouws AD, Andrews TJ. 2021. Power contours: optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods* **26**:295–314. DOI: <https://doi.org/10.1037/met0000337>, PMID: [32673043](https://pubmed.ncbi.nlm.nih.gov/32673043/)
- Barth M**. 2022. Tynylab: lightweight variable labels. R-Project. <https://cran.r-project.org/package=tynylab>
- Bechara A**, Tranel D, Damasio H, Adolphs R, Rockland C, Damasio AR. 1995. Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science* **269**:1115–1118. DOI: <https://doi.org/10.1126/science.7652558>, PMID: [7652558](https://pubmed.ncbi.nlm.nih.gov/7652558/)
- Behrendt S**. 2014. lm.beta: add standardized regression coefficients to lm-objects. R-Project. <https://CRAN.R-project.org/package=lm.beta>
- Bernstein DP**, Stein JA, Newcomb MD, Walker E, Pogge D, Ahluvalia T, Stokes J, Handelsman L, Medrano M, Desmond D, Zule W. 2003. Development and validation of a brief screening version of the childhood trauma questionnaire. *Child Abuse & Neglect* **27**:169–190. DOI: [https://doi.org/10.1016/s0145-2134\(02\)00541-0](https://doi.org/10.1016/s0145-2134(02)00541-0), PMID: [12615092](https://pubmed.ncbi.nlm.nih.gov/12615092/)
- Boucsein W**, Fowles DC, Grimnes S, Ben-Shakhar G, Roth WT, Dawson ME, Filion DL, Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. 2012. Publication recommendations for electrodermal measurements. *Psychophysiology* **49**:1017–1034. DOI: <https://doi.org/10.1111/j.1469-8986.2012.01384.x>, PMID: [22680988](https://pubmed.ncbi.nlm.nih.gov/22680988/)
- Bouton ME**. 2004. Context and behavioral processes in extinction: table 1. *Learning & Memory* **11**:485–494. DOI: <https://doi.org/10.1101/lm.78804>, PMID: [15466298](https://pubmed.ncbi.nlm.nih.gov/15466298/)
- Bouton ME**, García-Gutiérrez A, Ziiski J, Moody EW. 2006. Extinction in multiple contexts does not necessarily make extinction less vulnerable to relapse. *Behaviour Research and Therapy* **44**:983–994. DOI: <https://doi.org/10.1016/j.brat.2005.07.007>, PMID: [16198302](https://pubmed.ncbi.nlm.nih.gov/16198302/)
- Constantin AE**, Patil I. 2021. Ggsignif: R Package for Displaying Significance Brackets for “Ggplot2.” *PsyArXiv*. DOI: <https://doi.org/10.31234/osf.io/7awm6>
- Cooper SE**, van Dis EAM, Hagenaaars MA, Kryptos AM, Nemeroff CB, Lissek S, Engelhard IM, Dunsmoor JE. 2022a. A meta-analysis of conditioned fear generalization in anxiety-related disorders.

- Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology* **47**:1652–1661. DOI: <https://doi.org/10.1038/s41386-022-01332-2>, PMID: 35501429
- Cooper SE**, Dunsmoor JE, Koval K, Pino E, Steinman S. 2022b. Test-Retest Reliability of Human Threat Conditioning and Generalization. [PsyArXiv]. DOI: <https://doi.org/10.31234/osf.io/84uqz>
- Craske MG**, Kircanski K, Zelikowsky M, Mystkowski J, Chowdhury N, Baker A. 2008. Optimizing inhibitory learning during exposure therapy. *Behaviour Research and Therapy* **46**:5–27. DOI: <https://doi.org/10.1016/j.brat.2007.10.003>, PMID: 18005936
- Cronbach LJ**, Meehl PE. 1955. Construct validity in psychological tests. *Psychological Bulletin* **52**:281–302. DOI: <https://doi.org/10.1037/h0040957>, PMID: 13245896
- Dancey CP**, Reidy J. 2007. *Statistics without Maths for Psychology: Using SPSS for Windows*. Harlow, England ; New York: Pearson/Prentice Hall.
- Desikan RS**, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**:968–980. DOI: <https://doi.org/10.1016/j.neuroimage.2006.01.021>, PMID: 16530430
- DeYoung CG**, Sassenberg T, Abend R, Allen T, Beaty R, Bellgrove M, Blain SD, Bzdok D, Chavez R, Engel SA, Ma F, Fornito A, Genç E, Goghari V, Grazioplene R, Hanson JL, Haxby JV, Hilger K, Homan P, Joyner K, et al. 2022. Reproducible Between-Person Brain-Behavior Associations Do Not Always Require Thousands of Individuals. *PsyArXiv*. DOI: <https://doi.org/10.31234/osf.io/sfnmk>
- Duits P**, Cath DC, Lissek S, Hox JJ, Hamm AO, Engelhard IM, van den Hout MA, Baas JMP. 2015. Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and Anxiety* **32**:239–253. DOI: <https://doi.org/10.1002/da.22353>, PMID: 25703487
- Dunsmoor JE**, Cisler JM, Fonzo GA, Creech SK, Nemeroff CB. 2022. Laboratory models of post-traumatic stress disorder: the elusive bridge to translation. *Neuron* **110**:1754–1776. DOI: <https://doi.org/10.1016/j.neuron.2022.03.001>, PMID: 35325617
- Ehlers MR**, Nold J, Kuhn M, Klingelhöfer-Jens M, Lonsdorf TB. 2020. Revisiting potential associations between brain morphology, fear acquisition and extinction through new data and a literature review. *Scientific Reports* **10**:19894. DOI: <https://doi.org/10.1038/s41598-020-76683-1>, PMID: 33199738
- Elliott ML**, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, Sison ML, Moffitt TE, Caspi A, Hariri AR. 2020. What is the test-retest reliability of common task-functional MRI measures? new empirical evidence and a meta-analysis. *Psychological Science* **31**:792–806. DOI: <https://doi.org/10.1177/0956797620916786>, PMID: 32489141
- Essau CA**, Lewinsohn PM, Lim JX, Ho MR, Rohde P. 2018. Incidence, recurrence and comorbidity of anxiety disorders in four major developmental stages. *Journal of Affective Disorders* **228**:248–253. DOI: <https://doi.org/10.1016/j.jad.2017.12.014>, PMID: 29304469
- Feilong M**, Guntupalli JS, Haxby JV. 2021. The neural basis of intelligence in fine-grained cortical topographies. *eLife* **10**:e64058. DOI: <https://doi.org/10.7554/eLife.64058>, PMID: 33683205
- Fisher AJ**, Medaglia JD, Jeronimus BF. 2018. Lack of group-to-individual generalizability is a threat to human subjects research. *PNAS* **115**:E6106–E6115. DOI: <https://doi.org/10.1073/pnas.1711978115>, PMID: 29915059
- Foa EB**, Grayson JB, Steketee GS, Doppelt HG, Turner RM, Latimer PR. 1983. Success and failure in the behavioral treatment of obsessive-compulsives. *Journal of Consulting and Clinical Psychology* **51**:287–297. DOI: <https://doi.org/10.1037//0022-006x.51.2.287>, PMID: 6841773
- Fox J**, Weisberg S. 2019. An R companion to applied regression (Third). <https://socialsciences.mcmaster.ca/jfox/Books/Companion/> [Accessed May 17, 2020].
- Fox J**, Weisberg S, Price B. 2020. CarData: companion to applied regression data sets. R-Project. <https://CRAN.R-project.org/package=carData>
- Fredrikson M**, Annas P, Georgiades A, Hursti T, Tersman Z. 1993. Internal consistency and temporal stability of classically conditioned skin conductance responses. *Biological Psychology* **35**:153–163. DOI: [https://doi.org/10.1016/0301-0511\(93\)90011-v](https://doi.org/10.1016/0301-0511(93)90011-v), PMID: 8507744
- Fröhner JH**, Teckentrup V, Smolka MN, Kroemer NB. 2019. Addressing the reliability fallacy in fMRI: similar group effects may arise from unreliable individual effects. *NeuroImage* **195**:174–189. DOI: <https://doi.org/10.1016/j.neuroimage.2019.03.053>, PMID: 30930312
- Fullana MA**, Harrison BJ, Soriano-Mas C, Vervliet B, Cardoner N, Àvila-Parcet A, Radua J. 2016. Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry* **21**:500–508. DOI: <https://doi.org/10.1038/mp.2015.88>, PMID: 26122585
- Fullana MA**, Dunsmoor JE, Schruers KRJ, Savage HS, Bach DR, Harrison BJ. 2020. Human fear conditioning: from neuroscience to the clinic. *Behaviour Research and Therapy* **124**:103528. DOI: <https://doi.org/10.1016/j.brat.2019.103528>
- Galatzer-Levy IR**, Bonanno GA, Bush DEA, LeDoux JE. 2013a. Heterogeneity in threat extinction learning: substantive and methodological considerations for identifying individual difference in response to stress. *Frontiers in Behavioral Neuroscience* **7**:55. DOI: <https://doi.org/10.3389/fnbeh.2013.00055>
- Galatzer-Levy IR**, Bryant RA. 2013b. 636,120 ways to have posttraumatic stress disorder. *Perspectives on Psychological Science* **8**:651–662. DOI: <https://doi.org/10.1177/1745691613504115>, PMID: 26173229
- Gershman SJ**, Hartley CA. 2015. Individual differences in learning predict the return of fear. *Learning & Behavior* **43**:243–250. DOI: <https://doi.org/10.3758/s13420-015-0176-z>, PMID: 26100524
- Gohel D**. 2021. Flextable: functions for tabular reporting. 0.8.1. R-Project. <https://CRAN.R-project.org/package=flextable>

- Gohel D**, Ross N. 2022. officedown: Enhanced 'R Markdown' format for 'Word' and 'PowerPoint' 0.2.4. CRAN. <https://CRAN.R-project.org/package=officedown>
- Graham BM**, Milad MR. 2011. The study of fear extinction: implications for anxiety disorders. *American Journal of Psychiatry* **168**:1255–1265. DOI: <https://doi.org/10.1176/appi.ajp.2011.11040557>, PMID: 21865528
- Gromer D**. 2020. Apa: format outputs of statistical tests according to APA guidelines. R-Project. <https://CRAN.R-project.org/package=apa>
- Gulliksen H**. 1950. Effect of group heterogeneity on test reliability. Gulliksen H (Ed). *Theory of Mental Tests*. Hoboken, NJ: Wiley. p. 108–127.
- Haaker J**, Golkar A, Hermans D, Lonsdorf TB. 2014. A review on human reinstatement studies: an overview and methodological challenges. *Learning & Memory* **21**:424–440. DOI: <https://doi.org/10.1101/lm.036053.114>, PMID: 25128533
- Harder JA**. 2020. The multiverse of methods: extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science* **15**:1158–1177. DOI: <https://doi.org/10.1177/1745691620917678>, PMID: 32598854
- Häuser W**, Schmutzer G, Brähler E, Glaesmer H. 2011. Maltreatment in childhood and adolescence: results from a survey of a representative sample of the German population. *Deutsches Arzteblatt International* **108**:287–294. DOI: <https://doi.org/10.3238/arztebl.2011.0287>, PMID: 21629512
- Hayes AF**, Cai L. 2007. Using heteroskedasticity-consistent standard error estimators in OLS regression: an introduction and software implementation. *Behavior Research Methods* **39**:709–722. DOI: <https://doi.org/10.3758/bf03192961>, PMID: 18183883
- Hedge C**, Powell G, Sumner P. 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods* **50**:1166–1186. DOI: <https://doi.org/10.3758/s13428-017-0935-1>, PMID: 28726177
- Herting MM**, Gautam P, Chen Z, Mezher A, Vetter NC. 2018. Test-Retest reliability of longitudinal task-based fMRI: implications for developmental studies. *Developmental Cognitive Neuroscience* **33**:17–26. DOI: <https://doi.org/10.1016/j.dcn.2017.07.001>, PMID: 29158072
- Infantolino ZP**, Luking KR, Sauder CL, Curtin JJ, Hajcak G. 2018. Robust is not necessarily reliable: from within-subjects fMRI contrasts to between-subjects comparisons. *NeuroImage* **173**:146–152. DOI: <https://doi.org/10.1016/j.neuroimage.2018.02.024>, PMID: 29458188
- Kalisch R**, Korenfeld E, Stephan KE, Weiskopf N, Seymour B, Dolan RJ. 2006. Context-Dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. *The Journal of Neuroscience* **26**:9503–9511. DOI: <https://doi.org/10.1523/JNEUROSCI.2021-06.2006>, PMID: 16971534
- Kassambara A**. 2020. Ggpubr: 'ggplot2' based publication ready plots. R-Project. <https://CRAN.R-project.org/package=ggpubr>
- Kennedy JT**, Harms MP, Korucuoglu O, Astafiev SV, Barch DM, Thompson WK, Bjork JM, Anokhin AP. 2021. Reliability and Stability Challenges in ABCD Task FMRI Data. *bioRxiv*. DOI: <https://doi.org/10.1101/2021.10.08.463750>
- Kline P**. 2013. *Handbook of Psychological Testing*. Routledge. DOI: <https://doi.org/10.4324/9781315812274>
- Kong R**, Yang Q, Gordon E, Xue A, Yan X, Orban C, Zuo XN, Spreng N, Ge T, Holmes A, Eickhoff S, Yeo BTT. 2021. Individual-Specific areal-level parcellations improve functional connectivity prediction of behavior. *Cerebral Cortex* **31**:4477–4500. DOI: <https://doi.org/10.1093/cercor/bhab101>, PMID: 33942058
- Koo TK**, Li MY. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* **15**:155–163. DOI: <https://doi.org/10.1016/j.jcm.2016.02.012>, PMID: 27330520
- Kozak MJ**, Foa EB, Steketee G. 1988. Process and outcome of exposure treatment with obsessive-compulsives: psychophysiological indicators of emotional processing. *Behavior Therapy* **19**:157–169. DOI: [https://doi.org/10.1016/S0005-7894\(88\)80039-X](https://doi.org/10.1016/S0005-7894(88)80039-X)
- Kragel PA**, Han X, Kraynak TE, Gianaros PJ, Wager TD. 2021. Functional MRI can be highly reliable, but it depends on what you measure: a commentary on Elliott et al. (2020). *Psychological Science* **32**:622–626. DOI: <https://doi.org/10.1177/0956797621989730>, PMID: 33685310
- Kriegeskorte N**, Mur M, Bandettini P. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* **2**:4. DOI: <https://doi.org/10.3389/neuro.06.004.2008>, PMID: 19104670
- Kuhn M**, Gerlicher AMV, Lonsdorf TB. 2022. Navigating the manyverse of skin conductance response quantification approaches—a direct comparison of trough-to-peak, baseline correction, and model-based approaches in ledalab and pspm. *Psychophysiology* **59**:e14058. DOI: <https://doi.org/10.1111/psyp.14058>, PMID: 35365863
- Lang PJ**, Melamed BG, Hart J. 1970. A psychophysiological analysis of fear modification using an automated desensitization procedure. *Journal of Abnormal Psychology* **76**:220–234. DOI: <https://doi.org/10.1037/h0029875>, PMID: 5483369
- Levine DW**, Dunlap WP. 1982. Power of the F test with skewed data: should one transform or not? *Psychological Bulletin* **92**:272–280. DOI: <https://doi.org/10.1037/0033-2909.92.1.272>
- Lonsdorf TB**, Haaker J, Kalisch R. 2014. Long-Term expression of human contextual fear and extinction memories involves amygdala, hippocampus and ventromedial prefrontal cortex: a reinstatement study in two independent samples. *Social Cognitive and Affective Neuroscience* **9**:1973–1983. DOI: <https://doi.org/10.1093/scan/nsu018>, PMID: 24493848

- Lonsdorf TB**, Menz MM, Andreatta M, Fullana MA, Golkar A, Haaker J, Heitland I, Hermann A, Kuhn M, Kruse O, Meir Drexler S, Meulders A, Nees F, Pittig A, Richter J, Römer S, Shiban Y, Schmitz A, Straube B, Vervliet B, et al. 2017a. Don't fear "fear conditioning": methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience and Biobehavioral Reviews* **77**:247–285. DOI: <https://doi.org/10.1016/j.neubiorev.2017.02.026>, PMID: 28263758
- Lonsdorf TB**, Merz CJ. 2017b. More than just noise: inter-individual differences in fear acquisition, extinction and return of fear in humans—biological, experiential, temperamental factors, and methodological pitfalls. *Neuroscience and Biobehavioral Reviews* **80**:703–728. DOI: <https://doi.org/10.1016/j.neubiorev.2017.07.007>, PMID: 28764976
- Lonsdorf TB**, Klingelhöfer-Jens M, Andreatta M, Beckers T, Chalkia A, Gerlicher A, Jentsch VL, Meir Drexler S, Mertens G, Richter J, Sjouwerman R, Wendt J, Merz CJ. 2019a. Navigating the garden of forking paths for data exclusions in fear conditioning research. *eLife* **8**:e52465. DOI: <https://doi.org/10.7554/eLife.52465>, PMID: 31841112
- Lonsdorf TB**, Merz CJ, Fullana MA. 2019b. Fear extinction retention: is it what we think it is? *Biological Psychiatry* **85**:1074–1082. DOI: <https://doi.org/10.1016/j.biopsych.2019.02.011>, PMID: 31005240
- Lonsdorf TB**, Gerlicher A, Klingelhöfer-Jens M, Kryptos AM. 2022. Multiverse analyses in fear conditioning research. *Behaviour Research and Therapy* **153**:104072. DOI: <https://doi.org/10.1016/j.brat.2022.104072>, PMID: 35500540
- Lykken DT**, Venables PH. 1971. Direct measurement of skin conductance: a proposal for standardization. *Psychophysiology* **8**:656–672. DOI: <https://doi.org/10.1111/j.1469-8986.1971.tb00501.x>, PMID: 5116830
- Lykken DT**. 1972. Range correction applied to heart rate and to GSR data. *Psychophysiology* **9**:373–379. DOI: <https://doi.org/10.1111/j.1469-8986.1972.tb03222.x>, PMID: 5034126
- Lynam DR**, Hoyle RH, Newman JP. 2006. The perils of partialling: cautionary tales from aggression and psychopathy. *Assessment* **13**:328–341. DOI: <https://doi.org/10.1177/1073191106290562>, PMID: 16880283
- Maitra R**. 2010. A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps. *NeuroImage* **50**:124–135. DOI: <https://doi.org/10.1016/j.neuroimage.2009.11.070>, PMID: 19963068
- Månsson KNT**, Waschke L, Manzouri A, Furmark T, Fischer H, Garrett DD. 2021. Moment-to-moment brain signal variability reliably predicts psychiatric treatment outcome. *Biological Psychiatry* **91**:658–666. DOI: <https://doi.org/10.1016/j.biopsych.2021.09.026>, PMID: 34961621
- Marek S**, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, Dosenbach NUF. 2020. Towards Reproducible Brain-Wide Association Studies. *bioRxiv*. DOI: <https://doi.org/10.1101/2020.08.21.257758>
- MATLAB**. 2016. Matlab, natick. 0.1. The MathWorks, Inc. <https://www.mathworks.com/company/jobs/resources/locations/us-natick.html>
- MATLAB**. 2019. Matlab, sherborn. 4.1. The MathWorks, Inc. <https://www.indeed.com/jobs?q=The+Mathworks&l=Sherborn,+MA&redirected=1>
- Milad MR**, Wright CI, Orr SP, Pitman RK, Quirk GJ, Rauch SL. 2007. Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biological Psychiatry* **62**:446–454. DOI: <https://doi.org/10.1016/j.biopsych.2006.10.011>, PMID: 17217927
- Milad MR**, Pitman RK, Ellis CB, Gold AL, Shin LM, Lasko NB, Zeidan MA, Handwerker K, Orr SP, Rauch SL. 2009. Neurobiological basis of failure to recall extinction memory in posttraumatic stress disorder. *Biological Psychiatry* **66**:1075–1082. DOI: <https://doi.org/10.1016/j.biopsych.2009.06.026>, PMID: 19748076
- Milad MR**, Quirk GJ. 2012. Fear extinction as a model for translational neuroscience: ten years of progress. *Annual Review of Psychology* **63**:129–151. DOI: <https://doi.org/10.1146/annurev.psych.121208.131631>, PMID: 22129456
- Moriarty DP**, Alloy LB. 2021. Back to basics: the importance of measurement properties in biological psychiatry. *Neuroscience and Biobehavioral Reviews* **123**:72–82. DOI: <https://doi.org/10.1016/j.neubiorev.2021.01.008>, PMID: 33497789
- Müller K**. 2020. Here: a simpler way to find your files. R-Project. <https://CRAN.R-project.org/package=here>
- Myers KM**, Davis M. 2007. Mechanisms of fear extinction. *Molecular Psychiatry* **12**:120–150. DOI: <https://doi.org/10.1038/sj.mp.4001939>, PMID: 17160066
- Ney LJ**, Laing PAF, Steward T, Zuj DV, Dymond S, Felmingham KL. 2020. Inconsistent analytic strategies reduce robustness in fear extinction via skin conductance response. *Psychophysiology* **57**:11. DOI: <https://doi.org/10.1111/psyp.13650>
- Ney LJ**, Laing PAF, Steward T, Zuj DV, Dymond S, Harrison B, Graham B, Felmingham KL. 2022. Methodological implications of sample size and extinction gradient on the robustness of fear conditioning across different analytic strategies. *PLOS ONE* **17**:e0268814. DOI: <https://doi.org/10.1371/journal.pone.0268814>, PMID: 35609058
- Nieuwenhuys R**. 2012. The insular cortex. *Progress in Brain Research* **195**:123–163. DOI: <https://doi.org/10.1016/B978-0-444-53860-4.00007-6>
- Noble S**, Scheinost D, Constable RT. 2021. A guide to the measurement and interpretation of fMRI test-retest reliability. *Current Opinion in Behavioral Sciences* **40**:27–32. DOI: <https://doi.org/10.1016/j.cobeha.2020.12.012>, PMID: 33585666
- Nosek BA**, Ebersole CR, DeHaven AC, Mellor DT. 2018. The preregistration revolution. *PNAS* **115**:2600–2606. DOI: <https://doi.org/10.1073/pnas.1708274114>, PMID: 29531091
- Parsons S**, Kruijt AW, Fox E. 2019. Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science* **2**:378–395. DOI: <https://doi.org/10.1177/2515245919879695>

- Parsons S.** 2020. Exploring Reliability Heterogeneity with Multiverse Analyses: Data Processing Decisions Unpredictably Influence Measurement Reliability. *PsyArXiv*. DOI: <https://doi.org/10.31234/osf.io/y6tzc>
- Pedersen TL.** 2020. Patchwork: the composer of plots. R-Project. <https://CRAN.R-project.org/package=patchwork>
- Pitman RK, Orr SP, Altman B, Longpre RE, Poiré RE, Macklin ML, Michaels MJ, Steketee GS.** 1996. Emotional processing and outcome of imaginal flooding therapy in Vietnam veterans with chronic posttraumatic stress disorder. *Comprehensive Psychiatry* **37**:409–418. DOI: [https://doi.org/10.1016/s0010-440x\(96\)90024-3](https://doi.org/10.1016/s0010-440x(96)90024-3), PMID: 8932965
- Plendl W, Wotjak CT.** 2010. Dissociation of within- and between-session extinction of conditioned fear. *The Journal of Neuroscience* **30**:4990–4998. DOI: <https://doi.org/10.1523/JNEUROSCI.6038-09.2010>, PMID: 20371819
- Plichta MM, Schwarz AJ, Grimm O, Morgen K, Mier D, Haddad L, Gerdes ABM, Sauer C, Tost H, Esslinger C, Colman P, Wilson F, Kirsch P, Meyer-Lindenberg A.** 2012. Test-Retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *NeuroImage* **60**:1746–1758. DOI: <https://doi.org/10.1016/j.neuroimage.2012.01.129>, PMID: 22330316
- Plichta MM, Grimm O, Morgen K, Mier D, Sauer C, Haddad L, Tost H, Esslinger C, Kirsch P, Schwarz AJ, Meyer-Lindenberg A.** 2014. Amygdala habituation: a reliable fMRI phenotype. *NeuroImage* **103**:383–390. DOI: <https://doi.org/10.1016/j.neuroimage.2014.09.059>, PMID: 25284303
- Prenoveau JM, Craske MG, Liao B, Ornitz EM.** 2013. Human fear conditioning and extinction: timing is everything...or is it? *Biological Psychology* **92**:59–68. DOI: <https://doi.org/10.1016/j.biopsycho.2012.02.005>, PMID: 22349998
- Presentation software.** 2010. Presentation software. Neurobehavioral Systems, Inc. <https://www.neurobs.com/>
- Rachman S.** 1989. The return of fear: review and prospect. *Clinical Psychology Review* **9**:147–168. DOI: [https://doi.org/10.1016/0272-7358\(89\)90025-1](https://doi.org/10.1016/0272-7358(89)90025-1)
- Rauch SAM, Foa EB, Furr JM, Filip JC.** 2004. Imagery vividness and perceived anxious arousal in prolonged exposure treatment for PTSD. *Journal of Traumatic Stress* **17**:461–465. DOI: <https://doi.org/10.1007/s10960-004-5794-8>, PMID: 15730064
- R Development Core Team.** 2020. R: a language and environment for statistical computing. Vienna, Austria. R Foundation for Statistical Computing. <https://www.r-project.org/index.html>
- Revelle W.** 1979. Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research* **14**:57–74. DOI: https://doi.org/10.1207/s15327906mbr1401_4, PMID: 26766619
- Revelle W.** 2020. Psych: procedures for psychological, psychometric, and personality research. R-Project. <https://CRAN.R-project.org/package=psych>
- Ridderbusch IC, Wroblewski A, Yang Y, Richter J, Hollandt M, Hamm AO, Wittchen HU, Ströhle A, Arolt V, Margraf J, Lueken U, Herrmann MJ, Kircher T, Straube B.** 2021. Neural adaptation of cingulate and insular activity during delayed fear extinction: a replicable pattern across assessment sites and repeated measurements. *NeuroImage* **237**:118157. DOI: <https://doi.org/10.1016/j.neuroimage.2021.118157>, PMID: 34020017
- Riley WT, McCormick MGF, Simon EM, Stack K, Pushkin Y, Overstreet MM, Carmona JJ, Magakian C.** 1995. Effects of alprazolam dose on the induction and habituation processes during behavioral panic induction treatment. *Journal of Anxiety Disorders* **9**:217–227. DOI: [https://doi.org/10.1016/0887-6185\(95\)00003-7](https://doi.org/10.1016/0887-6185(95)00003-7)
- Rothbaum BO, Price M, Jovanovic T, Norrholm SD, Gerardi M, Dunlop B, Davis M, Bradley B, Duncan EJ, Rizzo A, Ressler KJ.** 2014. A randomized, double-blind evaluation of D-cycloserine or alprazolam combined with virtual reality exposure therapy for posttraumatic stress disorder in Iraq and Afghanistan war veterans. *The American Journal of Psychiatry* **171**:640–648. DOI: <https://doi.org/10.1176/appi.ajp.2014.13121625>, PMID: 24743802
- Scharfenort R, Menz M, Lonsdorf TB.** 2016. Adversity-induced relapse of fear: neural mechanisms and implications for relapse prevention from a study on experimentally induced return-of-fear following fear conditioning and extinction. *Translational Psychiatry* **6**:e858. DOI: <https://doi.org/10.1038/tp.2016.126>, PMID: 27434492
- Schumann D, Joue G, Jordan P, Bayer J, Sommer T.** 2020. Test-Retest reliability of the emotional enhancement of memory. *Memory* **28**:49–59. DOI: <https://doi.org/10.1080/09658211.2019.1679837>, PMID: 31612770
- Seel NM.** 2012. Rescorla-Wagner model. Seel NM (Ed). *Encyclopedia of the Sciences of Learning*. Springer US. DOI: https://doi.org/10.1007/978-1-4419-1428-6_2377
- Shrout PE, Fleiss JL.** 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* **86**:420–428. DOI: <https://doi.org/10.1037//0033-2909.86.2.420>, PMID: 18839484
- Shumake J, Furgeson-Moreira S, Monfils MH.** 2014. Predictability and heritability of individual differences in fear learning. *Animal Cognition* **17**:1207–1221. DOI: <https://doi.org/10.1007/s10071-014-0752-1>, PMID: 24791664
- Sjowerman R, Lonsdorf TB.** 2019. Latency of skin conductance responses across stimulus modalities. *Psychophysiology* **56**:e13307. DOI: <https://doi.org/10.1111/psyp.13307>, PMID: 30461024
- Sjowerman R, Illius S, Kuhn M, Lonsdorf TB.** 2022. A data multiverse analysis investigating non-model based SCR quantification approaches. *Psychophysiology* **1**:e14130. DOI: <https://doi.org/10.1111/psyp.14130>, PMID: 35780077
- Smits JAJ, Rosenfield D, Otto MW, Marques L, Davis ML, Meuret AE, Simon NM, Pollack MH, Hofmann SG.** 2013a. D-Cycloserine enhancement of exposure therapy for social anxiety disorder depends on the success of exposure sessions. *Journal of Psychiatric Research* **47**:1455–1461. DOI: <https://doi.org/10.1016/j.jpsychires.2013.06.020>, PMID: 23870811

- Smits JAJ, Rosenfield D, Otto MW, Powers MB, Hofmann SG, Telch MJ, Pollack MH, Tart CD. 2013b. D-Cycloserine enhancement of fear extinction is specific to successful exposure sessions: evidence from the treatment of height phobia. *Biological Psychiatry* **73**:1054–1058. DOI: <https://doi.org/10.1016/j.biopsych.2012.12.009>, PMID: 23332511
- Spearman C. 1910. Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920* **3**:271–295. DOI: <https://doi.org/10.1111/j.2044-8295.1910.tb00206.x>
- Specht J, Egloff B, Schmukle SC. 2011. Stability and change of personality across the life course: the impact of age and major life events on mean-level and rank-order stability of the big five. *Journal of Personality and Social Psychology* **101**:862–882. DOI: <https://doi.org/10.1037/a0024950>, PMID: 21859226
- Spielberger CD. 1983. Manual for the State-Trait Inventory STAI. Palo Alto, CA: Mind Garden.
- Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**:702–712. DOI: <https://doi.org/10.1177/1745691616658637>, PMID: 27694465
- Thomas DR, Zumbo BD. 2012. Difference scores from the point of view of reliability and repeated-measures ANOVA: in defense of difference scores for data analysis. *Educational and Psychological Measurement* **72**:37–43. DOI: <https://doi.org/10.1177/0013164411409929>
- Tiedemann F. 2020. Gghalves: compose half-half plots using your favourite geoms. R-Project. <https://CRAN.R-project.org/package=gghalves>
- Torchiano M. 2020. Effsize: efficient effect size computation. Zenodo. <https://doi.org/10.5281/zenodo.1480624>
- Torrents-Rodas D, Fullana MA, Bonillo A, Andi6n O, Molinuevo B, Caseras X, Torrubia R. 2014. Testing the temporal stability of individual differences in the acquisition and generalization of fear. *Psychophysiology* **51**:697–705. DOI: <https://doi.org/10.1111/psyp.12213>, PMID: 24673651
- Ushey K. 2020. Renv: project environments. R-Project. <https://CRAN.R-project.org/package=renv>
- Vaidya JG, Gray EK, Haig J, Watson D. 2002. On the temporal stability of personality: evidence for differential stability and the role of life experiences. *Journal of Personality and Social Psychology* **83**:1469–1484. DOI: <https://doi.org/10.1037/0022-3514.83.6.1469>, PMID: 12500825
- Vervliet B, Baeyens F, Van den Bergh O, Hermans D. 2013a. Extinction, generalization, and return of fear: a critical review of renewal research in humans. *Biological Psychology* **92**:51–58. DOI: <https://doi.org/10.1016/j.biopsycho.2012.01.006>, PMID: 22285129
- Vervliet B, Craske MG, Hermans D. 2013b. Fear extinction and relapse: state of the art. *Annual Review of Clinical Psychology* **9**:215–248. DOI: <https://doi.org/10.1146/annurev-clinpsy-050212-185542>, PMID: 23537484
- Visser RM, de Haan MIC, Beemsterboer T, Haver P, Kindt M, Scholte HS. 2016. Quantifying learning-dependent changes in the brain: single-trial multivoxel pattern analysis requires slow event-related fMRI. *Psychophysiology* **53**:1117–1127. DOI: <https://doi.org/10.1111/psyp.12665>, PMID: 27153295
- Visser RM, Bathelt J, Scholte HS, Kindt M. 2021. Robust BOLD responses to faces but not to conditioned threat: challenging the amygdala ' S reputation in human fear and extinction learning. *The Journal of Neuroscience* **41**:10278–10292. DOI: <https://doi.org/10.1523/JNEUROSCI.0857-21.2021>, PMID: 34750227
- Werner F, Klingelh6fer-Jens M, Sch6umann D, Gamer M, Kalisch R, Sommer T, Lonsdorf TB. 2022. Limited Temporal Stability of the Spielberger State-Trait Inventory over 3.5 Years. *PsyArXiv*. DOI: <https://doi.org/10.31234/osf.io/mubgv>
- Wickham H. 2007. Reshaping data with the reshape package. *Journal of Statistical Software* **21**:1–20. DOI: <https://doi.org/10.18637/jss.v021.i12>
- Wickham H. 2016. Elegant graphics for data analysis. 3.3.5. Ggplot2. <https://ggplot2.tidyverse.org>
- Wickham H. 2019. Stringr: simple, consistent wrappers for common string operations. 1.4.1. R-Project. <https://CRAN.R-project.org/package=stringr>
- Wickham H. 2020. Tidy: tidy messy data. R-Project. <https://CRAN.R-project.org/package=tidy>
- Wickham H, Fran6ois R, Henry L, M6ller K. 2021. Dplyr: a grammar of data manipulation. R-Project. <https://CRAN.R-project.org/package=dplyr>
- Wilke CO. 2020. Cowplot: streamlined plot theme and plot annotations for ' ggplot2. R-Project. <https://CRAN.R-project.org/package=cowplot>
- Xie Y. 2015. Dynamic documents with R and knitr. Yihui. <https://yihui.org/knitr/>
- Xiong P, Liu M, Liu B, Hall BJ. 2022. Trends in the incidence and dalys of anxiety disorders at the global, regional, and national levels: estimates from the global burden of disease study 2019. *Journal of Affective Disorders* **297**:83–93. DOI: <https://doi.org/10.1016/j.jad.2021.10.022>, PMID: 34678404
- Yonkers KA, Bruce SE, Dyck IR, Keller MB. 2003. Chronicity, relapse, and illness? course of panic disorder, social phobia, and generalized anxiety disorder: findings in men and women from 8 years of follow-up. *Depression and Anxiety* **17**:173–179. DOI: <https://doi.org/10.1002/da.10106>, PMID: 12768651
- Zeidan MA, Lebron-Milad K, Thompson-Hollands J, Im JY, Dougherty DD, Holt DJ, Orr SP, Milad MR. 2012. Test–retest reliability during fear acquisition and fear extinction in humans. *CNS Neuroscience & Therapeutics* **18**:313–317. DOI: <https://doi.org/10.1111/j.1755-5949.2011.00238.x>, PMID: 21592319
- Zeileis A, Hothorn T. 2002. Diagnostic checking in regression relationships. *R News* **2**:7–10.
- Zeileis A. 2004. Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software* **11**:1–17. DOI: <https://doi.org/10.18637/jss.v011.i10>
- Zeileis A, Grothendieck G. 2005. Zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software* **14**:1–27. DOI: <https://doi.org/10.18637/jss.v014.i06>
- Zeileis A. 2006. Object-oriented computation of sandwich estimators. *Journal of Statistical Software* **16**:1–16. DOI: <https://doi.org/10.18637/jss.v016.i09>

- Zeileis A**, Köll S, Graham N. 2020. Various versatile variances: an object-oriented implementation of clustered covariances in R. *Journal of Statistical Software* **95**:1–36. DOI: <https://doi.org/10.18637/jss.v095.i01>
- Zhu H**. 2020. KableExtra: construct complex table with ' kable ' and pipe SYNTAX. R-Project. <https://CRAN.R-project.org/package=kableExtra>
- Zuo XN**, Xu T, Milham MP. 2019. Harnessing reliability for neuroscience research. *Nature Human Behaviour* **3**:768–771. DOI: <https://doi.org/10.1038/s41562-019-0655-x>, PMID: 31253883

SUPPLEMENT FOR**Robust group- but limited individual-level (longitudinal) reliability and insights into cross-phases response prediction of conditioned fear**

Maren Klingelhöfer-Jens, Mana R. Ehlers, Manuel Kuhn, Vincent Keyaniyan, & Tina B. Lonsdorf

Supplementary file 1: Overview of experimental specifications and results of five previous studies reporting test-retest reliabilities in human fear conditioning research.

	Fredrikson et al., 1993	Zeidan et al., 2012	Torrents-Rodas et al., 2014	Ridderbusch et al., 2021	Cooper et al., PREPRINT
N/female/male/age	28/14/14/M = 28.5 (\pm 1.42)	18/9/9/M = 38.0 (\pm 12.7)	71/52/19/M = 22.4 (\pm 2.61)	100/46/54/M = 33.1 (\pm 10.7)	51/39/12/ M = 20.0 (\pm 2.88)
Reinforcement rate (%)	100	100	Acquisition training: 75 Generalization: 50	60	40
Acquisition type	Not reported	Not reported	Uninstructed but informed about the existence of contingencies ³	Instructed (but not informed about the reinforcement rate)	Instructed (but not informed about the reinforcement rate)
Extinction type	Immediate	Immediate; Extinction training consisted	None	24h delayed; Extinction training consisted of 2 subphases (Ex1 and Ex2)	None

		of 2 subphases separated by a 1-min rest period			
Additional phase(s)	None	24h delayed extinction recall immediately followed by renewal	Generalization (10 min. after acquisition training)	One re-acquisition trial prior to extinction training Reinstatement-test (immediately after extinction training and reinstatement)	Generalization
CS quality	Geometric shapes	Lamp in a room (2 colors)	2 rings as CSs and 8 rings as GSs	Neutral faces on colored background (background color = CS type)	Auditory (pure tone sine waves < 60 decibels): CS+ and CS- = 1000 and 550 Hz; 6 GSs = 650, 800, 900, 1100, 1200 and 1350 Hz
CS duration (s)	8	12	8	6	2.5
ITI duration (s)	20 – 40	12 – 21	9 – 17	6 – 10	7 – 8
US type	Auditory (110 dB white noise)	Electrotactile	Electrotactile	Electrotactile	Electrotactile
# of habituation trials CS+/CS-	4/4	4/4	6/6	2/2	No habituation phase
# of acquisition trials CS+/CS-	8/8	5/5	12/12	10/10	20/12
# of extinction trials CS+/CS-	8/8	5/5 (in each of the 2 subphases)	No extinction phase	Extinction phase 1 (Ex1): 10/10 Extinction phase 2 (Ex2): 10/10	No extinction phase
# of trials add. phase CS+/CS-	No additional phase	Extinction recall: 5/5 Renewal: 5/5	12/12 6 times each GS	Re-acquisition: 1/0 Reinstatement-test: 10/10	Generalization: 12/7 7 times each GS
SCR	Yes	Yes	Yes	Yes	Yes
FPS	No	No	Yes	Yes	No

Ratings	No	No	Risk ratings	Expectancy, arousal, valence ratings	Shock risk ratings
fMRI	No	No	No	Yes	No
Reported measure(s)	SCR	SCR	SCR, FPS, ratings	fMRI, ratings	SCR, ratings
# of measurement time points	2	3	2	2	2
Time gap between measurement time points	20 days	Time points 1 and 2: 17.9 ± 2.1 weeks Time points 2 and 3: 14.5 ± 0.7 weeks	5.8 - 9.0 months (M = 7.7)	13 weeks	9 days
Same stimuli used in retest	Not reported	No ¹	Yes (half of the participants) New set (other half of the participants) (new stimuli = lines with varying slopes)	No ⁵	Yes
Same allocation of stimuli to CSs	Not reported	No ²	Yes (applies to the use of the same stimulus set)	No ⁵	Yes
Reliability measure	Pearson's r	ICC (no type specified)	G coefficient (range = 0 - 1) 4	ICC(1,1) ⁶	G coefficient (range = 0 - 1) 4
Included trials	All	All	All	See results and notes below	All
Test-Retest Habituation					
CS+	SCR (FIR): 0.62	SCR (time points 1-3): 0.10 SCR (time points 1-2): 0.16	Not reported	fMRI No fMRI data for habituation Ratings No rating data for habituation	No habituation phase
CS-	SCR (FIR): 0.72	Not reported			

Test-Retest Acquisition					
CS+	SCR (FIR): 0.85 SCR (SIR): 0.51 SCR (TIR): 0.65	SCR (time points 1-3): 0.68 SCR (time points 1-2): 0.64	Same stimulus set ⁴ SCR: 0.27 FPS: 0.34 Ratings: 0.23	fMRI No fMRI data for acquisition training	SCR ⁴ : 0.50 Ratings ⁴ : 0.47
CS-	SCR (FIR): 0.57 SCR (SIR): 0.27 SCR (TIR): 0.29	Not reported	New stimulus set ⁴ SCR: 0.39 FPS: 0.40 Ratings: 0.46	Ratings Expectancy: not reported Arousal: no data for acquisition training Valence: no data for acquisition training	
CS discrimination	Not included	SCR (time points 1-3): 0.43			
Test-Retest Extinction					
CS+	SCR (FIR): 0.62 SCR (SIR): 0.27 SCR (TIR): 0.83	SCR (time points 1-3): -0.19 SCR (time points 1-2): -0.24	No extinction phase	fMRI Ex1 Right insula: 0.54 Left insula: 0.57 Middle cingulate cortex: 0.40 Ex1 > Ex2 Left insula: 0.22 Right insula: 0.14 Middle cingulate cortex: 0.29 Ratings pre Ex1 Expectancy: no data for pre Ex1 Arousal: not reported Valence: not reported Post Re-Acq, post Ex1 and post Ex2 ⁷ Expectancy: 0.66 Arousal: 0.63 Valence: 0.56	No extinction phase

CS-	SCR (FIR): 0.37 SCR (SIR): -0.05 SCR (TIR): 0.09	Not reported		Not reported	
CS discrimination	Not included	Not reported		<p>fMRI</p> <p>Ex1 Right insula: 0.44 Left insula: 0.39 Middle cingulate cortex: 0.34</p> <p>Ex1 > Ex2 Left insula: 0.20 Right insula: 0.01 Middle cingulate cortex: 0.13</p> <p>Ratings</p> <p>pre Ex1 Expectancy: no data for pre Ex1 Arousal: 0.42 Valence: 0.02</p> <p>Post Re-Acq, post Ex1 and post Ex2⁷ Expectancy: 0.64 Arousal: 0.43 Valence: 0.25</p>	
Test-Retest additional phase		Extinction recall Renewal	Generalization	Re-Acquisition Reinstatement-Test	Generalization
CS+	No additional phase	<p>Extinction recall: SCR (time points 1-3): 0.46 SCR (time points 1-2): 0.72</p> <p>Renewal: SCR (time points 1-3): 0.67 SCR (time points 1-2): 0.66</p>	<p>Same stimulus set⁴ SCR: 0.44 FPS: 0.22 Ratings: 0.22</p> <p>New stimulus set⁴ SCR: 0.21</p>	<p>fMRI Not reported</p> <p>Ratings</p> <p>post Re-Acquisition Expectancy: 0.51 Arousal: 0.53 Valence: 0.49</p>	<p>SCR⁴: 0.39</p> <p>Ratings⁴: 0.36</p>

CS-		Not reported	FPS: 0.16 Ratings: 0.25	Not reported	
CS discrimination		<p>Extinction Recall SCR (time points 1-3): 0.23</p> <p>Renewal SCR (time points 1-3): 0.50</p>		<p>fMRI RI-T: Cingulate cortex cluster pre RI⁸: 0.01 post RI⁹: -0.05 pre vs. post RI^{8,9}: -0.12</p> <p>Ratings post Re-Acquisition Expectancy: 0.49 Arousal: not reported Valence: not reported</p> <p>pre RI¹⁰ Expectancy: 0.67 Arousal: 0.53 Valence: 0.39</p> <p>post RI Expectancy: 0.52 Arousal: 0.55 Valence: 0.34</p> <p>pre vs. post RI¹⁰ Expectancy: 0.22 Arousal: 0.19 Valence: -0.03</p>	
Physiological response quantification					
SCR quantification	Trough-to-peak (TTP)	Baseline correction	Baseline correction	Trough-to-peak (TTP)	Trough-to-peak (TTP)
SCR scoring criteria	FIR: 1-4 s after CS onset SIR: 5-9 s after CS onset TIR: 1-4 s after CS termination	Baseline: means SCL during 2 s before trial onset subtracted from the	Value at stimulus onset subtracted from the maximum value during 1-5 s after stimulus onset (only	First response occurring 0.9-4 s after stimulus onset	First response occurring 0.5-3 s after stimulus onset, lasting for 0.5 and 5.0 s, > 0.02 μ S

		highest SCL within the 12 s CS duration	trials without risk ratings analyzed)		
FPS specifications	No FPS applied	No FPS applied	5s after onset of odd trials and during ITIs (6 times per phase, IPIs 18-25 s)	Either 4.5 or 5 s after CS onset and during ITI (2, 3, 4, 5, or 6 s after CS offset); presented during all CS trials during habituation and during 8 of 10 CS trials during fear acquisition training	No FPS applied
FPS quantification			Baseline correction	Trough-to-peak (TTP)	
FPS scoring criteria			Value at response peak (Response onset in a time window 20-100 ms after probe onset with a peak between 20 and 150 ms after probe onset) subtracted from a baseline value (averaged during the 50 ms preceding the probe)	Response in a time window 20-120 ms after probe onset with a maximum peak within 150 ms after onset	
Ratings provided	No ratings provided	No ratings provided	During even trials	Expectancy: before each CS trial Arousal and Valence: post Re-Acq, pre Ex1, post Ex1, post Ex2, post RI, post RIT	Trial-by-trial shock expectancy

Note. We are aware that there is another study by Savage et al (2019) which investigated the test-retest reliability of fear potentiated startle in a differential fear conditioning paradigm. But since the participants of this study were twins and relatively young ($age_M = 11.5$, $age_{SD} = 1.5$), we have not included them in the table due to reduced comparability. # = number; FIR = first interval response, occurring 1-4s after CS onset; SIR = second interval response, occurring 5-9s after CS onset; TIR = third interval response, occurring 1-4s after CS termination; GS = generalization stimulus; Ex1 = first extinction phase; Ex2 = second extinction phase; pre/post = prior and subsequent to respective phases; Re-Acq = re-acquisition; RI-T = reinstatement-test; RI = reinstatement.

¹“Conditioning context and color of the CS+ were different for each of the 3 sessions and counterbalanced across visits.” (Zeidan et al., 2012, p. 314)

²“The conditioning context and the color of the CS+ were different for each of the three test sessions and counterbalanced across visits.” (Zeidan et al., 2012, p. 315)

³“They were not instructed about the CS–US contingency, but were told that they might learn to predict the shock if they pay attention to the presented stimuli.” (Torrents-Rodas et al., 2014, p. 699)

⁴ The calculation of the G coefficient includes both responses to the CS+ and CS-. Thus, results are not separated for stimulus types.

⁵“The whole experimental protocol (t1) was repeated after an interval of an average of 13 weeks (second measurement: t2), using two different visual stimuli as CSs to avoid re-acquisition.” (Ridderbusch et al., 2021, p. 3)

⁶ One-way random effects model with single measures.

⁷ Post re-acq, post Ex1 and post Ex2 = “extinction training effect” (see Ridderbusch et al., 2021).

⁸ Pre RI means for fMRI: last half of Ex2 trials (5 trials).

⁹ Post RI means for fMRI: first half of RI-T (5 trials).

¹⁰ Pre RI means for ratings: post Ex2.

Deviations from the pre-registration

Supplementary file 2: Deviations from pre-registration.

Pre-registration	Deviation type	Manuscript	Justification
Not pre-registered	Additional analyses	Analyses of cross-sectional reliability	Considered to provide additional valuable information
Mixed-effects approach for calculation of ICCS	Changes analysis approach	ANOVA approach for calculation of ICCS	Statistical approach changed due to model non-convergence problems
Calculate ICCs for ranked and non-ranked data	Omitted pre-registered specification	Non-ranked ICCs only	During closer inspection of the conceptualization of ICC _{con} , we realised that it would be redundant to calculate both ICC _{abs} and ICC _{con} with ranked and non-ranked data as ICC _{con} itself ranks the data. Hence, we decided to calculate ICCs based on non-ranked data only.
Not pre-registered	Additional analyses	Inclusion of ICC for SCRs to the US and US aversiveness ratings	Considered to provide valuable information
Not pre-registered	Additional analyses	Additional phase operationalization: last two extinction trials	Considered to provide valuable information for completeness
Not pre-registered	Additional analyses	Trial-by trial ICCs for SCRs	Considered to provide additional valuable information
Not pre-registered	Additional analyses	Inclusion of analysis focusing on reliability at the group level for SCRs	Considered to provide additional valuable information
Not pre-registered	Additional outcome measure	Inclusion of fMRI as an outcome measure and corresponding reliability analyses as well as within-session predictability analyses	Considered to provide valuable information
Multiple linear regression with SCRs or fear ratings during both acquisition and extinction training as multiple predictors for responses at reinstatement-test	Changes in analysis approach	Simple linear regressions including SCRs or fear ratings during acquisition training as predictors and responding during reinstatement-test as criterion. Further checks of statistical assumptions revealed heteroscedasticity of the data. Therefore, we conducted simple linear regressions with robust standard errors instead of using classical OLS estimators	Due to multicollinearity of the predictors resulting from significant associations of responding during acquisition and extinction training the pre-registered analyses were not suitable
Not pre-registered	Additional analyses	We compare different patterns of SCR, fear rating and fMRI data after pooling them for certain data specifications/ROIs in predictability analyses	Considered to provide additional valuable information

Internal consistency of log-transformed as well as log-transformed and range corrected

SCRs

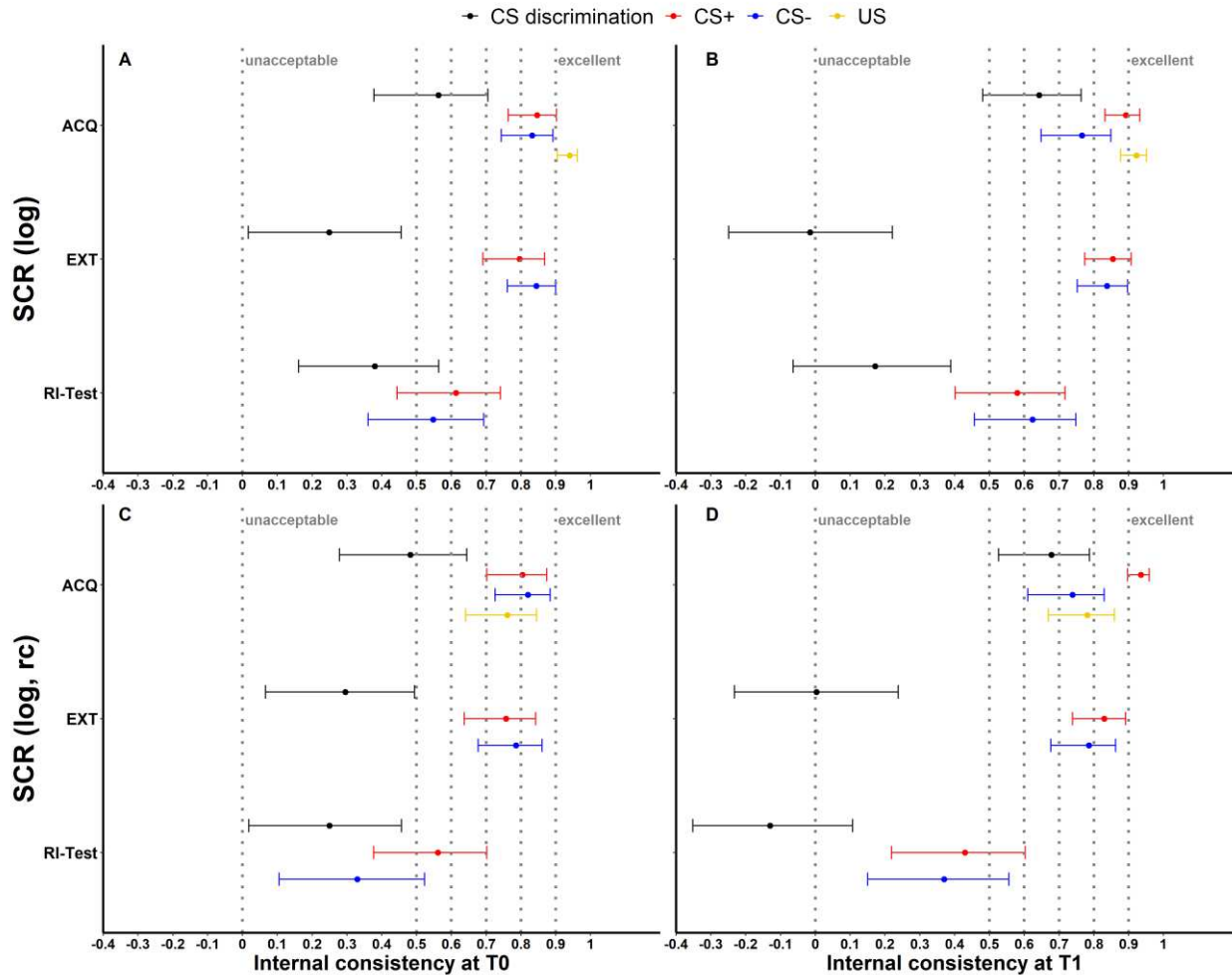


Figure 1 – figure supplement 1. Illustration of (A-B) internal consistency for log-transformed (log) as well as (C-D) log-transformed and range corrected (log rc) SCRs at T0 and T1 color coded for stimulus-type. Error bars represent 95% confidence intervals indicating significance, when zero is not included in the interval. The y-axis comprises the different experimental phases. Internal consistency is interpreted using benchmarks (Kline, 2013) for unacceptable (< 0.5), poor (> 0.5 but < 0.6), questionable (> 0.6 but < 0.7), acceptable (> 0.7 but < 0.8), good (> 0.8 but <

0.9) and excellent (≥ 0.9). ACQ = acquisition training, EXT = extinction training, RI-Test = reinstatement-test.

Detailed results of ICC calculations: SCR

Supplementary file 3: ICC_{abs} and ICC_{con} for all data specifications of SCRs.

Outcome	Ampl.-type	Stim.-type	Phase	Op.	ICC _{abs}				ICC _{con}			
					Value	Lower 95% CI	Upper 95% CI	p-value	Value	Lower 95% CI	Upper 95% CI	p-value
SCR	raw	CS dis.	Acq	average	0.160	-0.020	0.340	.077	0.170	-0.030	0.350	.077
SCR	raw	CS+	Acq	average	0.270	0.090	0.440	.006	0.300	0.110	0.460	.006
SCR	raw	CS-	Acq	average	0.390	0.210	0.540	.000	0.410	0.230	0.560	.000
SCR	raw	US	Acq	average	0.317	0.133	0.481	.003	0.322	0.135	0.487	.003
SCR	raw	CS dis.	Acq	last 2 trials	0.240	0.060	0.420	.017	0.250	0.060	0.430	.017
SCR	raw	CS+	Acq	last 2 trials	0.220	0.040	0.390	.025	0.230	0.040	0.410	.025
SCR	raw	CS-	Acq	last 2 trials	0.190	0.000	0.370	.055	0.190	-0.010	0.370	.055
SCR	raw	CS dis.	Ext	1st trial	0.190	0.010	0.360	.044	0.200	0.010	0.380	.044
SCR	raw	CS+	Ext	1st trial	0.320	0.140	0.490	.003	0.330	0.140	0.490	.003
SCR	raw	CS-	Ext	1st trial	0.040	-0.120	0.210	.331	0.050	-0.140	0.250	.331
SCR	raw	CS dis.	Ext	average	0.070	-0.130	0.270	.272	0.070	-0.120	0.260	.272
SCR	raw	CS+	Ext	average	0.480	0.320	0.620	.000	0.500	0.340	0.630	.000
SCR	raw	CS-	Ext	average	0.580	0.430	0.700	.000	0.610	0.470	0.720	.000
SCR	raw	CS dis.	Ext	last 2 trials	0.060	-0.140	0.250	.321	0.060	-0.140	0.250	.321
SCR	raw	CS+	Ext	last 2 trials	0.170	-0.020	0.360	.073	0.170	-0.020	0.360	.073
SCR	raw	CS-	Ext	last 2 trials	0.200	0.000	0.380	.048	0.200	0.000	0.380	.048
SCR	raw	CS dis.	RI-T	1st trial	0.140	-0.050	0.320	.119	0.140	-0.060	0.330	.119
SCR	raw	CS+	RI-T	1st trial	0.150	-0.040	0.330	.098	0.150	-0.040	0.340	.098
SCR	raw	CS-	RI-T	1st trial	0.030	-0.130	0.200	.389	0.030	-0.160	0.230	.389

Outcome	Ampl.-type	Stim.-type	Phase	Op.	ICC _{abs}				ICC _{con}			
					Value	Lower 95% CI	Upper 95% CI	p-value	Value	Lower 95% CI	Upper 95% CI	p-value
SCR	raw	US	RI	average	0.271	0.085	0.440	.009	0.278	0.087	0.449	.009
SCR	log	CS dis.	Acq	average	0.180	0.000	0.350	.054	0.190	0.000	0.370	.054
SCR	log	CS+	Acq	average	0.290	0.100	0.450	.004	0.310	0.130	0.480	.004
SCR	log	CS-	Acq	average	0.400	0.230	0.550	.000	0.420	0.250	0.570	.000
SCR	log	US	Acq	average	0.320	0.137	0.482	.003	0.327	0.140	0.491	.003
SCR	log	CS dis.	Acq	last 2 trials	0.230	0.040	0.400	.024	0.230	0.040	0.410	.024
SCR	log	CS+	Acq	last 2 trials	0.210	0.020	0.380	.032	0.220	0.030	0.400	.032
SCR	log	CS-	Acq	last 2 trials	0.190	0.000	0.370	.052	0.190	0.000	0.370	.052
SCR	log	CS dis.	Ext	1st trial	0.180	0.000	0.350	.052	0.190	0.000	0.370	.052
SCR	log	CS+	Ext	1st trial	0.310	0.120	0.470	.004	0.310	0.120	0.480	.004
SCR	log	CS-	Ext	1st trial	0.030	-0.130	0.200	.368	0.040	-0.160	0.230	.368
SCR	log	CS dis.	Ext	average	0.060	-0.140	0.260	.301	0.060	-0.130	0.250	.301
SCR	log	CS+	Ext	average	0.490	0.320	0.620	.000	0.500	0.340	0.640	.000
SCR	log	CS-	Ext	average	0.590	0.430	0.710	.000	0.620	0.480	0.720	.000
SCR	log	CS dis.	Ext	last 2 trials	0.070	-0.130	0.260	.283	0.070	-0.130	0.260	.283
SCR	log	CS+	Ext	last 2 trials	0.200	0.010	0.380	.044	0.200	0.010	0.380	.044
SCR	log	CS-	Ext	last 2 trials	0.230	0.040	0.410	.027	0.230	0.030	0.410	.027
SCR	log	CS dis.	RI-T	1st trial	0.170	-0.020	0.350	.076	0.170	-0.030	0.350	.076
SCR	log	CS+	RI-T	1st trial	0.160	-0.030	0.340	.083	0.160	-0.030	0.350	.083
SCR	log	CS-	RI-T	1st trial	0.040	-0.120	0.210	.337	0.050	-0.150	0.240	.337
SCR	log	US	RI	average	0.299	0.116	0.464	.004	0.308	0.120	0.475	.004
SCR	log rc	CS dis.	Acq	average	0.230	0.050	0.410	.018	0.250	0.050	0.420	.018
SCR	log rc	CS+	Acq	average	0.490	0.310	0.630	.000	0.530	0.370	0.660	.000
SCR	log rc	CS-	Acq	average	0.610	0.470	0.730	.000	0.630	0.500	0.740	.000
SCR	log rc	US	Acq	average	0.112	-0.086	0.301	.176	0.111	-0.086	0.300	.176
SCR	log rc	CS dis.	Acq	last 2 trials	0.050	-0.150	0.240	.337	0.050	-0.150	0.240	.337

Outcome	Ampl.- type	Stim.- type	Phase	Op.	ICC _{abs}				ICC _{con}			
					Value	Lower 95% CI	Upper 95% CI	p- value	Value	Lower 95% CI	Upper 95% CI	p- value
SCR	log rc	CS+	Acq	last 2 trials	0.300	0.120	0.470	.004	0.310	0.120	0.480	.004
SCR	log rc	CS-	Acq	last 2 trials	0.190	-0.010	0.370	.056	0.190	-0.010	0.370	.056
SCR	log rc	CS dis.	Ext	1st trial	0.200	0.020	0.370	.033	0.220	0.020	0.400	.033
SCR	log rc	CS+	Ext	1st trial	0.270	0.080	0.440	.012	0.270	0.070	0.440	.012
SCR	log rc	CS-	Ext	1st trial	-0.090	-0.250	0.090	.804	-0.100	-0.290	0.090	.804
SCR	log rc	CS dis.	Ext	average	0.350	0.160	0.510	.002	0.340	0.160	0.510	.002
SCR	log rc	CS+	Ext	average	0.540	0.380	0.670	.000	0.560	0.410	0.680	.000
SCR	log rc	CS-	Ext	average	0.620	0.440	0.740	.000	0.660	0.530	0.760	.000
SCR	log rc	CS dis.	Ext	last 2 trials	0.210	0.020	0.390	.037	0.210	0.020	0.390	.037
SCR	log rc	CS+	Ext	last 2 trials	0.360	0.170	0.520	.001	0.350	0.170	0.510	.001
SCR	log rc	CS-	Ext	last 2 trials	0.440	0.270	0.590	.000	0.440	0.270	0.590	.000
SCR	log rc	CS dis.	RI-T	1st trial	0.230	0.040	0.410	.023	0.240	0.040	0.410	.023
SCR	log rc	CS+	RI-T	1st trial	0.170	-0.020	0.360	.071	0.170	-0.020	0.360	.071
SCR	log rc	CS-	RI-T	1st trial	0.150	-0.030	0.330	.086	0.160	-0.030	0.350	.086
SCR	log rc	US	RI	average	0.093	-0.106	0.284	.221	0.092	-0.105	0.282	.221

Note. Ampl. = Amplitude, Stim. = Stimulus, Op. = Operationalization, CI = Confidence Interval, CS dis. = CS discrimination, log = log-transformed, log rc = log-transformed and range corrected, Acq = Acquisition training, Ext = Extinction training, RI = Reinstatement, RI-T = Reinstatement-Test.

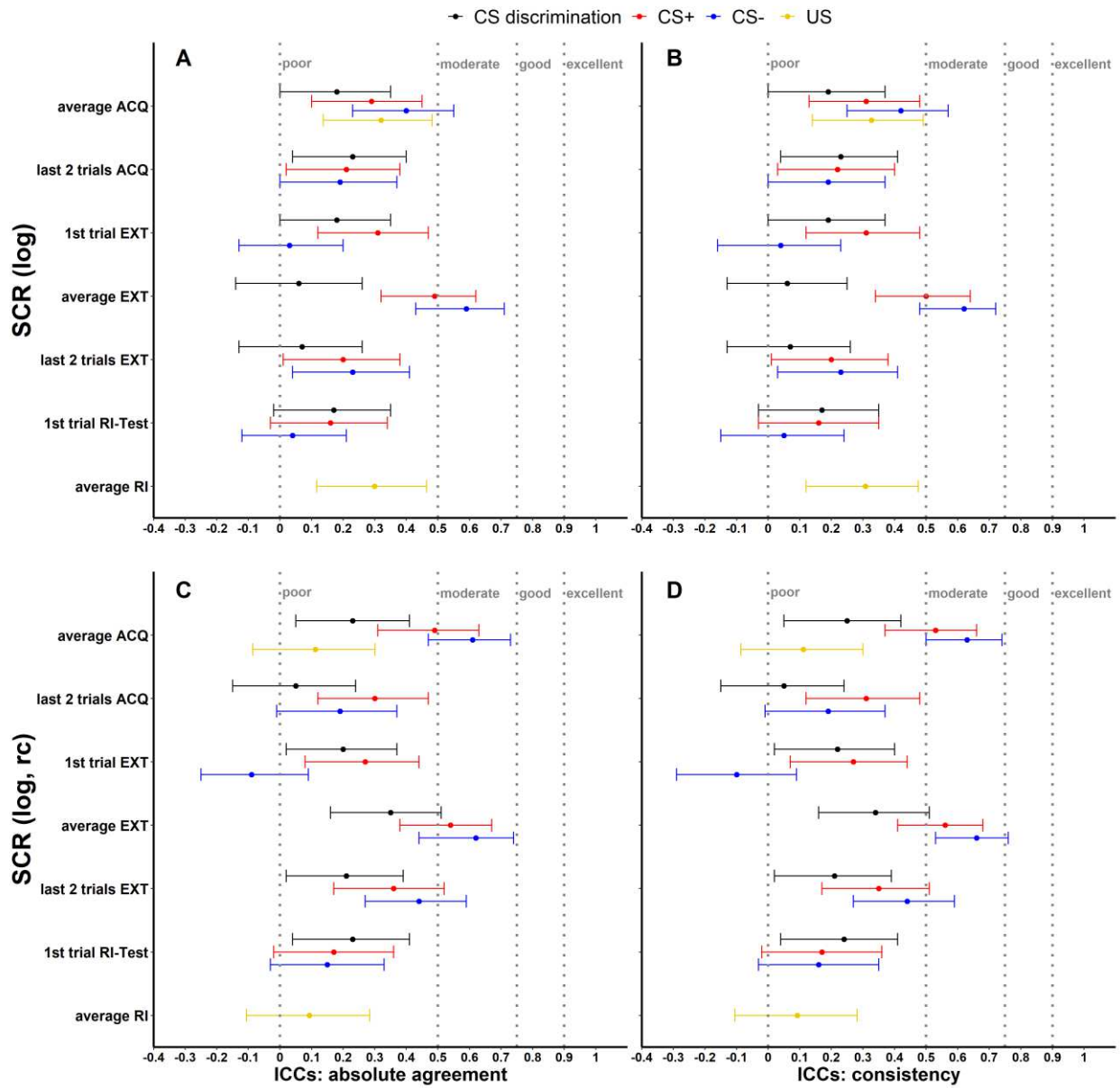


Figure 1-figure supplement 2. Illustration of (A-B) ICCs of log-transformed (log) as well as (C-D) log-transformed and range corrected (log, rc) SCRs color coded for stimulus-type. The y-axis comprises the different phase operationalizations. A and C display ICC_{abs} , B and D display ICC_{con} . ICCs < 0.5 , < 0.75 , < 0.9 and > 0.9 were interpreted as poor, moderate, good and excellent respectively (Koo & Li, 2016). Error bars represent 95% confidence intervals indicating

significance of ICCs, when zero is not included in the interval. ACQ = acquisition training, EXT = extinction training, RI = reinstatement, RI-Test = reinstatement-test.

ICCs of trial-by-trial SCRs

ICCs of trial-by-trial raw SCRs

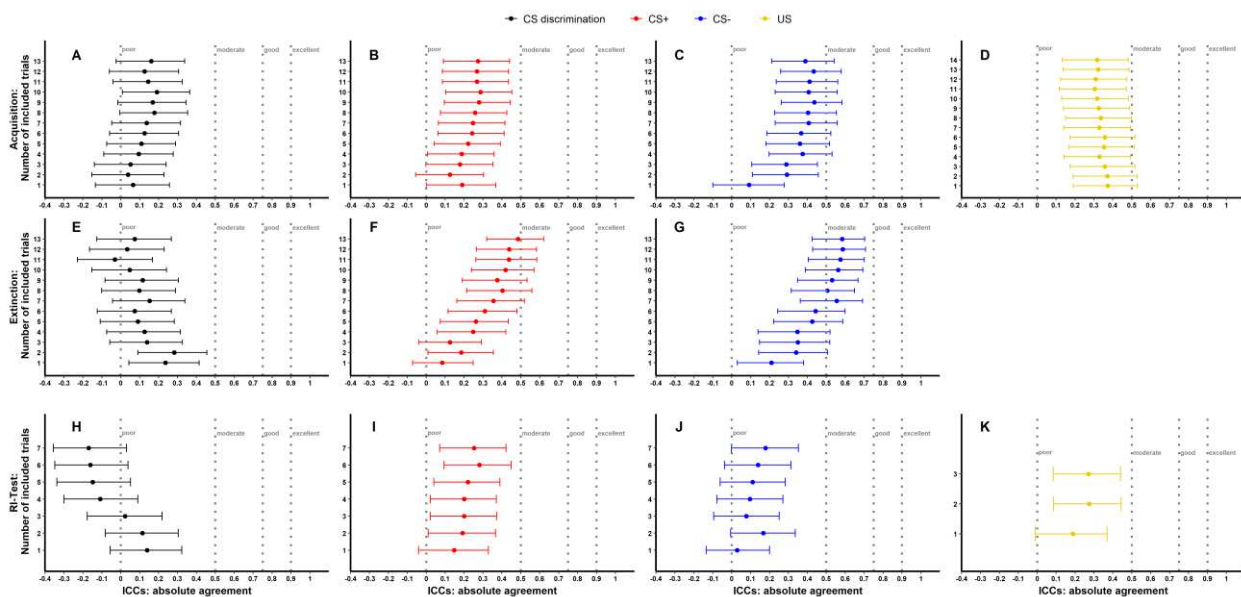


Figure 1-figure supplement 3. Illustration of ICC_{abs} of trial-by-trial raw SCRs for phases (A-D: Acquisition, E-G: Extinction, H-J: Reinstatement-Test, K: Reinstatement) and stimulus-types separately. Trials were averaged starting with the first (i.e., reinstatement-test and US trials) or second trial (i.e., acquisition and extinction training), adding all preceding trials trial-by-trial and averaged. ICCs < 0.5 , < 0.75 , < 0.9 and > 0.9 (Koo & Li, 2016) were interpreted as poor, moderate, good and excellent respectively. Error bars represent 95% confidence intervals. Non-overlapping error bars indicate significant differences between ICCs within one figure. RI-Test = reinstatement-test.

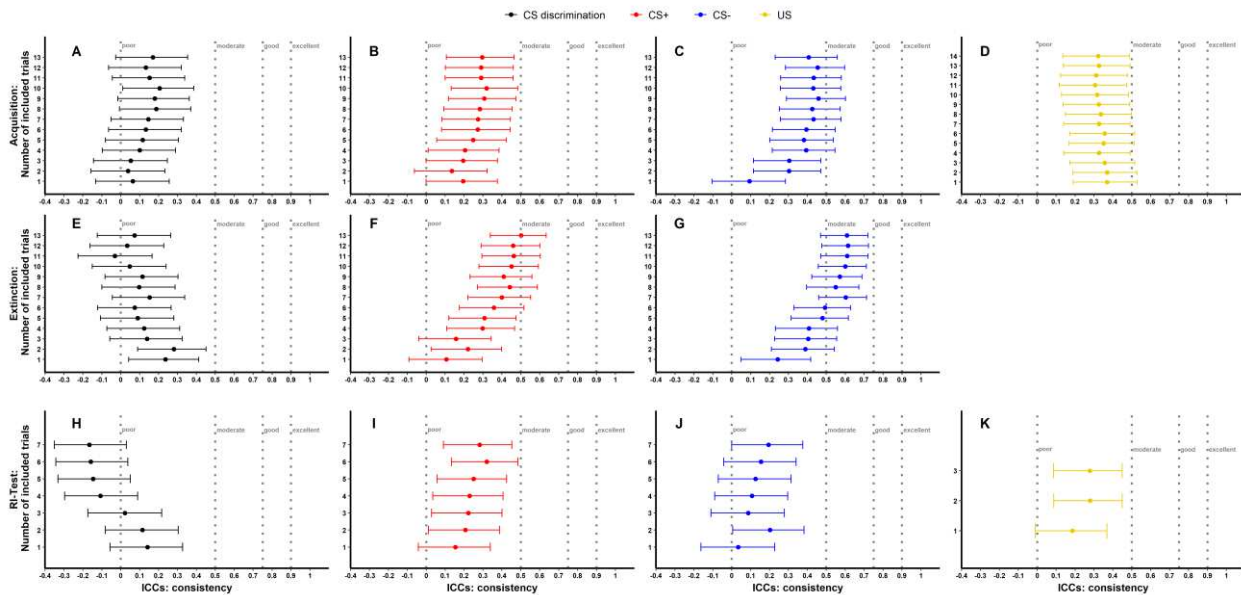


Figure 1-figure supplement 4. Illustration of ICC_{CON} of trial-by-trial raw SCRs for phases (A-D: Acquisition, E-G: Extinction, H-J: Reinstatement-Test, K: Reinstatement) and stimulus-types separately. Trials were averaged starting with the first (i.e., reinstatement-test and US trials) or second trial (i.e., acquisition and extinction training), adding all preceding trials trial-by-trial and averaged. ICCs < 0.5 , < 0.75 , < 0.9 and > 0.9 (Koo & Li, 2016) were interpreted as poor, moderate, good and excellent respectively. Error bars represent 95% confidence intervals. Non-overlapping error bars indicate significant differences between ICCs within one figure. RI-Test = reinstatement-test.

ICCs of trial-by-trial log-transformed SCRs

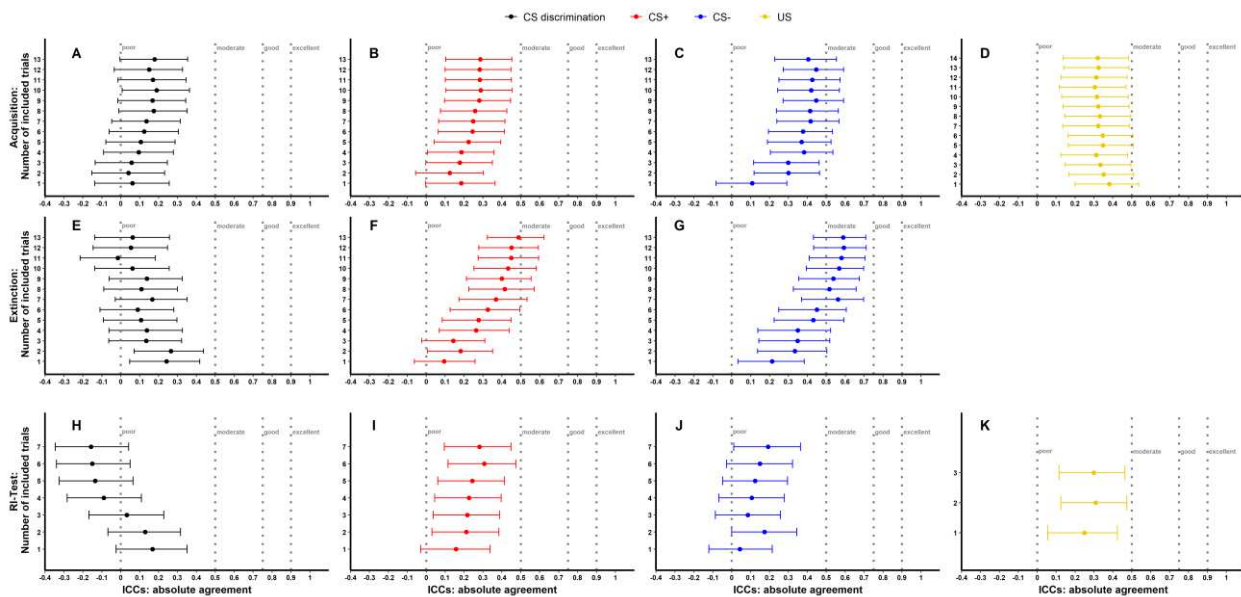


Figure 1-figure supplement 5. Illustration of ICC_{abs} of trial-by-trial log-transformed SCRs for phases (A-D: Acquisition, E-G: Extinction, H-J: Reinstatement-Test, K: Reinstatement) and stimulus-types separately. Trials were averaged starting with the first (i.e., reinstatement-test and US trials) or second trial (i.e., acquisition and extinction training), adding all preceding trials trial-by-trial and averaged. ICCs < 0.5, < 0.75, < 0.9 and > 0.9 (Koo & Li, 2016) were interpreted as poor, moderate, good and excellent respectively. Error bars represent 95% confidence intervals. Non-overlapping error bars indicate significant differences between ICCs within one figure. RI-Test = reinstatement-test.

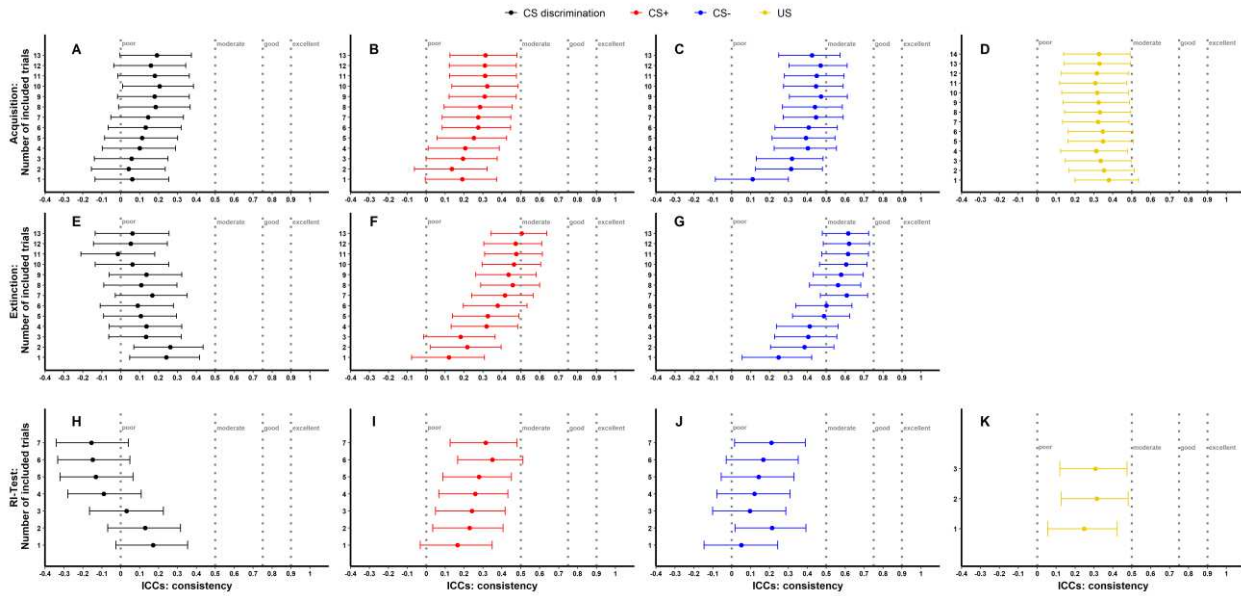


Figure 1-figure supplement 6. Illustration of ICC_{con} of trial-by-trial log-transformed SCRs for phases (A-D: Acquisition, E-G: Extinction, H-J: Reinstatement-Test, K: Reinstatement) and stimulus types separately. Trials were averaged starting with the first (i.e., reinstatement-test and US trials) or second trial (i.e., acquisition and extinction training). All preceding trials were added trial-by-trial and averaged. ICCs < 0.5, < 0.75, < 0.9 and > 0.9 (Koo & Li, 2016) were interpreted as poor, moderate, good and excellent respectively. Error bars represent 95% confidence intervals. Non-overlapping error bars indicate significant differences between ICCs within one figure. RI-Test = reinstatement-test.

ICCs of trial-by-trial log-transformed and range corrected SCRs

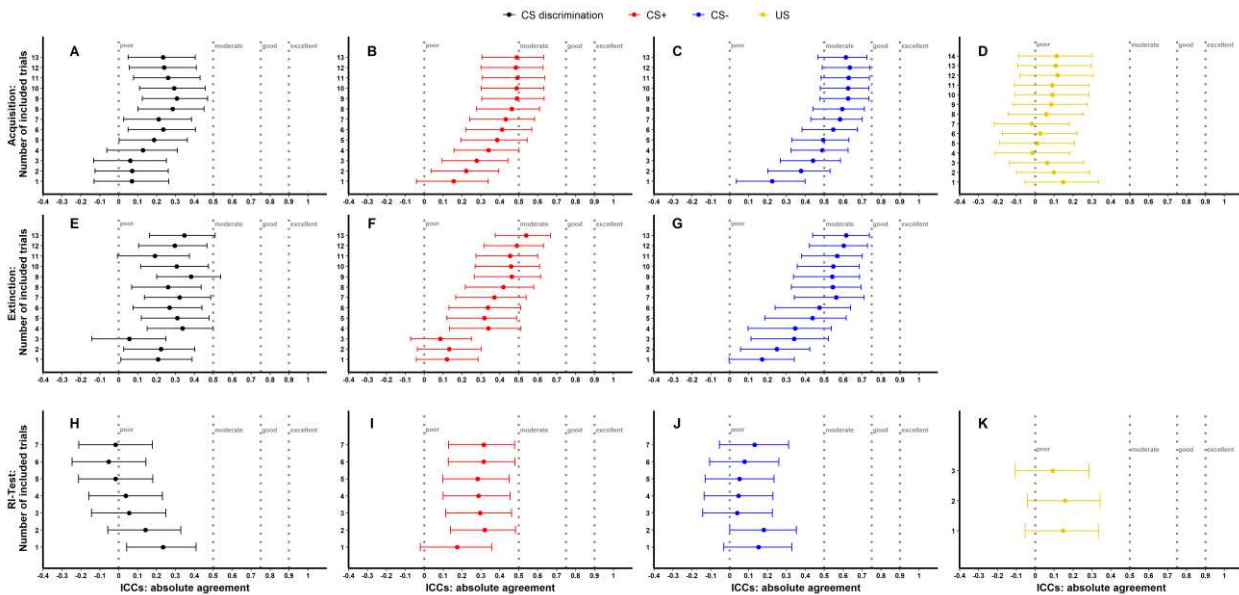


Figure 1-figure supplement 7. Illustration of ICC_{abs} of trial-by-trial log-transformed and range corrected SCRs for phases (A-D: Acquisition, E-G: Extinction, H-J: Reinstatement-Test, K: Reinstatement) and stimulus types separately. Trials were averaged starting with the first (i.e., reinstatement-test and US trials) or second trial (i.e., acquisition and extinction training). All preceding trials were added trial-by-trial and averaged. ICCs < 0.5, < 0.75, < 0.9 and > 0.9 (Koo & Li, 2016) were interpreted as poor, moderate, good and excellent respectively. Error bars represent 95% confidence intervals. Non-overlapping error bars indicate significant differences between ICCs within one figure. RI-Test = reinstatement-test.

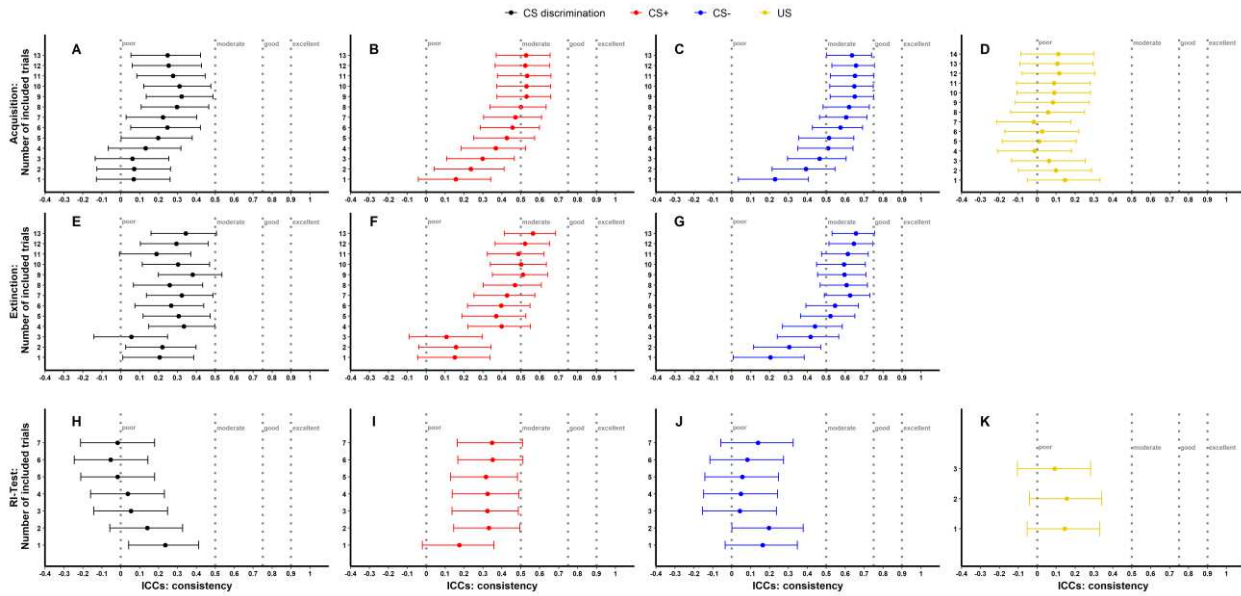


Figure 1-figure supplement 8. Illustration of ICC_{con} of trial-by-trial log-transformed and range corrected SCRs for phases (A-D: Acquisition, E-G: Extinction, H-J: Reinstatement-Test, K: Reinstatement) and stimulus types separately. Trials were averaged starting with the first (i.e., reinstatement-test and US trials) or second trial (i.e., acquisition and extinction training). All preceding trials were added trial-by-trial and averaged. ICCs < 0.5 , < 0.75 , < 0.9 and > 0.9 (Koo & Li, 2016) were interpreted as poor, moderate, good and excellent respectively. Error bars represent 95% confidence intervals. Non-overlapping error bars indicate significant differences between ICCs within one figure. RI-Test = reinstatement-test.

Detailed results of ICC calculations: fear ratings*Supplementary file 4: ICC_{abs} and ICC_{con} for all data specifications of fear ratings.*

Outcome	Stim.-type	Phase	Op.	ICC _{abs}				ICC _{con}			
				Value	Lower 95% CI	Upper 95% CI	p-value	Value	Lower 95% CI	Upper 95% CI	p-value
Ratings	CS dis.	Acq	post-pre	0.190	0.007	0.364	.043	0.203	0.008	0.384	.043
Ratings	CS+	Acq	post-pre	0.436	0.262	0.582	< .001	0.433	0.260	0.579	< .001
Ratings	CS-	Acq	post-pre	-0.162	-0.328	0.018	.945	-0.190	-0.372	0.005	.945
Ratings	CS dis.	Acq	post	0.424	0.230	0.581	< .001	0.470	0.302	0.609	< .001
Ratings	CS+	Acq	post	0.343	0.163	0.502	.001	0.362	0.179	0.521	.001
Ratings	CS-	Acq	post	0.228	0.045	0.400	.020	0.242	0.049	0.417	.020
Ratings	US	Acq	post	0.310	0.120	0.470	.005	0.300	0.110	0.470	.005
Ratings	CS dis.	Ext	pre	0.459	0.250	0.617	< .001	0.516	0.357	0.646	< .001
Ratings	CS+	Ext	pre	0.485	0.266	0.643	< .001	0.548	0.395	0.671	< .001
Ratings	CS-	Ext	pre	0.702	0.587	0.789	< .001	0.700	0.585	0.788	< .001
Ratings	CS dis.	Ext	pre-post	0.482	0.308	0.623	< .001	0.512	0.352	0.643	< .001
Ratings	CS+	Ext	pre-post	0.494	0.282	0.648	< .001	0.552	0.400	0.675	< .001
Ratings	CS-	Ext	pre-post	0.188	0.003	0.363	.048	0.198	0.002	0.378	.048
Ratings	CS dis.	Ext	post	0.165	-0.025	0.345	.078	0.169	-0.027	0.353	.078
Ratings	CS+	Ext	post	0.474	0.307	0.613	< .001	0.472	0.305	0.611	< .001
Ratings	CS-	Ext	post	0.686	0.563	0.779	< .001	0.700	0.584	0.787	< .001
Ratings	US	RI	post	0.430	0.260	0.580	< .001	0.450	0.280	0.590	< .001
Ratings	CS dis.	RI-T	pre	0.172	-0.021	0.354	.072	0.174	-0.022	0.357	.072
Ratings	CS+	RI-T	pre	0.437	0.264	0.582	< .001	0.436	0.263	0.582	< .001
Ratings	CS-	RI-T	pre	0.538	0.382	0.663	< .001	0.536	0.381	0.662	< .001

Note. Stim. = Stimulus, Op. = Operationalization, CI = Confidence Interval, CS dis. = CS discrimination, Acq = Acquisition training, Ext = Extinction training, RI = Reinstatement, RI-T = Reinstatement-Test, pre = prior to the experimental phase, post = subsequent to the experimental phase.

Detailed results of ICC and similarity calculations: BOLD fMRI

Supplementary File 5: ICC_{abs} and ICC_{con} for CS discrimination during fear acquisition (Acq) and extinction training (Ext).

Phase	ICC-type	Whole Brain	Anterior Insula	Amygdala	Hippocampus	Caudate Nucleus	Putamen	Pallidum	NAcc	Thalamus	dACC	dIPFC	vmPFC
Acq	ICCabs	0.175	0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.001
	ICCcon	0.175	0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.001
Ext	ICCabs	0.008	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	ICCcon	0.008	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Note. NAcc = nucleus accumbens; dACC = dorsal anterior cingulate cortex; dIPFC = dorsolateral prefrontal cortex; vmPFC = ventromedial prefrontal cortex.

Supplementary File 6: Paired sample t-tests comparing between- and within-subject similarity for whole brain activation pattern as well as activation pattern in the ROIs for acquisition training (Acq) and extinction training (Ext).

Phase	ROI	<i>t</i>	<i>df</i>	<i>p</i>	<i>Cohen's d</i>
Acq	Whole Brain	4.09	70	< .001	0.49
	Anterior Insula	4.33	70	< .001	0.51
	Amygdala	2.01	70	.048	0.24
	Hippocampus	2.18	70	.033	0.26
	Caudate Nucleus	2.27	70	.026	0.27
	Putamen	2.42	70	.018	0.29
	Pallidum	1.84	70	.070	0.22
	NAcc	-0.18	70	.857	-0.02
	Thalamus	3.20	70	.002	0.38
	dACC	3.75	70	< .001	0.44
	dIPFC	4.71	70	< .001	0.56
vmPFC	2.39	70	.019	0.28	
Ext	Whole Brain	1.44	70	.154	0.17
	Anterior Insula	0.63	70	.531	0.07
	Amygdala	-0.35	70	.726	-0.04
	Hippocampus	0.89	70	.379	0.11
	Caudate Nucleus	-0.65	70	.520	-0.08
	Putamen	-0.63	70	.528	-0.08
	Pallidum	0.34	70	.733	0.04
	NAcc	1.03	70	.306	0.12
	Thalamus	-0.84	70	.401	-0.10
	dACC	0.05	70	.956	0.01

Phase	ROI	<i>t</i>	<i>df</i>	<i>p</i>	<i>Cohen's d</i>
	dIPFC	-0.39	70	.697	-0.05
	vmPFC	-0.06	70	.955	-0.01

Note. NAcc = nucleus accumbens; dACC = dorsal anterior cingulate cortex; dIPFC = dorsolateral prefrontal cortex; vmPFC = ventromedial prefrontal cortex.

Longitudinal reliability at the group-level: log-transformed as well as log-transformed and range corrected SCRs

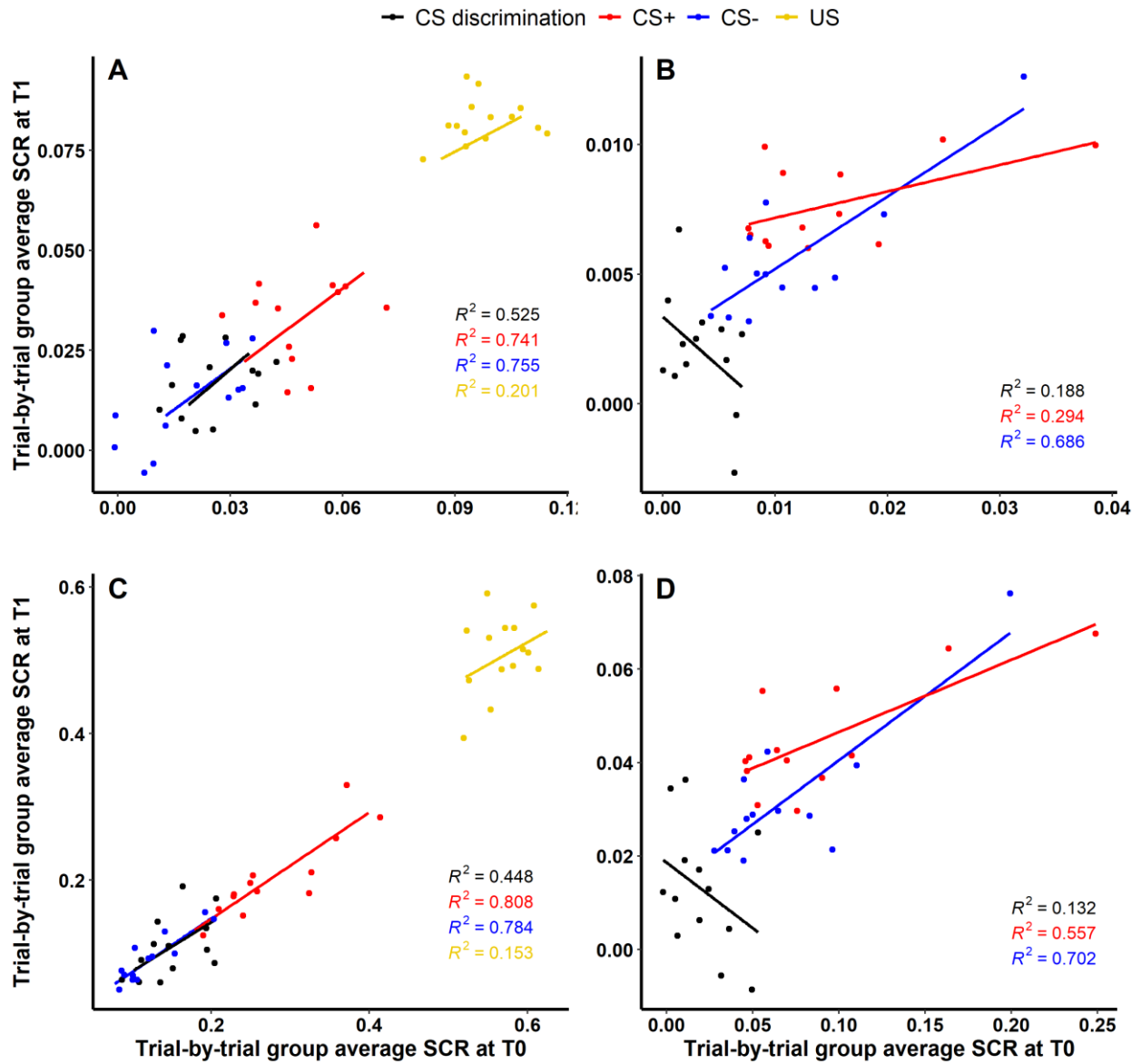


Figure 4-figure supplement 1. Scatter plots illustrating longitudinal reliability at the group level during (A,C) acquisition and (B,D) extinction training for log-transformed (A,B) as well as log-transformed and range corrected (C,D) SCRs. Longitudinal reliability at the group level refers to

the explained variance in linear regressions comprising SCRs at T0 as independent and SCRs at T1 as dependent variable. Results are shown for trial-by-trial group average SCRs to the CS+ (red), CS- (blue), the US (yellow) and CS discrimination (black). Single data points represent pairs of single trials at T0 and T1 averaged across participants. Note that no US was presented during extinction training and hence, no reliability of the US is shown in (B) and (D).

Detailed results of predictability analysis: SCR and fear ratings

Cohen's f^2 (formula: $f^2 = R^2/1 - R^2$) was calculated as effect size. According to the guidelines of Cohen (1988), $f^2 \geq .02$, $f^2 \geq .15$ and $f^2 \geq .34$ represent small, medium and large effect sizes respectively. Since Cohen's f^2 is informative, but less common (Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012), additionally R squared is reported as effect size.

Supplementary File 7: Detailed results of linear regressions: SCR.

Outcome	Stim.-type	Ampl.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
SCR	CS dis.	raw	not ranked	AVE ACQ	1st trial EXT	0.329	0.129	0.076	0.582	2.543	105	0.012	0.038	0.040
SCR	CS dis.	raw	not ranked	AVE last 2 trials ACQ	1st trial EXT	0.264	0.080	0.107	0.421	3.288	105	0.001	0.066	0.071
SCR	CS dis.	raw	not ranked	AVE ACQ	AVE EXT	0.109	0.062	-0.013	0.231	1.762	105	0.081	0.050	0.052
SCR	CS dis.	raw	not ranked	AVE last 2 trials ACQ	AVE EXT	0.031	0.031	-0.030	0.092	0.986	105	0.327	0.011	0.011
SCR	CS dis.	raw	not ranked	AVE ACQ	AVE last 2 trials EXT	0.081	0.115	-0.144	0.306	0.705	105	0.483	0.007	0.007
SCR	CS dis.	raw	not ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	-0.039	0.105	-0.245	0.167	-0.371	105	0.711	0.005	0.005
SCR	CS dis.	raw	not ranked	AVE ACQ	1st trial RI-Test	0.195	0.276	-0.346	0.736	0.708	105	0.480	0.008	0.008
SCR	CS dis.	raw	not ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.218	0.230	-0.233	0.669	0.945	105	0.347	0.028	0.029
SCR	CS dis.	raw	not ranked	1st trial EXT	1st trial RI-Test	0.038	0.165	-0.285	0.361	0.231	105	0.817	0.001	0.001
SCR	CS dis.	raw	not ranked	AVE EXT	1st trial RI-Test	0.222	0.501	-0.760	1.204	0.443	105	0.659	0.003	0.003
SCR	CS dis.	raw	not ranked	AVE last 2 trials EXT	1st trial RI-Test	-0.316	0.824	-1.931	1.299	-0.384	105	0.702	0.020	0.020
SCR	CS+	raw	not ranked	AVE ACQ	1st trial EXT	0.686	0.128	0.435	0.937	5.347	105	0.000	0.291	0.410
SCR	CS+	raw	not ranked	AVE last 2 trials ACQ	1st trial EXT	0.508	0.130	0.253	0.763	3.909	105	0.000	0.212	0.269
SCR	CS+	raw	not ranked	AVE ACQ	AVE EXT	0.283	0.076	0.134	0.432	3.705	105	0.000	0.273	0.375
SCR	CS+	raw	not ranked	AVE last 2 trials ACQ	AVE EXT	0.216	0.077	0.065	0.367	2.827	105	0.006	0.212	0.270
SCR	CS+	raw	not ranked	AVE ACQ	AVE last 2 trials EXT	0.200	0.099	0.006	0.394	2.006	105	0.047	0.120	0.137
SCR	CS+	raw	not ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.143	0.092	-0.037	0.323	1.550	105	0.124	0.082	0.089

Outcome	Stim.-type	Ampl.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
SCR	CS+	raw	not ranked	AVE ACQ	1st trial RI-Test	0.676	0.132	0.417	0.935	5.099	105	0.000	0.213	0.270
SCR	CS+	raw	not ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.434	0.147	0.146	0.722	2.956	105	0.004	0.117	0.132
SCR	CS+	raw	not ranked	1st trial EXT	1st trial RI-Test	0.608	0.101	0.410	0.806	5.993	105	0.000	0.279	0.386
SCR	CS+	raw	not ranked	AVE EXT	1st trial RI-Test	1.123	0.250	0.633	1.613	4.486	105	0.000	0.172	0.208
SCR	CS+	raw	not ranked	AVE last 2 trials EXT	1st trial RI-Test	0.386	0.238	-0.080	0.852	1.627	105	0.107	0.023	0.024
SCR	CS-	raw	not ranked	AVE ACQ	1st trial EXT	0.728	0.193	0.350	1.106	3.774	105	0.000	0.132	0.152
SCR	CS-	raw	not ranked	AVE last 2 trials ACQ	1st trial EXT	0.520	0.150	0.226	0.814	3.469	105	0.001	0.086	0.094
SCR	CS-	raw	not ranked	AVE ACQ	AVE EXT	0.369	0.082	0.208	0.530	4.518	105	0.000	0.280	0.390
SCR	CS-	raw	not ranked	AVE last 2 trials ACQ	AVE EXT	0.231	0.074	0.086	0.376	3.126	105	0.002	0.140	0.163
SCR	CS-	raw	not ranked	AVE ACQ	AVE last 2 trials EXT	0.370	0.148	0.080	0.660	2.495	105	0.014	0.165	0.197
SCR	CS-	raw	not ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.265	0.107	0.055	0.475	2.475	105	0.015	0.107	0.120
SCR	CS-	raw	not ranked	AVE ACQ	1st trial RI-Test	0.640	0.240	0.170	1.110	2.661	105	0.009	0.086	0.094
SCR	CS-	raw	not ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.449	0.239	-0.019	0.917	1.878	105	0.063	0.054	0.057
SCR	CS-	raw	not ranked	1st trial EXT	1st trial RI-Test	0.336	0.118	0.105	0.567	2.849	105	0.005	0.096	0.106
SCR	CS-	raw	not ranked	AVE EXT	1st trial RI-Test	0.584	0.299	-0.002	1.170	1.953	105	0.054	0.035	0.036
SCR	CS-	raw	not ranked	AVE last 2 trials EXT	1st trial RI-Test	0.145	0.319	-0.480	0.770	0.453	105	0.651	0.004	0.004
SCR	CS dis.	log	not ranked	AVE ACQ	1st trial EXT	0.328	0.134	0.065	0.591	2.446	105	0.016	0.037	0.039
SCR	CS dis.	log	not ranked	AVE last 2 trials ACQ	1st trial EXT	0.260	0.085	0.093	0.427	3.045	105	0.003	0.062	0.066
SCR	CS dis.	log	not ranked	AVE ACQ	AVE EXT	0.109	0.056	-0.001	0.219	1.956	105	0.053	0.047	0.050
SCR	CS dis.	log	not ranked	AVE last 2 trials ACQ	AVE EXT	0.031	0.031	-0.030	0.092	1.000	105	0.320	0.010	0.010
SCR	CS dis.	log	not ranked	AVE ACQ	AVE last 2 trials EXT	0.074	0.103	-0.128	0.276	0.719	105	0.474	0.006	0.006
SCR	CS dis.	log	not ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	-0.039	0.105	-0.245	0.167	-0.373	105	0.710	0.004	0.004
SCR	CS dis.	log	not ranked	AVE ACQ	1st trial RI-Test	0.135	0.259	-0.373	0.643	0.521	105	0.603	0.004	0.004
SCR	CS dis.	log	not ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.173	0.221	-0.260	0.606	0.784	105	0.435	0.018	0.018
SCR	CS dis.	log	not ranked	1st trial EXT	1st trial RI-Test	0.043	0.149	-0.249	0.335	0.291	105	0.771	0.001	0.001

Outcome	Stim.-type	Ampl.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
SCR	CS dis.	log	not ranked	AVE EXT	1st trial RI-Test	0.149	0.450	-0.733	1.031	0.331	105	0.741	0.001	0.001
SCR	CS dis.	log	not ranked	AVE last 2 trials EXT	1st trial RI-Test	-0.282	0.685	-1.625	1.061	-0.412	105	0.681	0.017	0.017
SCR	CS+	log	not ranked	AVE ACQ	1st trial EXT	0.679	0.115	0.454	0.904	5.906	105	0.000	0.297	0.423
SCR	CS+	log	not ranked	AVE last 2 trials ACQ	1st trial EXT	0.502	0.117	0.273	0.731	4.305	105	0.000	0.210	0.265
SCR	CS+	log	not ranked	AVE ACQ	AVE EXT	0.294	0.074	0.149	0.439	3.995	105	0.000	0.277	0.383
SCR	CS+	log	not ranked	AVE last 2 trials ACQ	AVE EXT	0.232	0.074	0.087	0.377	3.145	105	0.002	0.223	0.287
SCR	CS+	log	not ranked	AVE ACQ	AVE last 2 trials EXT	0.202	0.094	0.018	0.386	2.158	105	0.033	0.117	0.133
SCR	CS+	log	not ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.149	0.088	-0.023	0.321	1.686	105	0.095	0.082	0.089
SCR	CS+	log	not ranked	AVE ACQ	1st trial RI-Test	0.659	0.123	0.418	0.900	5.361	105	0.000	0.216	0.275
SCR	CS+	log	not ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.418	0.135	0.153	0.683	3.106	105	0.002	0.112	0.126
SCR	CS+	log	not ranked	1st trial EXT	1st trial RI-Test	0.603	0.096	0.415	0.791	6.255	105	0.000	0.280	0.390
SCR	CS+	log	not ranked	AVE EXT	1st trial RI-Test	1.032	0.219	0.603	1.461	4.706	105	0.000	0.165	0.198
SCR	CS+	log	not ranked	AVE last 2 trials EXT	1st trial RI-Test	0.364	0.214	-0.055	0.783	1.701	105	0.092	0.023	0.024
SCR	CS-	log	not ranked	AVE ACQ	1st trial EXT	0.712	0.183	0.353	1.071	3.904	105	0.000	0.133	0.154
SCR	CS-	log	not ranked	AVE last 2 trials ACQ	1st trial EXT	0.518	0.146	0.232	0.804	3.546	105	0.001	0.089	0.098
SCR	CS-	log	not ranked	AVE ACQ	AVE EXT	0.384	0.081	0.225	0.543	4.729	105	0.000	0.295	0.418
SCR	CS-	log	not ranked	AVE last 2 trials ACQ	AVE EXT	0.245	0.073	0.102	0.388	3.356	105	0.001	0.152	0.179
SCR	CS-	log	not ranked	AVE ACQ	AVE last 2 trials EXT	0.379	0.145	0.095	0.663	2.608	105	0.010	0.175	0.213
SCR	CS-	log	not ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.285	0.106	0.077	0.493	2.687	105	0.008	0.126	0.144
SCR	CS-	log	not ranked	AVE ACQ	1st trial RI-Test	0.612	0.216	0.189	1.035	2.833	105	0.006	0.086	0.094
SCR	CS-	log	not ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.449	0.213	0.032	0.866	2.108	105	0.037	0.058	0.062
SCR	CS-	log	not ranked	1st trial EXT	1st trial RI-Test	0.341	0.113	0.120	0.562	3.021	105	0.003	0.101	0.113
SCR	CS-	log	not ranked	AVE EXT	1st trial RI-Test	0.578	0.280	0.029	1.127	2.066	105	0.041	0.038	0.040
SCR	CS-	log	not ranked	AVE last 2 trials EXT	1st trial RI-Test	0.163	0.287	-0.400	0.726	0.568	105	0.571	0.005	0.005
SCR	CS dis.	log rc	not ranked	AVE ACQ	1st trial EXT	0.378	0.200	-0.014	0.770	1.894	105	0.061	0.030	0.031

Outcome	Stim.-type	Ampl.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
SCR	CS dis.	log rc	not ranked	AVE last 2 trials ACQ	1st trial EXT	0.298	0.109	0.084	0.512	2.733	105	0.007	0.046	0.049
SCR	CS dis.	log rc	not ranked	AVE ACQ	AVE EXT	0.071	0.054	-0.035	0.177	1.328	105	0.187	0.017	0.017
SCR	CS dis.	log rc	not ranked	AVE last 2 trials ACQ	AVE EXT	0.022	0.038	-0.052	0.096	0.572	105	0.568	0.004	0.004
SCR	CS dis.	log rc	not ranked	AVE ACQ	AVE last 2 trials EXT	0.103	0.106	-0.105	0.311	0.974	105	0.332	0.010	0.010
SCR	CS dis.	log rc	not ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	-0.007	0.093	-0.189	0.175	-0.073	105	0.942	0.000	0.000
SCR	CS dis.	log rc	not ranked	AVE ACQ	1st trial RI-Test	-0.212	0.225	-0.653	0.229	-0.943	105	0.348	0.009	0.009
SCR	CS dis.	log rc	not ranked	AVE last 2 trials ACQ	1st trial RI-Test	-0.041	0.158	-0.351	0.269	-0.259	105	0.796	0.001	0.001
SCR	CS dis.	log rc	not ranked	1st trial EXT	1st trial RI-Test	0.032	0.112	-0.188	0.252	0.290	105	0.773	0.001	0.001
SCR	CS dis.	log rc	not ranked	AVE EXT	1st trial RI-Test	-0.025	0.381	-0.772	0.722	-0.066	105	0.948	0.000	0.000
SCR	CS dis.	log rc	not ranked	AVE last 2 trials EXT	1st trial RI-Test	-0.360	0.376	-1.097	0.377	-0.958	105	0.340	0.027	0.028
SCR	CS+	log rc	not ranked	AVE ACQ	1st trial EXT	0.435	0.122	0.196	0.674	3.564	105	0.001	0.108	0.121
SCR	CS+	log rc	not ranked	AVE last 2 trials ACQ	1st trial EXT	0.320	0.111	0.102	0.538	2.886	105	0.005	0.069	0.074
SCR	CS+	log rc	not ranked	AVE ACQ	AVE EXT	0.267	0.058	0.153	0.381	4.578	105	0.000	0.215	0.274
SCR	CS+	log rc	not ranked	AVE last 2 trials ACQ	AVE EXT	0.239	0.059	0.123	0.355	4.069	105	0.000	0.204	0.256
SCR	CS+	log rc	not ranked	AVE ACQ	AVE last 2 trials EXT	0.181	0.073	0.038	0.324	2.495	105	0.014	0.079	0.086
SCR	CS+	log rc	not ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.157	0.069	0.022	0.292	2.268	105	0.025	0.070	0.075
SCR	CS+	log rc	not ranked	AVE ACQ	1st trial RI-Test	0.233	0.138	-0.037	0.503	1.681	105	0.096	0.023	0.024
SCR	CS+	log rc	not ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.070	0.130	-0.185	0.325	0.543	105	0.588	0.003	0.003
SCR	CS+	log rc	not ranked	1st trial EXT	1st trial RI-Test	0.403	0.103	0.201	0.605	3.910	105	0.000	0.122	0.138
SCR	CS+	log rc	not ranked	AVE EXT	1st trial RI-Test	0.500	0.206	0.096	0.904	2.425	105	0.017	0.035	0.037
SCR	CS+	log rc	not ranked	AVE last 2 trials EXT	1st trial RI-Test	-0.021	0.196	-0.405	0.363	-0.106	105	0.916	0.000	0.000
SCR	CS-	log rc	not ranked	AVE ACQ	1st trial EXT	0.247	0.193	-0.131	0.625	1.278	105	0.204	0.014	0.014
SCR	CS-	log rc	not ranked	AVE last 2 trials ACQ	1st trial EXT	0.192	0.150	-0.102	0.486	1.275	105	0.205	0.010	0.011
SCR	CS-	log rc	not ranked	AVE ACQ	AVE EXT	0.370	0.078	0.217	0.523	4.719	105	0.000	0.272	0.375
SCR	CS-	log rc	not ranked	AVE last 2 trials ACQ	AVE EXT	0.246	0.069	0.111	0.381	3.577	105	0.001	0.151	0.178

Outcome	Stim.-type	Ampl.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
SCR	CS-	log rc	not ranked	AVE ACQ	AVE last 2 trials EXT	0.310	0.104	0.106	0.514	2.971	105	0.004	0.125	0.143
SCR	CS-	log rc	not ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.246	0.078	0.093	0.399	3.143	105	0.002	0.099	0.110
SCR	CS-	log rc	not ranked	AVE ACQ	1st trial RI-Test	0.397	0.240	-0.073	0.867	1.654	105	0.101	0.031	0.032
SCR	CS-	log rc	not ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.255	0.216	-0.168	0.678	1.179	105	0.241	0.016	0.017
SCR	CS-	log rc	not ranked	1st trial EXT	1st trial RI-Test	0.192	0.118	-0.039	0.423	1.619	105	0.108	0.032	0.033
SCR	CS-	log rc	not ranked	AVE EXT	1st trial RI-Test	0.178	0.278	-0.367	0.723	0.639	105	0.524	0.003	0.003
SCR	CS-	log rc	not ranked	AVE last 2 trials EXT	1st trial RI-Test	-0.108	0.189	-0.478	0.262	-0.569	105	0.571	0.002	0.002
SCR	CS dis.	raw	ranked	AVE ACQ	1st trial EXT	0.180	0.089	0.006	0.354	2.009	105	0.047	0.032	0.033
SCR	CS dis.	raw	ranked	AVE last 2 trials ACQ	1st trial EXT	0.273	0.086	0.104	0.442	3.167	105	0.002	0.087	0.095
SCR	CS dis.	raw	ranked	AVE ACQ	AVE EXT	0.211	0.097	0.021	0.401	2.169	105	0.032	0.043	0.045
SCR	CS dis.	raw	ranked	AVE last 2 trials ACQ	AVE EXT	0.228	0.092	0.048	0.408	2.489	105	0.014	0.059	0.063
SCR	CS dis.	raw	ranked	AVE ACQ	AVE last 2 trials EXT	0.125	0.120	-0.110	0.360	1.045	105	0.299	0.012	0.012
SCR	CS dis.	raw	ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.200	0.106	-0.008	0.408	1.888	105	0.062	0.036	0.038
SCR	CS dis.	raw	ranked	AVE ACQ	1st trial RI-Test	0.037	0.103	-0.165	0.239	0.362	105	0.718	0.001	0.001
SCR	CS dis.	raw	ranked	AVE last 2 trials ACQ	1st trial RI-Test	-0.071	0.096	-0.259	0.117	-0.740	105	0.461	0.006	0.006
SCR	CS dis.	raw	ranked	1st trial EXT	1st trial RI-Test	0.034	0.112	-0.186	0.254	0.303	105	0.763	0.001	0.001
SCR	CS dis.	raw	ranked	AVE EXT	1st trial RI-Test	0.017	0.109	-0.197	0.231	0.154	105	0.878	0.000	0.000
SCR	CS dis.	raw	ranked	AVE last 2 trials EXT	1st trial RI-Test	0.068	0.087	-0.103	0.239	0.773	105	0.442	0.006	0.006
SCR	CS+	raw	ranked	AVE ACQ	1st trial EXT	0.594	0.075	0.447	0.741	7.958	105	0.000	0.319	0.469
SCR	CS+	raw	ranked	AVE last 2 trials ACQ	1st trial EXT	0.381	0.078	0.228	0.534	4.860	105	0.000	0.187	0.230
SCR	CS+	raw	ranked	AVE ACQ	AVE EXT	0.607	0.071	0.468	0.746	8.500	105	0.000	0.324	0.480
SCR	CS+	raw	ranked	AVE last 2 trials ACQ	AVE EXT	0.451	0.077	0.300	0.602	5.852	105	0.000	0.256	0.343
SCR	CS+	raw	ranked	AVE ACQ	AVE last 2 trials EXT	0.364	0.125	0.119	0.609	2.912	105	0.004	0.072	0.078
SCR	CS+	raw	ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.281	0.108	0.069	0.493	2.608	105	0.010	0.061	0.065
SCR	CS+	raw	ranked	AVE ACQ	1st trial RI-Test	0.485	0.083	0.322	0.648	5.828	105	0.000	0.215	0.274
SCR	CS+	raw	ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.216	0.088	0.044	0.388	2.441	105	0.016	0.061	0.064

Outcome	Stim.- type	Ampl.- type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
SCR	CS+	raw	ranked	1st trial EXT	1st trial RI-Test	0.518	0.083	0.355	0.681	6.282	105	0.000	0.272	0.374
SCR	CS+	raw	ranked	AVE EXT	1st trial RI-Test	0.340	0.097	0.150	0.530	3.507	105	0.001	0.120	0.136
SCR	CS+	raw	ranked	AVE last 2 trials EXT	1st trial RI-Test	0.009	0.075	-0.138	0.156	0.113	105	0.910	0.000	0.000
SCR	CS-	raw	ranked	AVE ACQ	1st trial EXT	0.388	0.096	0.200	0.576	4.057	105	0.000	0.129	0.148
SCR	CS-	raw	ranked	AVE last 2 trials ACQ	1st trial EXT	0.196	0.078	0.043	0.349	2.507	105	0.014	0.056	0.060
SCR	CS-	raw	ranked	AVE ACQ	AVE EXT	0.670	0.070	0.533	0.807	9.586	105	0.000	0.384	0.623
SCR	CS-	raw	ranked	AVE last 2 trials ACQ	AVE EXT	0.353	0.072	0.212	0.494	4.905	105	0.000	0.184	0.225
SCR	CS-	raw	ranked	AVE ACQ	AVE last 2 trials EXT	0.427	0.115	0.202	0.652	3.702	105	0.000	0.109	0.122
SCR	CS-	raw	ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.388	0.094	0.204	0.572	4.117	105	0.000	0.155	0.183
SCR	CS-	raw	ranked	AVE ACQ	1st trial RI-Test	0.340	0.104	0.136	0.544	3.281	105	0.001	0.099	0.110
SCR	CS-	raw	ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.206	0.080	0.049	0.363	2.567	105	0.012	0.062	0.067
SCR	CS-	raw	ranked	1st trial EXT	1st trial RI-Test	0.327	0.100	0.131	0.523	3.265	105	0.001	0.107	0.119
SCR	CS-	raw	ranked	AVE EXT	1st trial RI-Test	0.298	0.096	0.110	0.486	3.096	105	0.003	0.089	0.097
SCR	CS-	raw	ranked	AVE last 2 trials EXT	1st trial RI-Test	0.110	0.069	-0.025	0.245	1.583	105	0.117	0.017	0.018
SCR	CS dis.	log	ranked	AVE ACQ	1st trial EXT	0.177	0.090	0.001	0.353	1.971	105	0.051	0.031	0.032
SCR	CS dis.	log	ranked	AVE last 2 trials ACQ	1st trial EXT	0.269	0.086	0.100	0.438	3.135	105	0.002	0.084	0.092
SCR	CS dis.	log	ranked	AVE ACQ	AVE EXT	0.206	0.098	0.014	0.398	2.108	105	0.037	0.041	0.043
SCR	CS dis.	log	ranked	AVE last 2 trials ACQ	AVE EXT	0.214	0.092	0.034	0.394	2.319	105	0.022	0.052	0.055
SCR	CS dis.	log	ranked	AVE ACQ	AVE last 2 trials EXT	0.132	0.120	-0.103	0.367	1.102	105	0.273	0.014	0.014
SCR	CS dis.	log	ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.194	0.106	-0.014	0.402	1.838	105	0.069	0.034	0.036
SCR	CS dis.	log	ranked	AVE ACQ	1st trial RI-Test	0.039	0.103	-0.163	0.241	0.380	105	0.704	0.002	0.002
SCR	CS dis.	log	ranked	AVE last 2 trials ACQ	1st trial RI-Test	-0.081	0.096	-0.269	0.107	-0.845	105	0.400	0.008	0.008
SCR	CS dis.	log	ranked	1st trial EXT	1st trial RI-Test	0.030	0.110	-0.186	0.246	0.270	105	0.787	0.001	0.001
SCR	CS dis.	log	ranked	AVE EXT	1st trial RI-Test	0.009	0.109	-0.205	0.223	0.084	105	0.933	0.000	0.000
SCR	CS dis.	log	ranked	AVE last 2 trials EXT	1st trial RI-Test	0.060	0.087	-0.111	0.231	0.696	105	0.488	0.005	0.005
SCR	CS+	log	ranked	AVE ACQ	1st trial EXT	0.591	0.075	0.444	0.738	7.906	105	0.000	0.316	0.462

Outcome	Stim.-type	Ampl.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
SCR	CS+	log	ranked	AVE last 2 trials ACQ	1st trial EXT	0.382	0.078	0.229	0.535	4.880	105	0.000	0.188	0.231
SCR	CS+	log	ranked	AVE ACQ	AVE EXT	0.606	0.073	0.463	0.749	8.363	105	0.000	0.324	0.479
SCR	CS+	log	ranked	AVE last 2 trials ACQ	AVE EXT	0.455	0.077	0.304	0.606	5.942	105	0.000	0.260	0.351
SCR	CS+	log	ranked	AVE ACQ	AVE last 2 trials EXT	0.375	0.125	0.130	0.620	3.003	105	0.003	0.077	0.083
SCR	CS+	log	ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.285	0.108	0.073	0.497	2.646	105	0.009	0.063	0.067
SCR	CS+	log	ranked	AVE ACQ	1st trial RI-Test	0.484	0.084	0.319	0.649	5.789	105	0.000	0.214	0.272
SCR	CS+	log	ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.221	0.088	0.049	0.393	2.514	105	0.013	0.063	0.068
SCR	CS+	log	ranked	1st trial EXT	1st trial RI-Test	0.518	0.083	0.355	0.681	6.282	105	0.000	0.272	0.374
SCR	CS+	log	ranked	AVE EXT	1st trial RI-Test	0.337	0.097	0.147	0.527	3.488	105	0.001	0.118	0.134
SCR	CS+	log	ranked	AVE last 2 trials EXT	1st trial RI-Test	0.008	0.075	-0.139	0.155	0.110	105	0.912	0.000	0.000
SCR	CS-	log	ranked	AVE ACQ	1st trial EXT	0.387	0.095	0.201	0.573	4.057	105	0.000	0.128	0.147
SCR	CS-	log	ranked	AVE last 2 trials ACQ	1st trial EXT	0.197	0.078	0.044	0.350	2.520	105	0.013	0.057	0.060
SCR	CS-	log	ranked	AVE ACQ	AVE EXT	0.674	0.070	0.537	0.811	9.688	105	0.000	0.388	0.634
SCR	CS-	log	ranked	AVE last 2 trials ACQ	AVE EXT	0.356	0.072	0.215	0.497	4.959	105	0.000	0.187	0.230
SCR	CS-	log	ranked	AVE ACQ	AVE last 2 trials EXT	0.432	0.115	0.207	0.657	3.751	105	0.000	0.111	0.125
SCR	CS-	log	ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.391	0.094	0.207	0.575	4.151	105	0.000	0.157	0.186
SCR	CS-	log	ranked	AVE ACQ	1st trial RI-Test	0.341	0.104	0.137	0.545	3.289	105	0.001	0.099	0.110
SCR	CS-	log	ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.206	0.080	0.049	0.363	2.573	105	0.011	0.063	0.067
SCR	CS-	log	ranked	1st trial EXT	1st trial RI-Test	0.327	0.100	0.131	0.523	3.265	105	0.001	0.107	0.119
SCR	CS-	log	ranked	AVE EXT	1st trial RI-Test	0.298	0.096	0.110	0.486	3.108	105	0.002	0.089	0.097
SCR	CS-	log	ranked	AVE last 2 trials EXT	1st trial RI-Test	0.109	0.069	-0.026	0.244	1.578	105	0.118	0.017	0.017
SCR	CS dis.	log rc	ranked	AVE ACQ	1st trial EXT	0.150	0.096	-0.038	0.338	1.571	105	0.119	0.023	0.023
SCR	CS dis.	log rc	ranked	AVE last 2 trials ACQ	1st trial EXT	0.248	0.085	0.081	0.415	2.930	105	0.004	0.071	0.077
SCR	CS dis.	log rc	ranked	AVE ACQ	AVE EXT	0.136	0.096	-0.052	0.324	1.411	105	0.161	0.018	0.018
SCR	CS dis.	log rc	ranked	AVE last 2 trials ACQ	AVE EXT	0.164	0.094	-0.020	0.348	1.739	105	0.085	0.031	0.032
SCR	CS dis.	log rc	ranked	AVE ACQ	AVE last 2 trials EXT	0.135	0.120	-0.100	0.370	1.130	105	0.261	0.014	0.015

Outcome	Stim.-type	Ampl.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
SCR	CS dis.	log rc	ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.167	0.105	-0.039	0.373	1.594	105	0.114	0.025	0.026
SCR	CS dis.	log rc	ranked	AVE ACQ	1st trial RI-Test	-0.038	0.100	-0.234	0.158	-0.381	105	0.704	0.001	0.001
SCR	CS dis.	log rc	ranked	AVE last 2 trials ACQ	1st trial RI-Test	-0.099	0.093	-0.281	0.083	-1.064	105	0.290	0.011	0.012
SCR	CS dis.	log rc	ranked	1st trial EXT	1st trial RI-Test	0.040	0.100	-0.156	0.236	0.399	105	0.691	0.002	0.002
SCR	CS dis.	log rc	ranked	AVE EXT	1st trial RI-Test	-0.014	0.101	-0.212	0.184	-0.137	105	0.892	0.000	0.000
SCR	CS dis.	log rc	ranked	AVE last 2 trials EXT	1st trial RI-Test	-0.010	0.084	-0.175	0.155	-0.121	105	0.904	0.000	0.000
SCR	CS+	log rc	ranked	AVE ACQ	1st trial EXT	0.358	0.096	0.170	0.546	3.722	105	0.000	0.116	0.131
SCR	CS+	log rc	ranked	AVE last 2 trials ACQ	1st trial EXT	0.244	0.082	0.083	0.405	2.957	105	0.004	0.076	0.083
SCR	CS+	log rc	ranked	AVE ACQ	AVE EXT	0.558	0.089	0.384	0.732	6.264	105	0.000	0.274	0.377
SCR	CS+	log rc	ranked	AVE last 2 trials ACQ	AVE EXT	0.437	0.076	0.288	0.586	5.786	105	0.000	0.240	0.316
SCR	CS+	log rc	ranked	AVE ACQ	AVE last 2 trials EXT	0.397	0.131	0.140	0.654	3.044	105	0.003	0.086	0.094
SCR	CS+	log rc	ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.299	0.110	0.083	0.515	2.729	105	0.007	0.069	0.075
SCR	CS+	log rc	ranked	AVE ACQ	1st trial RI-Test	0.200	0.097	0.010	0.390	2.058	105	0.042	0.037	0.038
SCR	CS+	log rc	ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.074	0.085	-0.093	0.241	0.869	105	0.387	0.007	0.007
SCR	CS+	log rc	ranked	1st trial EXT	1st trial RI-Test	0.349	0.097	0.159	0.539	3.587	105	0.001	0.124	0.141
SCR	CS+	log rc	ranked	AVE EXT	1st trial RI-Test	0.161	0.096	-0.027	0.349	1.681	105	0.096	0.027	0.028
SCR	CS+	log rc	ranked	AVE last 2 trials EXT	1st trial RI-Test	-0.067	0.071	-0.206	0.072	-0.937	105	0.351	0.007	0.008
SCR	CS-	log rc	ranked	AVE ACQ	1st trial EXT	0.244	0.100	0.048	0.440	2.446	105	0.016	0.051	0.053
SCR	CS-	log rc	ranked	AVE last 2 trials ACQ	1st trial EXT	0.111	0.078	-0.042	0.264	1.418	105	0.159	0.018	0.018
SCR	CS-	log rc	ranked	AVE ACQ	AVE EXT	0.682	0.072	0.541	0.823	9.479	105	0.000	0.397	0.659
SCR	CS-	log rc	ranked	AVE last 2 trials ACQ	AVE EXT	0.347	0.071	0.208	0.486	4.913	105	0.000	0.177	0.215
SCR	CS-	log rc	ranked	AVE ACQ	AVE last 2 trials EXT	0.487	0.117	0.258	0.716	4.148	105	0.000	0.141	0.164
SCR	CS-	log rc	ranked	AVE last 2 trials ACQ	AVE last 2 trials EXT	0.383	0.093	0.201	0.565	4.107	105	0.000	0.150	0.177
SCR	CS-	log rc	ranked	AVE ACQ	1st trial RI-Test	0.251	0.097	0.061	0.441	2.582	105	0.011	0.054	0.057
SCR	CS-	log rc	ranked	AVE last 2 trials ACQ	1st trial RI-Test	0.146	0.080	-0.011	0.303	1.815	105	0.072	0.031	0.032
SCR	CS-	log rc	ranked	1st trial EXT	1st trial RI-Test	0.189	0.104	-0.015	0.393	1.822	105	0.071	0.036	0.037

Outcome	Stim.-type	Ampl.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
SCR	CS-	log rc	ranked	AVE EXT	1st trial RI-Test	0.145	0.100	-0.051	0.341	1.454	105	0.149	0.021	0.022
SCR	CS-	log rc	ranked	AVE last 2 trials EXT	1st trial RI-Test	0.039	0.070	-0.098	0.176	0.556	105	0.579	0.002	0.002

Note. Ampl. = Amplitude, Stim. = Stimulus, CI = Confidence Interval, CS dis. = CS discrimination, log = log-transformed, log rc = log-transformed and range corrected, AVE = average, ACQ = Acquisition training, EXT = Extinction training, RI = Reinstatement, RI-Test = Reinstatement-Test.

Supplementary File 8: Detailed results of linear regressions: fear ratings.

Outcome	Stim.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
Fear Ratings	CS dis.	not ranked	post-pre ACQ	pre EXT	0.519	0.094	0.335	0.703	5.543	77	0.000	0.270	0.369
Fear Ratings	CS dis.	not ranked	post ACQ	pre EXT	0.570	0.088	0.398	0.742	6.454	92	0.000	0.265	0.360
Fear Ratings	CS dis.	not ranked	post-pre ACQ	pre-post EXT	0.372	0.132	0.113	0.631	2.827	76	0.006	0.175	0.213
Fear Ratings	CS dis.	not ranked	post ACQ	pre-post EXT	0.413	0.107	0.203	0.623	3.851	91	0.000	0.177	0.215
Fear Ratings	CS dis.	not ranked	post-pre ACQ	post EXT	0.144	0.087	-0.027	0.315	1.648	79	0.103	0.076	0.082
Fear Ratings	CS dis.	not ranked	post ACQ	post EXT	0.139	0.077	-0.012	0.290	1.818	98	0.072	0.063	0.068
Fear Ratings	CS dis.	not ranked	post ACQ	1st trial RI-Test	0.257	0.084	0.092	0.422	3.056	74	0.003	0.099	0.110
Fear Ratings	CS dis.	not ranked	post-pre ACQ	1st trial RI-Test	0.240	0.081	0.081	0.399	2.967	60	0.004	0.117	0.132
Fear Ratings	CS dis.	not ranked	pre EXT	1st trial RI-Test	0.236	0.086	0.067	0.405	2.739	69	0.008	0.112	0.127
Fear Ratings	CS dis.	not ranked	pre-post EXT	1st trial RI-Test	0.187	0.104	-0.017	0.391	1.793	68	0.077	0.052	0.055
Fear Ratings	CS dis.	not ranked	post EXT	1st trial RI-Test	0.301	0.187	-0.066	0.668	1.610	71	0.112	0.043	0.045
Fear Ratings	CS+	not ranked	post-pre ACQ	pre EXT	0.509	0.095	0.323	0.695	5.365	91	0.000	0.208	0.263
Fear Ratings	CS+	not ranked	post ACQ	pre EXT	0.655	0.090	0.479	0.831	7.307	97	0.000	0.319	0.469
Fear Ratings	CS+	not ranked	post-pre ACQ	pre-post EXT	0.425	0.081	0.266	0.584	5.218	90	0.000	0.194	0.241
Fear Ratings	CS+	not ranked	post ACQ	pre-post EXT	0.461	0.083	0.298	0.624	5.547	96	0.000	0.209	0.263
Fear Ratings	CS+	not ranked	post-pre ACQ	post EXT	0.084	0.069	-0.051	0.219	1.212	92	0.229	0.011	0.011
Fear Ratings	CS+	not ranked	post ACQ	post EXT	0.172	0.073	0.029	0.315	2.370	101	0.020	0.042	0.044
Fear Ratings	CS+	not ranked	post ACQ	1st trial RI-Test	0.503	0.102	0.303	0.703	4.928	85	0.000	0.171	0.207
Fear Ratings	CS+	not ranked	post-pre ACQ	1st trial RI-Test	0.330	0.105	0.124	0.536	3.153	79	0.002	0.091	0.100

Outcome	Stim.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	<i>Cohen's f</i> ²
Fear Ratings	CS+	not ranked	pre EXT	1st trial RI-Test	0.430	0.093	0.248	0.612	4.630	82	0.000	0.184	0.226
Fear Ratings	CS+	not ranked	pre-post EXT	1st trial RI-Test	0.293	0.119	0.060	0.526	2.455	81	0.016	0.060	0.064
Fear Ratings	CS+	not ranked	post EXT	1st trial RI-Test	0.352	0.134	0.089	0.615	2.615	84	0.011	0.072	0.077
Fear Ratings	CS-	not ranked	post-pre ACQ	pre EXT	0.077	0.099	-0.117	0.271	0.773	86	0.442	0.019	0.020
Fear Ratings	CS-	not ranked	post ACQ	pre EXT	0.244	0.101	0.046	0.442	2.423	98	0.017	0.130	0.150
Fear Ratings	CS-	not ranked	post-pre ACQ	pre-post EXT	-0.197	0.120	-0.432	0.038	-1.638	85	0.105	0.061	0.065
Fear Ratings	CS-	not ranked	post ACQ	pre-post EXT	-0.147	0.123	-0.388	0.094	-1.198	97	0.234	0.030	0.031
Fear Ratings	CS-	not ranked	post-pre ACQ	post EXT	0.236	0.138	-0.034	0.506	1.708	88	0.091	0.090	0.099
Fear Ratings	CS-	not ranked	post ACQ	post EXT	0.386	0.140	0.112	0.660	2.753	100	0.007	0.217	0.278
Fear Ratings	CS-	not ranked	post ACQ	1st trial RI-Test	0.270	0.176	-0.075	0.615	1.534	88	0.129	0.039	0.040
Fear Ratings	CS-	not ranked	post-pre ACQ	1st trial RI-Test	0.082	0.160	-0.232	0.396	0.513	78	0.610	0.004	0.004
Fear Ratings	CS-	not ranked	pre EXT	1st trial RI-Test	0.493	0.275	-0.046	1.032	1.788	86	0.077	0.043	0.045
Fear Ratings	CS-	not ranked	pre-post EXT	1st trial RI-Test	-0.279	0.213	-0.696	0.138	-1.308	85	0.194	0.028	0.029
Fear Ratings	CS-	not ranked	post EXT	1st trial RI-Test	0.582	0.193	0.204	0.960	3.010	87	0.003	0.109	0.122
Fear Ratings	CS dis.	ranked	post-pre ACQ	pre EXT	0.664	0.112	0.444	0.884	5.936	77	0.000	0.299	0.427
Fear Ratings	CS dis.	ranked	post ACQ	pre EXT	0.534	0.081	0.375	0.693	6.569	92	0.000	0.300	0.428
Fear Ratings	CS dis.	ranked	post-pre ACQ	pre-post EXT	0.595	0.117	0.366	0.824	5.082	76	0.000	0.241	0.317
Fear Ratings	CS dis.	ranked	post ACQ	pre-post EXT	0.461	0.084	0.296	0.626	5.461	91	0.000	0.231	0.301
Fear Ratings	CS dis.	ranked	post-pre ACQ	post EXT	0.269	0.167	-0.058	0.596	1.613	79	0.111	0.033	0.034
Fear Ratings	CS dis.	ranked	post ACQ	post EXT	0.241	0.120	0.006	0.476	2.003	98	0.048	0.040	0.042
Fear Ratings	CS dis.	ranked	post ACQ	1st trial RI-Test	0.213	0.086	0.044	0.382	2.471	74	0.016	0.072	0.078
Fear Ratings	CS dis.	ranked	post-pre ACQ	1st trial RI-Test	0.236	0.116	0.009	0.463	2.043	60	0.045	0.066	0.071
Fear Ratings	CS dis.	ranked	pre EXT	1st trial RI-Test	0.217	0.094	0.033	0.401	2.309	69	0.024	0.075	0.082
Fear Ratings	CS dis.	ranked	pre-post EXT	1st trial RI-Test	0.138	0.097	-0.052	0.328	1.417	68	0.161	0.029	0.030
Fear Ratings	CS dis.	ranked	post EXT	1st trial RI-Test	0.143	0.081	-0.016	0.302	1.769	71	0.081	0.047	0.050
Fear Ratings	CS+	ranked	post-pre ACQ	pre EXT	0.516	0.101	0.318	0.714	5.107	91	0.000	0.215	0.274
Fear Ratings	CS+	ranked	post ACQ	pre EXT	0.584	0.087	0.413	0.755	6.675	97	0.000	0.326	0.484

Outcome	Stim.-type	Ranking	Predictor	Criterion	<i>b</i>	<i>SE_b</i>	Lower 95% CI	Upper 95% CI	<i>t</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	Cohen's <i>f</i> ²
Fear Ratings	CS+	ranked	post-pre ACQ	pre-post EXT	0.497	0.094	0.313	0.681	5.267	90	0.000	0.202	0.253
Fear Ratings	CS+	ranked	post ACQ	pre-post EXT	0.442	0.090	0.266	0.618	4.930	96	0.000	0.191	0.236
Fear Ratings	CS+	ranked	post-pre ACQ	post EXT	0.070	0.136	-0.197	0.337	0.517	92	0.607	0.003	0.003
Fear Ratings	CS+	ranked	post ACQ	post EXT	0.208	0.119	-0.025	0.441	1.744	101	0.084	0.029	0.030
Fear Ratings	CS+	ranked	post ACQ	1st trial RI-Test	0.364	0.087	0.193	0.535	4.192	85	0.000	0.162	0.193
Fear Ratings	CS+	ranked	post-pre ACQ	1st trial RI-Test	0.286	0.096	0.098	0.474	2.985	79	0.004	0.092	0.102
Fear Ratings	CS+	ranked	pre EXT	1st trial RI-Test	0.382	0.081	0.223	0.541	4.732	82	0.000	0.198	0.247
Fear Ratings	CS+	ranked	pre-post EXT	1st trial RI-Test	0.228	0.089	0.054	0.402	2.568	81	0.012	0.066	0.071
Fear Ratings	CS+	ranked	post EXT	1st trial RI-Test	0.166	0.076	0.017	0.315	2.199	84	0.031	0.056	0.059
Fear Ratings	CS-	ranked	post-pre ACQ	pre EXT	0.430	0.169	0.099	0.761	2.552	86	0.012	0.090	0.098
Fear Ratings	CS-	ranked	post ACQ	pre EXT	0.558	0.093	0.376	0.740	5.965	98	0.000	0.276	0.381
Fear Ratings	CS-	ranked	post-pre ACQ	pre-post EXT	0.086	0.136	-0.181	0.353	0.629	85	0.531	0.006	0.006
Fear Ratings	CS-	ranked	post ACQ	pre-post EXT	0.192	0.080	0.035	0.349	2.405	97	0.018	0.059	0.062
Fear Ratings	CS-	ranked	post-pre ACQ	post EXT	0.250	0.167	-0.077	0.577	1.500	88	0.137	0.030	0.031
Fear Ratings	CS-	ranked	post ACQ	post EXT	0.443	0.098	0.251	0.635	4.512	100	0.000	0.171	0.206
Fear Ratings	CS-	ranked	post ACQ	1st trial RI-Test	0.144	0.081	-0.015	0.303	1.775	88	0.079	0.037	0.038
Fear Ratings	CS-	ranked	post-pre ACQ	1st trial RI-Test	0.050	0.123	-0.191	0.291	0.411	78	0.682	0.002	0.002
Fear Ratings	CS-	ranked	pre EXT	1st trial RI-Test	0.148	0.075	0.001	0.295	1.979	86	0.051	0.043	0.045
Fear Ratings	CS-	ranked	pre-post EXT	1st trial RI-Test	0.003	0.103	-0.199	0.205	0.025	85	0.980	0.000	0.000
Fear Ratings	CS-	ranked	post EXT	1st trial RI-Test	0.249	0.071	0.110	0.388	3.503	87	0.001	0.126	0.145

Note. Stim. = Stimulus, CI = Confidence Interval, CS dis. = CS discrimination, pre = prior to the experimental phase, post = subsequent to the experimental phase, ACQ = Acquisition training, EXT = Extinction training, RI = Reinstatement, RI-Test = Reinstatement-Test.

Regressions including extinction learning rates by using different operationalizations

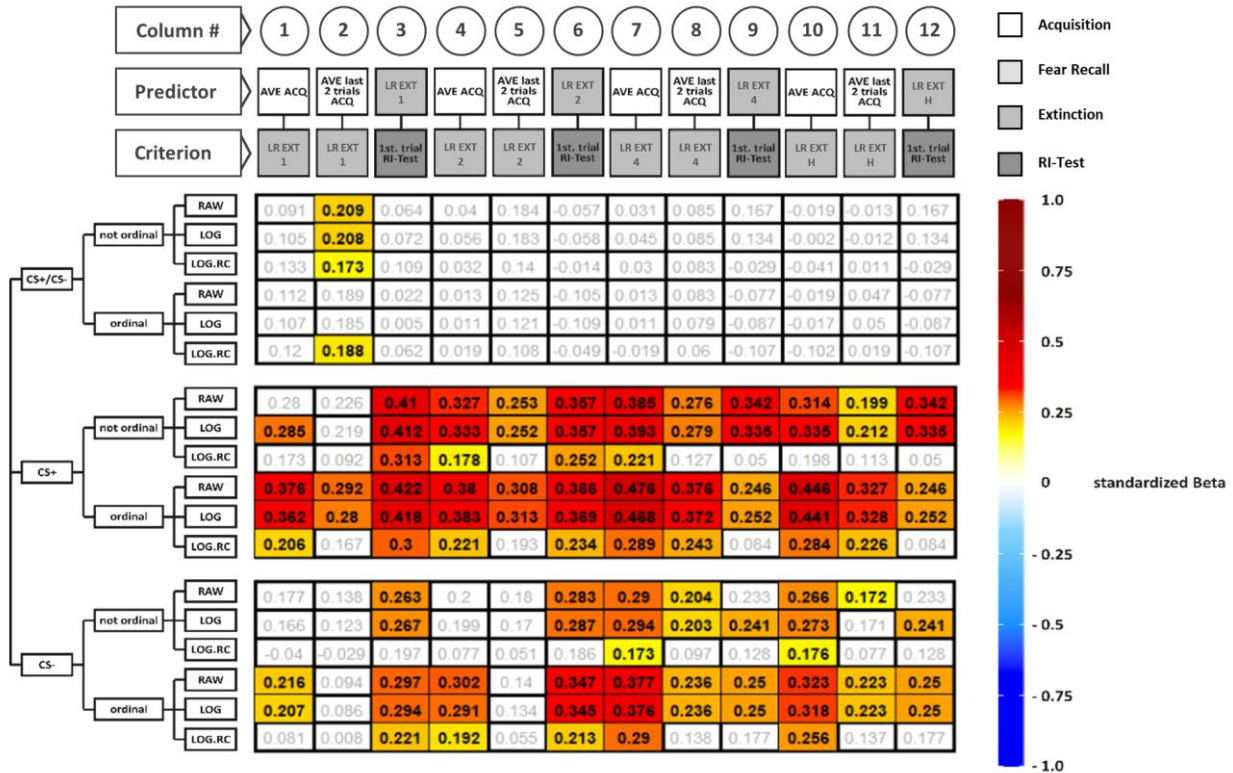


Figure 5-figure supplement 1. As per reviewer’s request, we illustrate standardized betas derived from non-preregistered regressions including SCR extinction training learning rates (LR EXT). As there is no agreed upon approach, we provide a small manyverse of approximations of extinction learning rates: We subtracted i) the last extinction trial from the first extinction trial (i.e., for CS-discrimination during the first and last trial, for CS+ and for CS- respectively; LR EXT 1, columns 1 - 3), ii) the last two extinction trials from the first two extinction trials (LR EXT 2, columns 4 - 6), iii) the last quarter of trials from the first quarter of trials (i.e., 4 trials; LR EXT 4, columns 7 - 9) and iv) the last half from the first half of trials (i.e., 7 trials; LR EXT H, columns 10 - 12). We acknowledge that learning rates have been inferred through different approaches in the literature (see e.g., Ney et al., 2020, Ney et al., 2022) and are often calculated from model-based approaches such as Rescorla Wagner Model (Seel, 2012) and hence our

operationalizations are only four out of multiple equally justifiable options. Colored cells indicate statistical significance of standardized betas, non-colored cells indicate non-significance.

Standardized betas are color-coded for their direction and magnitude showing positive values from yellow to red and negative values from light blue to dark blue. Darker colors indicate higher betas. AVE = average, LOG = log-transformed data, LOG.RC = log-transformed and range corrected data, not ordinal = not ordinally ranked data, ordinal = ordinally ranked data.

References

- Cooper, S. E., Dunsmoor, J. E., Koval, K., Pino, E., & Steinman, S. (2022). *Test-Retest Reliability of Human Threat Conditioning and Generalization* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/84uqz>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.
- Fredrikson, M., Annas, P., Georgiades, A., Hursti, T., & Tersman, Z. (1993). Internal consistency and temporal stability of classically conditioned skin conductance responses. *Biological Psychology*, 35(2), 153–163. [https://doi.org/10.1016/0301-0511\(93\)90011-V](https://doi.org/10.1016/0301-0511(93)90011-V)
- Ridderbusch, I. C., Wroblewski, A., Yang, Y., Richter, J., Hollandt, M., Hamm, A. O., Wittchen, H.-U., Ströhle, A., Arolt, V., Margraf, J., Lueken, U., Herrmann, M. J., Kircher, T., & Straube, B. (2021). Neural adaptation of cingulate and insular activity during delayed fear extinction: A replicable pattern across assessment sites and repeated measurements. *NeuroImage*, 237, 118157. <https://doi.org/10.1016/j.neuroimage.2021.118157>

- Savage, J. E., Moore, A. A., Sawyers, C. K., Bourdon, J. L., Verhulst, B., Carney, D. M., Moroney, E., Machlin, L., Kaabi, O., Vrana, S., Grillon, C., Brotman, M. A., Leibenluft, E., Pine, D. S., Roberson-Nay, R., & Hettema, J. M. (2019). Fear-potentiated startle response as an endophenotype: Evaluating metrics and methods for genetic applications. *Psychophysiology*, *56*(5), e13325. <https://doi.org/10.1111/psyp.13325>
- Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A Practical Guide to Calculating Cohen's f^2 , a Measure of Local Effect Size, from PROC MIXED. *Frontiers in Psychology*, *3*. <https://doi.org/10.3389/fpsyg.2012.00111>
- Torrents-Rodas, D., Fullana, M. A., Bonillo, A., Andi3n, O., Molinuevo, B., Caseras, X., & Torrubia, R. (2014). Testing the temporal stability of individual differences in the acquisition and generalization of fear: Stability acquisition and generalization of fear. *Psychophysiology*, *51*(7), 697–705. <https://doi.org/10.1111/psyp.12213>
- Zeidan, M. A., Lebron-Milad, K., Thompson-Hollands, J., Im, J. J. Y., Dougherty, D. D., Holt, D. J., Orr, S. P., & Milad, M. R. (2012). Test–Retest Reliability during Fear Acquisition and Fear Extinction in Humans. *CNS Neuroscience & Therapeutics*, *18*(4), 313–317. <https://doi.org/10.1111/j.1755-5949.2011.00238.x>

9 Study III

This article was published in *eLife*, 8, Lonsdorf, T. B., Klingelhöfer-Jens, M., Andreatta, M., Beckers, T., Chalkia, A., Gerlicher, A., Jentsch, V. L., Meir Drexler, S., Mertens, G., Richter, J., Sjouwerman, R., Wendt, J., & Merz, C. J., Navigating the garden of forking paths for data exclusions in fear conditioning research, e52465, License: CC BY 4.0 DEED, <https://creativecommons.org/licenses/by/4.0/>, no changes have been implemented (2019).

Navigating the garden of forking paths for data exclusions in fear conditioning research

Tina B Lonsdorf^{1*}, Maren Klingelhöfer-Jens¹, Marta Andreatta^{2,3}, Tom Beckers⁴, Anastasia Chalkia⁴, Anna Gerlicher⁵, Valerie L Jentsch⁶, Shira Meir Drexler⁶, Gaetan Mertens⁷, Jan Richter⁸, Rachel Sjouwerman¹, Julia Wendt⁹, Christian J Merz⁶

¹Department of Systems Neuroscience, University Medical Center Hamburg Eppendorf, Hamburg, Germany; ²Department of Psychology, Biological Psychology, Clinical Psychology and Psychotherapy, University of Würzburg, Würzburg, Germany; ³Institute of Psychology, Education & Child Studies, Erasmus University Rotterdam, Rotterdam, Netherlands; ⁴Centre for the Psychology of Learning and Experimental Psychopathology and Leuven Brain Institute, KU Leuven, Leuven, Belgium; ⁵Faculty of Social and Behavioural Sciences, Programme group Clinical Psychology, University of Amsterdam, Amsterdam, Netherlands; ⁶Institute of Cognitive Neuroscience, Department of Cognitive Psychology, Ruhr University Bochum, Bochum, Germany; ⁷Department of Psychology, Utrecht University, Utrecht, Netherlands; ⁸Department of Physiological and Clinical Psychology/Psychotherapy, University of Greifswald, Greifswald, Germany; ⁹Biological Psychology and Affective Science, University of Potsdam, Potsdam, Germany

*For correspondence:
t.lonsdorf@uke.de

Competing interests: The authors declare that no competing interests exist.

Funding: See page 20

Received: 04 October 2019

Accepted: 16 December 2019

Published: 16 December 2019

Reviewing editor: Alexander Shackman, University of Maryland, United States

© Copyright Lonsdorf et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Abstract In this report, we illustrate the considerable impact of researcher degrees of freedom with respect to exclusion of participants in paradigms with a learning element. We illustrate this empirically through case examples from human fear conditioning research, in which the exclusion of ‘non-learners’ and ‘non-responders’ is common – despite a lack of consensus on how to define these groups. We illustrate the substantial heterogeneity in exclusion criteria identified in a systematic literature search and highlight the potential problems and pitfalls of different definitions through case examples based on re-analyses of existing data sets. On the basis of these studies, we propose a consensus on evidence-based rather than idiosyncratic criteria, including clear guidelines on reporting details. Taken together, we illustrate how flexibility in data collection and analysis can be avoided, which will benefit the robustness and replicability of research findings and can be expected to be applicable to other fields of research that involve a learning element.

Introduction

In the past decade, efforts to understand the impact of undisclosed flexibility in data collection and analysis on research findings have gained momentum – for instance in defining and excluding ‘outliers’ (*Simmons et al., 2011*). This flexibility has been referred to as ‘researcher degrees of freedom’ (*Simmons et al., 2011*) or ‘the garden of forking paths’ (*Gelman and Loken, 2013*) to reflect the fact that each decision during data processing and/or analysis will take the researcher down a different ‘path’. Importantly and concerningly, these different paths can lead to fundamentally different end-points (i.e., results and associated conclusions) despite an identical starting point (i.e., raw data) (*Silberzahn et al., 2018*). Often, researchers take a certain path without malicious intent to obtain

favorable results (e.g., 'p-hacking'; *Head et al., 2015*): the decision to follow a certain path may be based on unawareness of alternative paths (due to lack of specific background knowledge) or the researcher following the most obvious path from an individual perspective. The latter is influenced by the scientific environment, the research question at stake or practices previously published by researchers in the field.

Admittedly, there is substantial ambiguity in what constitutes 'the best decision' for data analysis, and none of the available options may be necessarily incorrect (*Simmons et al., 2011; Silberzahn et al., 2018*). More precisely, different paths in the garden of forking paths may be more or less appropriate for different research questions, experimental designs, outcome measures or samples. Consequently, it is notoriously difficult for researchers, particularly those new to a field, to make informed and hence appropriate decisions. As a matter of fact, it is difficult to anticipate the number of different paths available and the consequences of choosing one over the other, or to come up with facts that truly justify choosing one path over the other – even for experts in a field. However, simply choosing a particular path because others chose it before (i.e., adopting published exclusion criteria) can also be highly problematic, as decisions often hinge on study-specific characteristics that do not invariantly apply to other studies.

We argue that it is important to raise awareness to this issue. Specifically, we think that it is critical to discuss both the rationale behind and the consequences associated with taking different analytical paths in general and in specific sub-fields of research. Here, we exemplarily take up this discussion for human fear conditioning research as a case example for tasks with a learning element grounded in recent discussions in science in general (*Flake and Fried, 2019; John et al., 2012*) and in fear conditioning research specifically (*Lonsdorf et al., 2017; Lonsdorf et al., 2019*). Fear conditioning is a typical paradigm employed to study (emotional) learning and memory processes with a particularly strong translational perspective (*Lonsdorf et al., 2017; Vervliet et al., 2013*). Questions addressed in the field of human fear conditioning are often concerned with consolidation, retrieval, generalization or modification of conditioned responses. Hence, it has often been claimed that the study of these processes requires the acquisition of a robust conditioned response as a precondition. Therefore, participants are often (routinely) excluded from analyses if they appear to not have learned ('non-learners') or not have been responsive to the experimental stimuli ('non-responders') during fear acquisition training, in which one conditional stimulus (CS+) predicts an upcoming aversive unconditioned stimulus (US) and another conditional stimulus does not (CS–) (*Lonsdorf et al., 2017; Pavlov, 1927*).

Critically, 'non-learning' is most often defined as a failure to show discrimination between the CS+ and CS– in skin conductance responses (SCRs) – the most common outcome measure in the field (*Lonsdorf et al., 2017*). This practice may seem trivial at first glance and has been referred to as exclusion of 'non-learners', 'performance-based exclusion' or even 'exclusion of outliers'. Yet, defining a set of characteristics to identify individuals who 'did not learn' is operationalized in very heterogeneous ways across studies. The same applies to the criteria that determine what constitutes a 'non-responder' during fear acquisition training.

In addition to the heterogeneity in operationalization, other problems of performance-based exclusion of participants are worth noting: definitions of 'non-learners' are typically based on SCRs only (for exceptions see *Ahmed and Lovibond, 2019; Oyarzún et al., 2019*) and 'non-learners' are typically excluded from *all* analyses, that is, all experimental phases and outcome measures of a study. As SCRs are not a pure measure of either learning or fear, but rather reflect arousal levels (*Hamm et al., 1993*) that serve as proxies for fear learning, classification into 'learners' and 'non-learners' on the basis of this single outcome measure may induce substantial sample bias. First, defining 'non-learning' on one single outcome measure, such as SCRs, ignores the fact that successful CS+/CS– differentiation may be present in other outcome measures (*Hamm et al., 1993*) such as fear potentiated startle (FPS) or ratings of fear and contingencies (i.e., cognitive awareness of the CS+/US contingencies). As such, 'non-learning' as defined on a single outcome measure such as SCRs cannot comprehensively capture 'non-learning'. Second, the level of responding in SCRs and CS+/CS– discrimination has been shown to be associated with a vast number of individual difference factors (*Lonsdorf and Merz, 2017; Boucsein et al., 2012*) such as age and sex (for a discussion see *Boucsein et al., 2012*), ethnicity (*Alexandra Kredlow et al., 2017; Boucsein et al., 2012*), genetic make-up (*Garpenstrand et al., 2001*), use of oral contraceptives (*Merz et al., 2018b*) or personality traits (*Naveteur and Freixa I Baque, 1987*). Consequently, excluding participants from an

experiment as ‘non-learners’ may pre-select specific sub-samples and thus may thus severely hamper the generalizability and interpretation of the findings. Importantly, this practice may be a threat to and a limitation of the clinical translation of findings because it potentially leads to the selective exclusion of specific and highly relevant sub-groups. In fact, a recent meta-analysis suggests that patients suffering from anxiety disorders show overgeneralization of fear responding, which is enhanced when responding to the CS– (Duits *et al.*, 2015), which may lead to reduced CS+/CS– discrimination if the response to the CS+ is comparable.

The concerns discussed above are merely based on theoretical considerations. Below, we aim to address the important and controversial topic of exclusion of ‘non-learners’ and ‘non-responders’ in human fear conditioning research empirically. We set out to provide an overview and inventory of the exclusion criteria that are currently employed in the field by means of a systematic literature search following PRISMA guidelines (Moher *et al.*, 2009), covering a publication period of six months. Importantly, we distinguish between ‘non-learners’ (based on task performance, that is, CS +/CS– discrimination) and ‘non-responders’ (based on a lack of responsiveness) as assessed using SCRs. We expect the identified criteria for ‘non-learners’ and ‘non-responders’ to be characterized by noticeable heterogeneity (thus allowing for considerable researcher degrees of freedom) across studies. We thus aim to (1) raise awareness and (2) illustrate the impact of applying different exclusion criteria features (i.e., forking paths) on results and interpretation through case examples exemplified by the re-analyses of existing data sets. Finally, we aim to (3) provide

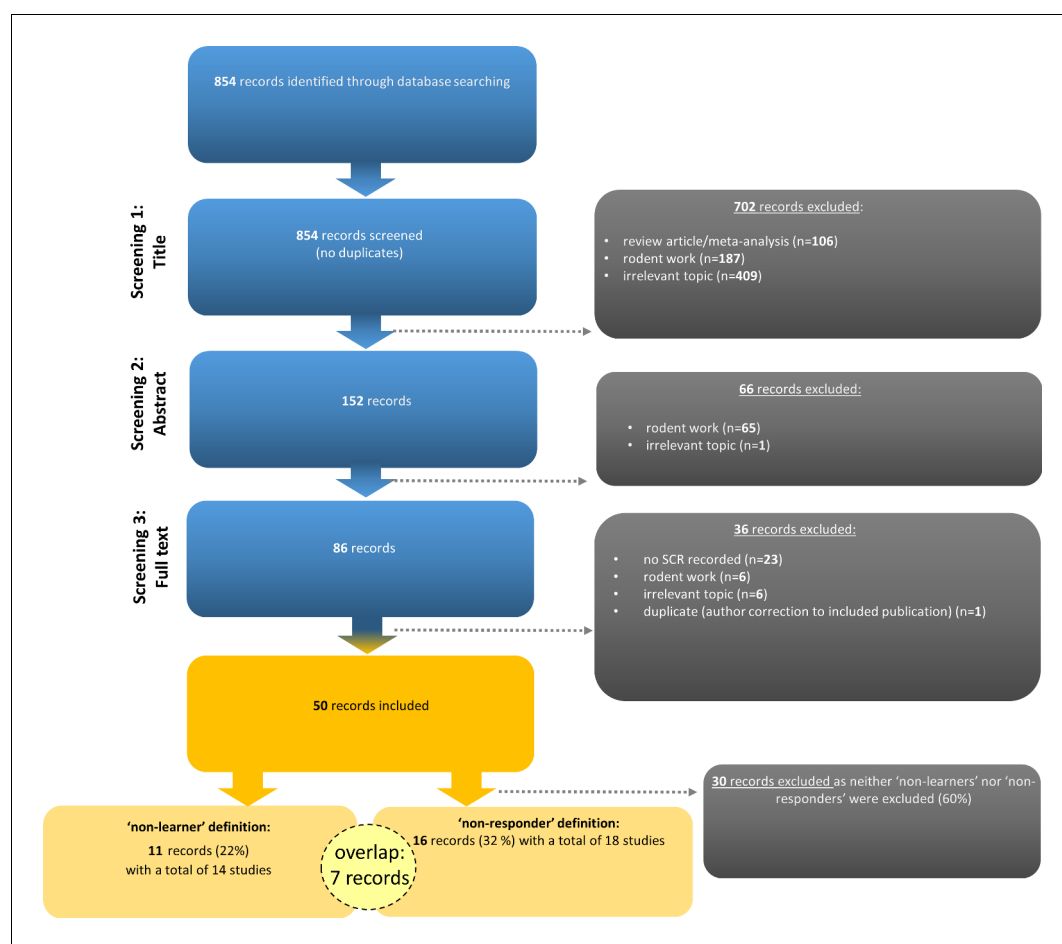


Figure 1. Flow chart illustrating the selection of records according to PRISMA guidelines (Moher *et al.*, 2009). Note that seven records (14%) employed the definition and exclusion of both ‘non-learners’ and ‘non-responders’. Examples of irrelevant topics included studies that did not use fear conditioning paradigms (see <https://osf.io/uxdhk/> for a documentation of excluded publications).

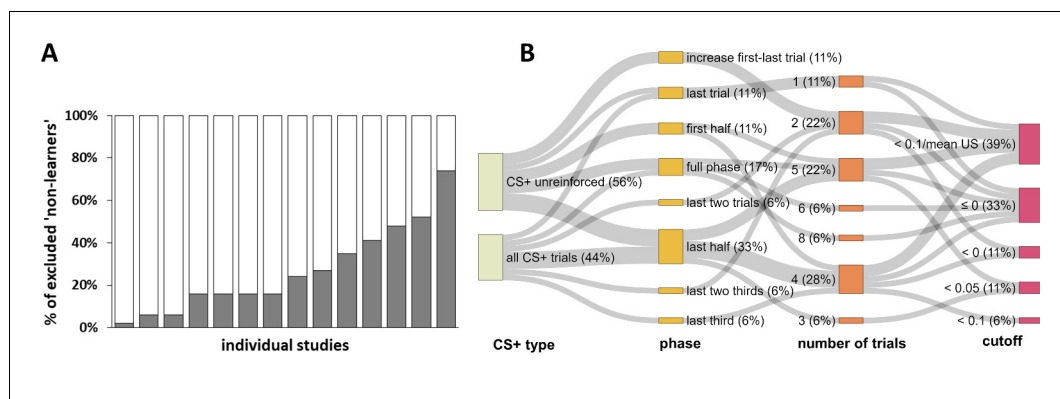


Figure 2. Graphical illustration of the percentage of 'non-learners' and forking path analysis across studies. (A) Illustration of the percentage of participants excluded ('non-learners') based on SCR CS+/CS− discrimination scores across studies included in the systematic literature search (note that these 14 individual studies are derived from 11 different records, as three records reported two individual studies each). Please note that some studies excluded participants on the basis of 'non-learning' as well as 'non-responding' (cf. **Figure 1**), and hence the percentages displayed here do not necessarily map onto the percentage of total participants excluded per study. Also note that the study with the highest percentage of excluded participants (i.e., 74%) reported the percentage of excluded participants as a single value that included 'non-learners' and 'non-responders'. This study is only included here because the largest proportion of exclusions can be expected to result from 'non-learning'. (B) Sankey plot showing the 'forking paths' of performance-based exclusion of participants as 'non-learners', illustrating differences in the experimental phase, number of trials, the SCR CS+/CS− discrimination score in μS used to define a 'non-learner', the CS+ type considered (illustrated as the nodes in graded colors) and their combinations used to define 'non-learners' across studies. Path width was scaled in relation to frequency of the combinations. Note that for some 'nodes' the percentages do not add up to 100% because of rounding.

methodologically informed, evidence-based recommendations for future studies with respect to defining and handling 'non-learners' and 'non-responders'.

Results

Definition of performance-based exclusion of participants ('non-learners') and number of participants excluded across studies

Slightly fewer than one fourth of the records (i.e., 22%; 11 out of 50 records comprising 14 individual studies as three records reported two studies each) included in the systematic literature search employed performance-based exclusion of participants (i.e., SCR 'non-learners', **Figure 1**).

Strikingly, every single one of these records used an idiosyncratic definition to define 'non-learners', yielding a total of eleven different definitions in the short period of six months (see **Appendix 1—table 1**). The percentages of excluded participants varied from 2% to 74% (**Figure 2A**) of the respective study sample. Definitions differed in i) the experimental (sub-)phases to which they were applied (i.e., whether the full phase or only the first half, second half or even single trials were considered), ii) the number of trials that the exclusion was based on (varying between one and eight single trials), iii) the CS+/CS− discrimination cutoff applied (varying between $<0 \mu\text{S}$ and $<0.1 \mu\text{S}$), and iv) the CS+ type (only non-reinforced or all CS+ trials) considered. The different forking paths and their frequency resulting from these combinations are displayed in **Figure 2B**.

The cutoff for CS+ versus CS− discrimination used to identify a 'non-learner' varied between $<0 \mu\text{S}$ and $<0.1 \mu\text{S}$, with most records excluding participants as 'non-learners' if they showed either a negative discrimination ($<0 \mu\text{S}$) and/or no discrimination ($\leq 0 \mu\text{S}$). These criteria apply if the SCR amplitude in response to the CS− was higher than and/or equal to the amplitude elicited by the CS+. Furthermore, most records required this criterion to be fulfilled only during the last half or the full fear acquisition training phase. Of note, the number of trials included in the same 'phase' category is contingent on the experimental design and hence does not represent a homogeneous category

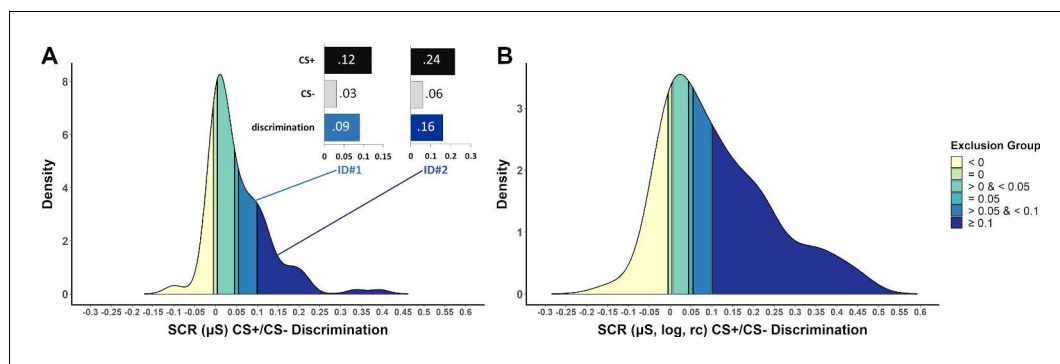


Figure 3. Density plots illustrating the frequency of CS+/CS- discrimination scores in a sample of $N = 116$ (Data set 1) based on the last half of the acquisition phase (including 7 CS+ and 7CS-, 100% reinforcement rate) for (A) SCR raw data and (B) logarithmized and range-corrected (rc; individual trial SCR/SCR_{max_across_all_trials}) SCR data (as it is typically not reported to which data exclusion criteria are applied). Color coding (yellow to blue) illustrates which part of the sample would be excluded when applying the performance-based exclusion criteria (i.e. CS+/CS- discrimination) as identified by the systematic literature search. Panel (A) also illustrates two case examples (ID#1 and ID#2) that differ in SCR amplitudes but importantly show the same discrimination ratio between CS+ and CS- (4:1). These two case examples illustrate that high CS+/CS- discrimination cutoffs favor individuals with high SCR amplitudes to remain in the final sub-sample. Data are based on a re-analysis of an unpublished data set recorded in the fMRI environment (Klingelhöfer-Jens M., Kuhn, M. and Lonsdorf, T.B.; unpublished). The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Percentages of participants excluded (Data set 1) when employing the different CS+/CS- discrimination cutoffs (as identified by the systematic literature search and graphically shown in **Figure 3B**) which are illustrated as density plots in **Figure 3**.

(‘last half’ may include five trials for one study comprising 10 trials in total but 10 trials for a different study employing 20 trials in total.

Applying the identified performance-based exclusion criteria to existing data: a case example

We applied the identified cutoff criteria to an existing data set (Data set 1) to exemplify the part of the sample that would be excluded when applying different cutoff criteria (shown in different colors from yellow to dark blue in **Figure 3**) based on the most frequently used phase restriction: the last half of fear acquisition training. CS+/CS- discrimination was calculated on the basis of raw (A) or log-transformed, range-corrected (log, rc) scores (B), because it is not usually reported which data are used to classify ‘learners’ vs. ‘non-learners’. Strikingly, the proportion of participants that are excluded is higher when CS+/CS- discrimination is calculated on the basis of raw data rather than log-transformed and range-corrected data (despite employing the same criteria) in particular for the highest ‘non-learner’ $< 0.01 \mu\text{S}$ cutoff (76.7% versus 52.6%, respectively) (see **Figure 3—figure supplement 1** for details).

In addition, we included a case example of two hypothetical individuals that differ in raw SCR amplitudes (ID#1: low and ID#2: high), but importantly show the same discrimination ratio (4:1) between CS+ and CS- (see **Figure 3A**). These two case examples illustrate that high CS+/CS- discrimination cutoffs, such as excluding individuals with discrimination scores $< 0.1 \mu\text{S}$ as ‘non-learners’, favor individuals with high SCR raw amplitudes.

Unsurprisingly, the exclusion group defined by a CS+/CS- discrimination cutoff $< 0 \mu\text{S}$ showed inverse discrimination (CS- \rightarrow CS+, not significant in raw SCRs [$p = 0.117$]; significant in log,rc SCRs [$p = 0.021$]). Strikingly and more importantly, most cumulative exclusion groups, as established by defining ‘non-learners’ by the CS+/CS- discrimination different cutoffs in SCRs in the literature, in fact show statistically significant CS+/CS- discrimination (see Appendix 2 for details and a brief discussion).

Note that despite the different color coding, which serves illustrative purposes only, the groups are in practice cumulative. More precisely, the groups illustrated by lighter colors are always

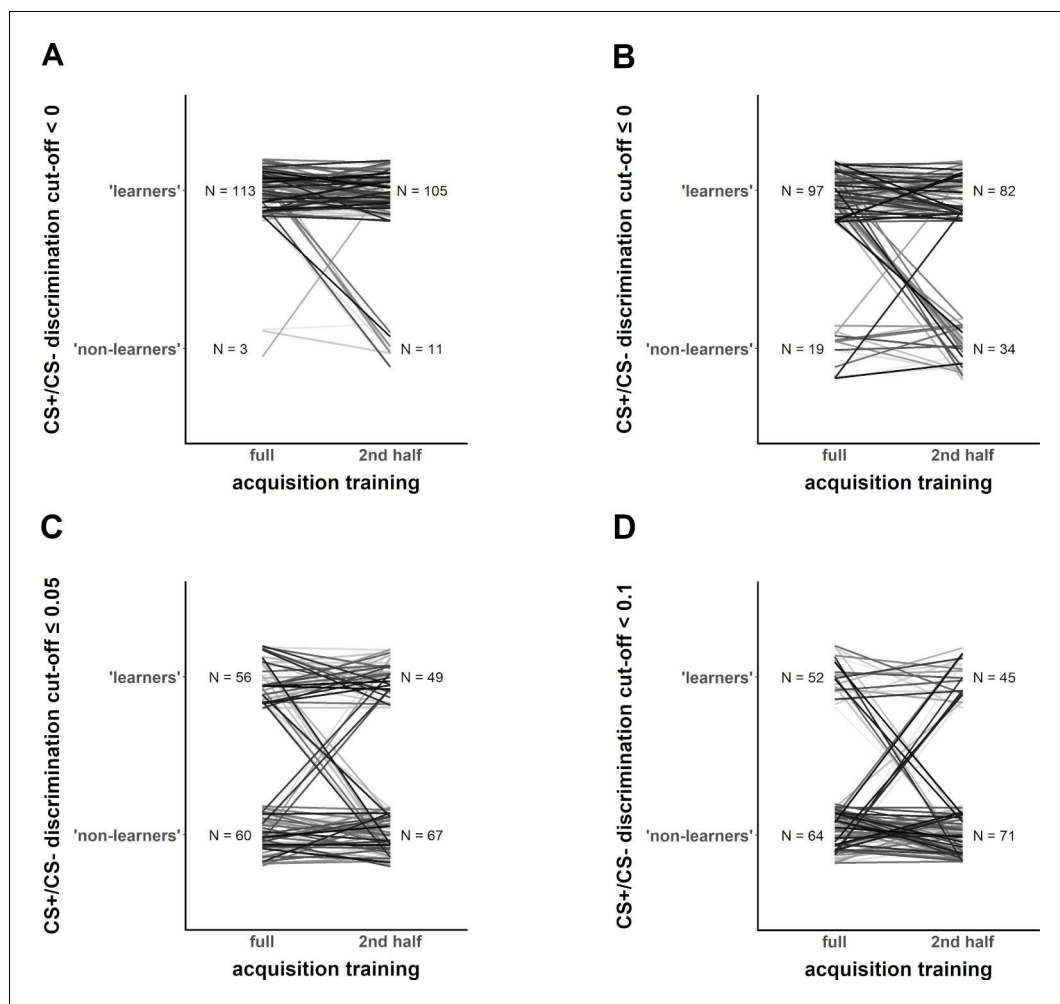


Figure 4. Exemplary illustration of individuals (Data set 1) that switch from being classified as ‘learners’ vs. ‘non-learners’ depending on the different CS+/CS– discrimination cutoff level (panels A–D), when calculation of CS+/CS– discrimination is based on either the full fear acquisition phase or the second half of the fear acquisition training (left and right part of each panel, respectively).

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Bar plots (mean \pm SE) on which the superimposed individual data points show CS+ and CS– amplitudes (of raw SCR values) and CS+/CS– discrimination in (A) fear ratings and (B) SCRs raw values in the group of ‘non-learners’, as exemplarily defined for this example as a group consisting of individuals in the two lowest SCR CS+/CS– discrimination cutoff groups (i.e., ≤ 0) in Data set 1.

contained in the darker colored groups when applying the respective cutoffs. For example, the group excluded when employing a cutoff of $< 0.1 \mu\text{S}$ (mid blue) also comprises the groups already excluded for the lower cutoffs of $= 0.05 \mu\text{S}$ (light blue), $< 0.05 \mu\text{S}$ (turquoise), $= 0 \mu\text{S}$ (light green) and $< 0 \mu\text{S}$ (yellow). For illustrative purposes, the different groups are treated as separate groups in this figure.

Exploratory analyses of consistency of classification (‘learners’ vs. ‘non-learners’) across outcome measures and criteria employed

The convergence of non-discrimination across different outcome measures was investigated by testing for CS+/CS– discrimination in fear ratings in individuals with different amounts of CS+/CS– discrimination in SCRs as defined by the criteria described above. In fact, individuals with non-significant and inverse CS+/CS– discrimination (i.e., $\leq 0 \mu\text{S}$) in SCRs showed significant CS+/CS– discrimination in fear ratings ($t_{31} = 9.69$, $p_{\text{bonf. corr}} < 0.000000001$, $d = 1.71$, see **Figure 4—figure**

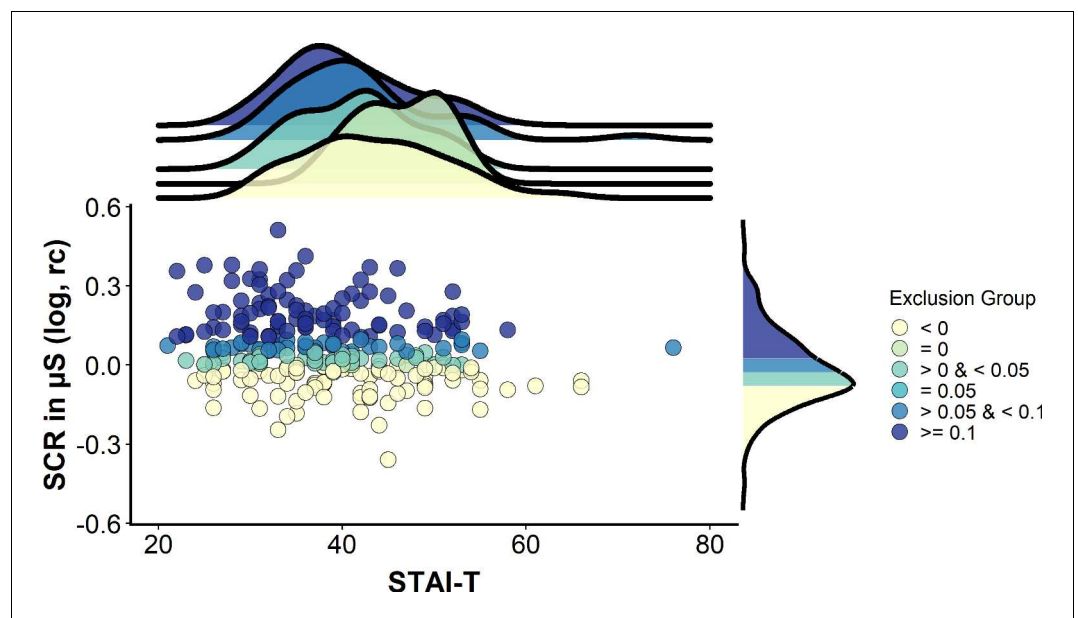


Figure 5. A case example illustrating potential sample bias induced by excluding individuals on the basis of CS+/CS- discrimination scores (based on logarithmized, range-corrected (rc) SCR data). Scatterplot illustrating the association between trait anxiety (measured via the trait version of the State-Trait Anxiety Inventory, STAI-T) and CS+/CS- discrimination scores in a sample of $N = 268$ (Data set 2). Color coding (yellow to blue) illustrates which part of the sample would be excluded when applying the performance-based exclusion criteria (i.e. CS+/CS- discrimination) as identified by the systematic literature search. Note that within this sample, no individuals were identified with CS+/CS- discrimination equaling $0.05 \mu\text{S}$. The upper panel illustrates densities for trait anxiety for the different CS+/CS- discrimination groups. The rightmost panel illustrates the density for CS+/CS- discrimination in the full sample. Data are based on a re-analysis of a data set recorded in the behavioral environment (Schiller et al., 2010). Note that despite the different color coding, which serves illustrative purposes only, the groups are in practice cumulative. More precisely, the groups illustrated by lighter colors are always contained in the darker colored groups when applying the respective cutoffs. For example, the group excluded when employing a cutoff of $<0.1 \mu\text{S}$ (mid blue) also comprises the groups already excluded for the lower cutoffs of $= 0.05 \mu\text{S}$ (light blue), $<0.05 \mu\text{S}$ (turquoise), $= 0 \mu\text{S}$ (light green) and $<0 \mu\text{S}$ (yellow). For illustrative purposes, the different groups are treated as separate groups in this figure.

supplement 1). Importantly, all cumulative exclusion groups showed significant CS+/CS- discrimination in fear ratings (all p 's < 0.002 , see **Appendix 3—table 1**).

We also illustrate (**Figure 4**) that the classification as 'learners' and 'non-learners' changes if two features (CS+/CS- discrimination cutoff and full vs. last half of acquisition training phase) of the criteria are changed (as illustrated in their full variation in **Figure 2B**).

The potential sample bias with respect to individual differences induced by employing different performance-based exclusion criteria: a re-analysis of existing data and a case example

Regarding the impact of performance-based exclusion on the pre-selection for certain individual differences, **Figure 5** shows that the distributions of trait anxiety were shifted to the left (i.e., towards lower scores) with higher SCR CS+/CS- discrimination cutoffs. More precisely, this means that, in this sample, highly anxious individuals display smaller CS+/CS- discrimination in SCRs, and that excluding individuals who display low discrimination scores will lead to the exclusion of anxious individuals.

In fact, we observed a main effect of 'Exclusion group' on trait anxiety score ($F_{[4,263]} = 219.2$, $p < 0.001$, $\eta_p^2 = 0.77$). All exclusion groups (corresponding to the color coding in **Figure 5**) differ significantly from each other in their trait anxiety scores (all $p_{\text{bonf_corr}} \leq 0.001$), except for the group that did not show any CS+/CS- discrimination ($= 0 \mu\text{S}$, light green, however $n = 6$ only), which showed significantly higher trait anxiety scores (mean \pm SD STAI score: 43.8 ± 6.1) than the group

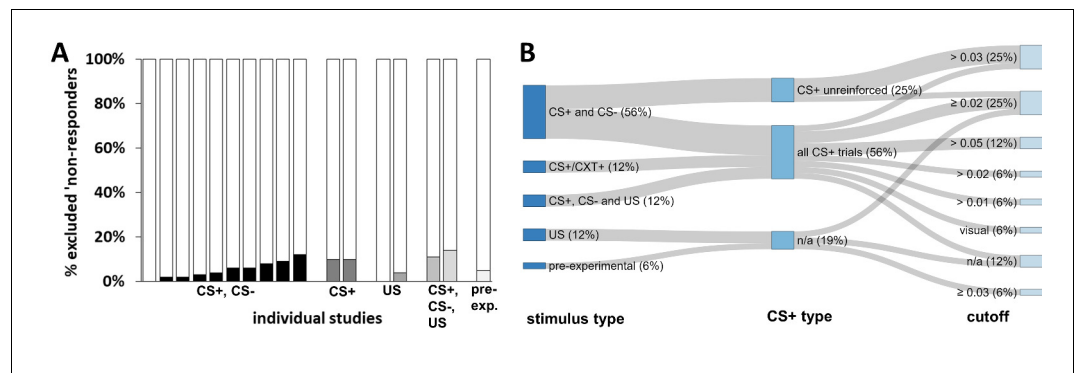


Figure 6. Graphical illustration of the percentage of ‘non-responders’ and forking path analysis across studies. (A) Illustration of the percentage of participants excluded from each study as a result of ‘SCR non-responding’ to (i) the conditioned stimuli (i.e., CS+ and CS-), (ii) the US, (iii) the CS+ (which also comprises a study that used the CXT+, i.e. context), (iv) the CS+, CS- and US or (v) a pre-experimental test. Note that these 18 individual studies are derived from 16 different records, two of which included two different studies that used the same criteria. Note that some studies excluded participants on the basis of ‘non-learning’ as well as ‘non-responding’, and hence the percentages displayed here do not necessarily map onto the percentage of total participants excluded from each study. Also note that a single study (*Schiller et al., 2018*) is not included in this visualization because it reported % ‘non-learners’ and % ‘non-responders’ as a single value. This value has been included in the visualization of ‘non-learners’ (*Figure 2*) as these are expected to represent the largest proportion. (B) Sanky plot illustrating the stimulus type (pre-experiment refers to determination of ‘responding’ in an unrelated phase prior to the experiment), the minimally required response amplitude in μ S (note that ‘visual’ refers to visual inspection of the data without a clear-cut amplitude cutoff, NA refers to no criterion applied) illustrated as the nodes in graded colors and their combinations that lead to classification as a ‘non-responder’. Path width was scaled in relation to frequency of the combinations. Note that for some ‘nodes’ the percentages do not add up to 100% because of rounding.

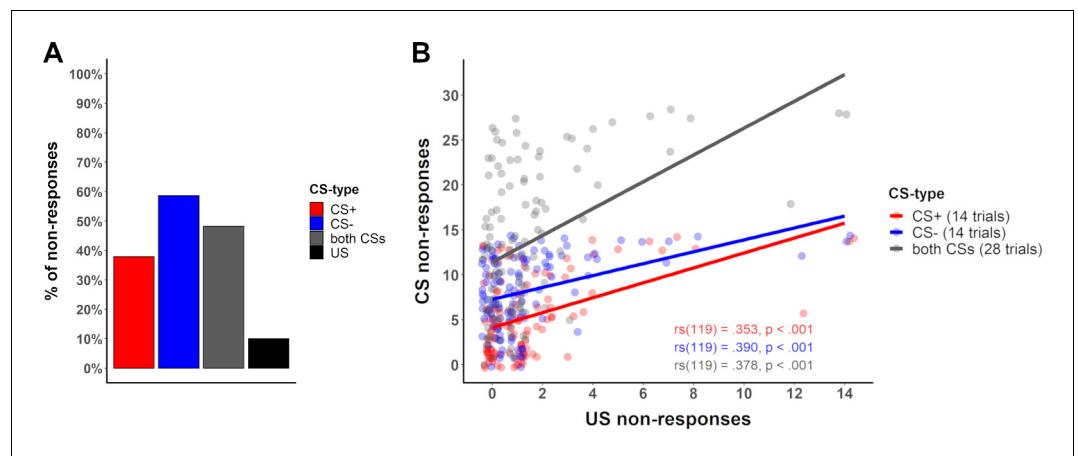


Figure 7. Percentage of no-responses across stimuli and correlation between CS and US non-responses. (A) Bar plot displaying the number of ‘non-responses’ to the CS+, CS-, across both CS and to the US across all participants in Data set 1 (see *Appendix 4—table 1* for percentages across different data sets). (B) Scatterplot illustrating the number of ‘non-responses’ (i.e., zero-responses, here defined by an amplitude $< 0.01 \mu$ S) to the US presentations (total of 14 presentations) and the CS+ (red) and CS- (blue) responses (14 presentations each) for each participant in Data set 1. For completeness sake, ‘non-responses’ across CS types are illustrated in gray (CS+ and CS- combined, total of 28 presentations). Lines illustrate the Spearman correlation (rs) between ‘non-responses’ to the US and ‘non-responses’ to the CS+, CS- and both CS, with corresponding correlation coefficients (font color corresponds to CS type) included in the figure.

with the largest CS+/CS− discrimination only (i.e., $\geq 0.1 \mu\text{S}$, dark blue, $n = 88$, mean \pm SD STAI score: 36.6 ± 8.5 , $p_{\text{bonf_corr}} \leq 0.001$, CI [0.133 to 0.279]). Nevertheless, trait anxiety scores in this group (light green) were not significantly larger than those in the group with the negative discrimination (i.e., $< 0 \mu\text{S}$, yellow, $n = 89$, mean \pm SD STAI score: 40.5 ± 9.7 , $p_{\text{bonf_corr}} = 0.10$, CI [−0.004 to 0.142]), the group with a small discrimination score (i.e., $> 0 \mu\text{S}$ but $< 0.05 \mu\text{S}$, light blue, $n = 43$, mean \pm SD STAI score 37.9 ± 7.9 , $p_{\text{bonf_corr}} = 1.0$, CI [−0.054 to 0.094]) or the group with the middle discrimination score (i.e., $> 0.05 \mu\text{S}$ but $< 0.1 \mu\text{S}$, mid blue, $n = 42$, mean \pm SD STAI score 38 ± 10.2 , $p_{\text{bonf_corr}} = 0.11$, CI [−0.005 to 0.145]).

Definition of ‘non-responders’ and number of participants excluded across studies

Thirty-two percent (i.e., 16 records) of the records in our systematic literature search included a definition and exclusion of ‘non-responders’, with percentages of participants excluded as a result of non-responding ranging between 0% and 14% (see **Figure 6A**). A single study (**Chauret et al., 2014; Oyarzún et al., 2012**) reported % ‘non-learners’ and % ‘non-responders’ as a single value (see **Appendix 1—table 2**). The definitions differed in: i) the stimulus type(s) used to define ‘non-responding’ (CS+ reinforced, CS+ unreinforced, all CS+s, CS−, US), ii) the SCR minimum amplitude criterion used to define a responder (varying between 0.01 μS and 0.05 μS ; visual inspection), and iii) the percentage of trials for which these criteria have to be met (see **Figure 6B** and **Appendix 1—table 2**), as well as a combination thereof.

‘Non-responding’ was most commonly defined as not showing a sufficient number of responses to the conditioned stimuli (CS+ and CS−), less frequently by the absence of responses to the US or any stimulus (CS+, CS− or US), and in two cases by the absence of responses to the CS+ or context (CXT+) specifically (see **Figure 6B**). Not surprisingly, the percentage of excluded participants differed substantially depending on the stimulus type used to define ‘non-responding’ (CS based, 0–10%; CS+/CXT+ based, 10–11%; US based, 0–4%; CS and US based, 11–14%; pre-experimental test based, 5%; **Figure 6A**).

Despite these differences in the stimulus types used to define ‘non-responding’ in the first place, studies differed widely in the amplitude cutoff criterion to be exceeded in order to qualify as a response (see **Figure 6B**) as well as in the percentage of trials in which this cutoff had to be met (see **Appendix 1—table 2**).

The question of what (physiological) ‘non-responders’ during fear acquisition training are and how to identify them might be elucidated by investigating the number of ‘non-responses’ across trial types (CS and US) across data sets, and whether ‘non-responding’ to the US predicts ‘non-responding’ to the CS or vice versa. As expected from **Figure 6A**, the number of ‘non-responses’ to the US was low (as was also the case in our data [10%, Data set 1]), while the number of ‘non-responses’ to the CS (48.29%) was substantially higher – in particular for the CS− (58.6%; CS+ ‘non-responses’: 37.9%, see **Figure 7A**). This pattern, exemplarily illustrated here in one data set is representative of a larger number of data sets (see Appendix 4, table 1 for details). Furthermore, in our data (Data set 1), all individuals that did not react to the US in more than two thirds of the US trials also showed no responses to the CS ($n = 3$ of $N = 119$). To summarize, this provides the first evidence that ‘non-responding’ to the US may predict ‘non-responding’ to the CS but not vice versa. Furthermore, our data also suggest a positive correlation between the number of ‘non-responses’ to the US and the number of ‘non-responses’ to the CS (see **Figure 7B** for statistics).

Discussion

In this article, we showed that participant exclusion in fear conditioning research is common (i.e., 40% of records included) and characterized by substantial operationalizational heterogeneity of definitions for ‘non-learners’ and (physiological) ‘non-responders’. Furthermore, we provide case-examples that illustrate: i) the futility of some definitions of ‘non-learners’ (i.e., when those classified as ‘non-learners’ in fact show significant discrimination on both ratings and SCRs as illustrated in Appendix 3 and Appendix 2, respectively) when applied to our data; and ii) the potential sample bias induced by excluding ‘non-learners’ with respect to individual differences. Furthermore, we provide an overview of SCR ‘non-responses’ to different stimulus types (CS+, CS− and US) across different data sets (see **Appendix 4—tables 1 and 2**) as a guide for developing evidence-based criteria

Box 1. List of reporting details, potential difficulties and recommendations when excluding ‘non-learners’ (performance-based exclusion) and/or ‘non-responders’ with a focus on SCRs.

Please note that this Box can be annotated online.

(A) General reporting details

What to report?	Why is this considered important?	What can go wrong or be ambiguous?	Recommendations on how to proceed
Details on data recording and response quantification pipeline	<ul style="list-style-type: none"> because differences in data recording and quantification (i.e., response scoring) can make a substantial difference 		<ul style="list-style-type: none"> report recording equipment and all settings used (e.g., filter) report software used for response quantification report precise details of response quantification
Minimal response criterion (μS) to define a valid SCR	<ul style="list-style-type: none"> to define valid responses 	<ul style="list-style-type: none"> minimally detectable amplitude (e.g., 0.01, 0.02, 0.03, 0.05 μS, etc.) may be sample- and equipment-specific no clear recommendations (existing guidelines provide a range of 0.01 to 0.05 μS) because this is influenced by noise level and equipment 	<ul style="list-style-type: none"> test different minimal response criteria in the data set and define the cutoff empirically. In our experience (Data set 1), a cutoff was easily determined empirically by visually inspecting responses at different cutoffs (e.g., <0.01 μS, between 0.01 μS and 0.02 μS) and by evaluating their discrimination from noise
Whether the first CS+ and/or the first CS- trial is included or not, and information on trial sequence	<ul style="list-style-type: none"> no learning can be evident in the first trial, as the first US may occur at the earliest at the end of the CS+ and hence after the scoring window for the CS+-induced SCR if the first trial is a CS-, no learning can have taken place as the US has not been presented yet inclusion of the first trial (or the first trials in partial reinforcement protocols) may thus artificially reduce CS+/CS- discrimination 	<ul style="list-style-type: none"> in fully randomized partial reinforcement protocols, US presentations may cluster in the first or last half of the acquisition training, which will impact on CS+/CS- discrimination in SCRs 	<ul style="list-style-type: none"> careful experimental design with respect to trial-sequences (in particular in partial reinforcement protocols) report whether the first trial for both CS+ and CS- is excluded because it may induce noise and bias CS+/CS- discrimination towards non-discrimination and as the first trial is sensitive to trial sequence effects
Precise number of trials considered (if applicable for each trial type including reinforced and non-reinforced CS+ trials in case of partial reinforcement)	<ul style="list-style-type: none"> often difficult/ambiguous to infer this information from the ‘Materials and methods’ section of a report^a number of trials that the ‘last half’ or ‘full phase’ refers to is contingent on experimental design and hence ambiguous and imprecise (see Figure 2B) 		<ul style="list-style-type: none"> precision in reporting rather than relying on the reader making the right inferences specify clearly the number of trials per stimulus type that are comprised in the ‘last half’ or ‘full phase’ provide a justification (theoretical and/or empirical) for this decision^b
Details of whether results were based on raw or transformed data	<ul style="list-style-type: none"> typically, transformations are required to allow interpretation of the reported results and to meet the assumptions of commonly statistical models 		<ul style="list-style-type: none"> report details of transformation (e.g., logarithmized [log/LN], range-corrected, square-root) including the number of trials considered (for each stimulus type) and the sequence of transformations applied and specific formula (e.g., for range-correction) provide justification for any applied transformation (e.g., violation of assumption of normal distribution of residuals)

Precise number of excluded participants and specific reasons	<ul style="list-style-type: none"> • often difficult/ambiguous to infer this information from the 'Materials and methods' section of a report^a 	<ul style="list-style-type: none"> • different researchers have different opinions on what 'exclusion' is (e.g., having individuals discontinue after a first experimental day based on performance should be considered and reported as exclusion) 	<ul style="list-style-type: none"> • report a breakdown of specific reasons for exclusions with respective <i>n</i>'s
--	--	--	--

(B) Specific reporting details for exclusion of 'non-learners'

What to report?	Why is this considered important?	What can go wrong or be ambiguous?	Recommendations on how to proceed
CS+/CS- discrimination is calculated on the basis of raw SCR or transformed (e.g., logarithmized [log/LN], range-corrected, square-root) scores	<ul style="list-style-type: none"> • the same criteria lead to different proportions of excluded individuals when applying them to raw or transformed data (see Figure 3A and B) 		<ul style="list-style-type: none"> • exact details of transformations (optimally calculation formulas) need to be included for full transparency and reproducibility
Minimal differential (CS+ vs. CS) cutoff for 'non-learning' in μ S	<ul style="list-style-type: none"> • different cutoffs lead to very different proportions of individuals excluded (see Figure 3) 		<ul style="list-style-type: none"> • exact details on cutoffs need to be included for full transparency and reproducibility
On what outcome measures is 'non-learning' determined?	<ul style="list-style-type: none"> • 'non-learners' do not necessarily converge across different outcome measures (Appendix 3, Figure 4—figure supplement 1) 	<ul style="list-style-type: none"> • all outcome measures recorded need to be reported 	<ul style="list-style-type: none"> • 'non-learning' should not be based on a single outcome measure or a clear justification needs to be provided as to why a single measure is considered meaningful
If 'non-learning' is determined by responding during fear acquisition training, which trial types and number of trials per trial type were considered?	<ul style="list-style-type: none"> • depending on the criteria employed, the same individual may be classified as 'learner' or 'non-learner' (see Figure 4) 		<ul style="list-style-type: none"> • classification as 'non-learner' should be based on differential scores (CS+ vs. CS-), and the number of trials included for this calculation should be clearly justified. Providing a generally valid recommendation regarding the number of trials to be included is difficult because it critically depends on experimental design choices
If 'non-learning' criteria are used, do they differ from criteria that the researcher or the research group used in previous publications? If yes, why were the criteria changed?	<ul style="list-style-type: none"> • provide explicit justifications on why different criteria were used previously and presently 		<ul style="list-style-type: none"> • report differences between present and previous criteria used including references and justifications
Did 'non-learners' really fail to learn?	<ul style="list-style-type: none"> • important as a manipulation check but note that the absence of a statistically significant CS+/CS- discrimination effect in a group on average cannot be taken to imply that all individuals in this group do not show meaningful CS+/CS- discrimination 	<ul style="list-style-type: none"> • individuals classified as 'non-learners' may in fact show significant CS+/CS- discrimination in SCRs (see Appendix 2) or in other outcome measures (see Figure 3—figure supplement 1 and Appendix 4) and hence fail the manipulation check 	<ul style="list-style-type: none"> • do the groups classified as 'non-learners' and 'learners' differ significantly in discrimination, and do 'non-learners' really not discriminate in SCRs and other outcome measures? Report the data on this group graphically and/or statistically in the supplementary material (do not report the full sample with and without exclusions only)

Are results contingent on the exclusion of 'non-learners'?	<ul style="list-style-type: none"> important to allow for transparency and to evaluate the impact of the results 	<ul style="list-style-type: none"> it is not clearly defined when results differ meaningfully when excluding and including 'non-learners' 	<ul style="list-style-type: none"> provide results with and without exclusion of 'non-learners' additional analyses can be provided as supplementary material. When results are not contingent on the exclusion of 'non-learners', it is sufficient to mention this briefly in the results of the main manuscript (e.g., results are not contingent on the exclusion of 'non-learners') if the results of the main analyses and hence the main conclusions change when 'non-learners' are excluded, this needs to be included in the main manuscript, and the implications need to be adequately discussed. Please note that this does not necessarily invalidate findings but can refine them
Descriptive statistics for excluded 'non-learners'	<ul style="list-style-type: none"> important to allow for transparency and evaluation of the potential sample biases introduced 		<ul style="list-style-type: none"> report sex, age, anxiety levels, awareness

(C) Specific reporting details for exclusions of 'non-responders'

What to report?	Why is this considered important?	What can go wrong or be ambiguous?	Recommendations on how to proceed
Whether 'non-responses' are calculated on the basis of raw SCR or transformed (e.g., logarithmized [log/LN], range-corrected, square-root) scores	<ul style="list-style-type: none"> the same criteria lead to different proportions of excluded individuals when applying to raw or transformed data (see Figure 3A and B) 		<ul style="list-style-type: none"> exact details of transformations (optimally calculation formulas) need to be included for full transparency and reproducibility
Minimal cutoff for 'non-responses' in μS	<ul style="list-style-type: none"> it is often difficult/ambiguous to infer this information from the 'Materials and methods' section of a report^a higher cutoffs could unnecessarily reduce the sample size 		<ul style="list-style-type: none"> exact details on cutoffs need to be included for full transparency and reproducibility
Was 'non-responding' determined in a pre-experimental phase such as forced-breathing or US calibration?	<ul style="list-style-type: none"> determining 'non-responding' during a pre-experimental phase may help to detect malfunctioning of the equipment and allow this to be corrected prior to data acquisition classification of 'non-responders' independent of the experimental task and its specifications (e.g., number of US presentations) 	<ul style="list-style-type: none"> electrodes may detach between the pre-experimental phase and fear acquisition training 	<ul style="list-style-type: none"> report details of pre-experimental phase classification in SCR 'non-responders' should be based on a pre-experimental phase if no US presentations occur during the experiment, such as in case of threat of shock experiments, observational conditioning, extinction or return of fear tests
If 'non-responding' is determined by responding during fear acquisition training, what trial types are considered?	<ul style="list-style-type: none"> frequency of 'non-responding' differs substantially between different stimuli (CS and US) but also between CS+ and CS- (see Figure 7A) 	<ul style="list-style-type: none"> 'non-responding' to the US may be due to technical failure (i.e., no US was administered) 	<ul style="list-style-type: none"> classification in SCR 'non-responders' should not be based on SCRS elicited by CS (CS+, CS- or both), but should be based on US responding a question on the estimated number of US presented during fear acquisition training (and all other phases) may serve as a manipulation check
Descriptive statistics for excluded 'non-responder'	<ul style="list-style-type: none"> important to allow for transparency and evaluation of the potential sample biases introduced 		<ul style="list-style-type: none"> report sex, age, anxiety levels, awareness

^a based on our experience with extracting this information from literature identified in the systematic literature search reported in this manuscript.

^b 'others have done this previously' is not an acceptable justification in our point of view.

to define 'non-responders'. Together, we believe that this work contributes to: i) raising awareness of some of the problems associated with performance-based exclusion of participants ('non-learners') and of how this exclusion is implemented, ii) facilitating decision-making on which criteria to employ and not to employ, iii) enhancing transparency and clarity in future publications, and thereby iv) fostering reproducibility and robustness as well as clinical translation in the field of fear conditioning research and beyond.

'Non-learners': conclusions, caveats and considerations

Operationalizational heterogeneity is illustrated by every single record in our systematic literature search (covering a six months period) that employed definitions of 'non-learners' using a set of idiosyncratic criteria. The true number of definitions in the field applied over decades will be even substantially larger. In the records included here, 6–52% of participants were excluded (disregarding one study reporting percentages of 'non-learners' and 'non-responders' together with 74%; cf. **Figure 2A**), which substantially exceed the percentages recently put forward for 'non-learning' exclusions (**Marin et al., 2019**) that were suggested to lie between 4% (**Chauret et al., 2014**) and 19% (**Oyarzún et al., 2012**).

If several thousand analytical pipelines can be applied, the likelihood of false positives is high (**Munafò et al., 2017**) and the temptation of their opportunistic (ab)use must be considered a threat. Hence, a constructive discussion on where to go from here and how to not get lost in the garden of forking paths is important. This being said, we do acknowledge that certain research questions or the use of different recording equipment (robust lab equipment vs. novel mobile devices such as smartwatches) may potentially require distinct data-processing pipelines and the exclusion of certain observations (**Silberzahn et al., 2018; Simmons et al., 2011**), and hence it is not desirable to propose rigid and fixed rules for generic adoption. Procedural differences, in particular the inclusion of outcome measures that require certain triggers to elicit a response (such as startle responses or ratings) have also been shown to impact on the learning process itself (**Sjouwerman et al., 2016**). Rather, we call for a reconsideration of methods in the field and want to raise awareness to the pitfalls of adopting exclusion criteria from previously published work without critical evaluation of whether these apply meaningfully to one's own research. Furthermore, we want to promote the adoption of transparent reporting of data processing, recording and analyses and strive to suggest standards in the field to reduce heterogeneity based on idiosyncratic customs rather than methodological and theoretical considerations (see **Box 1**).

Yet, there are many other critical considerations worth discussing beyond the heterogeneous criteria used to define 'non-learners' and their impact on the outcome of statistical tests:

First, 'performance-based exclusion of participants' is often based on a single outcome measure (typically SCRs), despite multiple measures being recorded (for exceptions see **Ahmed and Lovibond, 2019; Belleau et al., 2018; Oyarzún et al., 2012**). Importantly, 'fear learning' cannot be reliably inferred by means of SCRs, because SCRs capture arousal-related processes and can only be used as a proxy to infer 'fear learning' as fear is closely linked to arousal (**Hamm and Weike, 2005**). Relatedly, the fact that physiological proxies of 'fear' do not map onto 'fear' itself has been discussed extensively (**LeDoux, 2012; LeDoux, 2014**).

Second, but related, individuals that fail to show CS+/CS− discrimination in SCRs may show substantial discrimination, as an indicator of successful learning, in other outcome measures such as ratings of fear, US expectancy or fear potentiated startle (**Hamm and Weike, 2005; Marin et al., 2019**), as illustrated here for fear ratings (see **Figure 4—figure supplement 1** and **Appendix 3—table 1**).

Third, a common justification for excluding 'non-learners' is that it is not possible to investigate extinction- or return-of-fear-related phenomena in individuals who 'did not learn'. To our knowledge, there is some evidence (**Craske et al., 2008; Plendl and Wotjak, 2010; Prenoveau et al., 2013**) that this theoretical assumption does not necessarily hold true, (i.e., CS+/CS− discrimination during fear acquisition training does not necessarily predict CS+/CS− discrimination during other experimental phases) (**Gerlicher et al., 2019**). An empirical investigation of this, however, would go beyond this manuscript's scope.

Fourth, we provided empirical evidence that those classified as a group of 'non-learners' in SCRs in the literature (sometimes referred to as 'outliers') on the basis of the identified definitions in fact displayed significant CS+/CS− discrimination when applied to our own data. An exception to this

was using cut offs in differential responding of $<0.05 \mu\text{S}$ (note, however, that a non-significant CS+/CS- discrimination effect in the group of 'non-learners' as a whole cannot be taken as evidence that all individuals in this group do not in fact display meaningful or statistically significant CS+/CS- discrimination). Hence, in addition to the many conceptual problems we raised here, the operationalization of 'non-learning' in the field failed its critical manipulation check given that those classified as 'non-learners' show clear evidence of learning as a group (i.e., CS+/CS- discrimination, see **Appendix 2—table 1**).

Fifth, we illustrate a concerning sample bias that is introduced by performance-based participant exclusion. CS+/CS- discrimination in SCRs during fear acquisition training has been linked to a number of individual difference factors (**Lonsdorf and Merz, 2017**) and, naturally, selecting participants on the basis of SCR CS+/CS- discrimination will also select them on the basis of these individual differences (illustrated by our case example on trait anxiety, **Figure 5**). In our case example, we illustrate that excluding 'non-learners' biases the sample towards low anxiety scores, which hampers the generalizability and replicability of findings: i) the effect may only exist in low-anxiety individuals but not in the general population, and ii) as fear acquisition is a clinically relevant paradigm, pre-selection in favor of low-anxiety individuals might represent a threat to the clinical translation of the findings. Many studies in the field of fear conditioning research aim to develop behavioral or pharmacological manipulations to enhance treatment effects or aim to study mechanisms that are relevant for clinical fear and anxiety. Hence, it is highly problematic that these studies may exclude individuals who show response patterns that mimic responses typically observed in anxiety patients when excluding 'non-learners'. In fact, patients suffering from anxiety disorders have been shown to be characterized by generalization of fear from the CS+ to the CS- (**Duits et al., 2015**).

Sixth, as illustrated by our case example (**Figure 3**), high CS+/CS- discrimination cutoffs generally favor individuals with high SCR amplitudes despite potentially identical ratios between CS+ and CS- amplitudes, which may introduce a sampling bias for individuals characterized by high arousal levels that probably have biological underpinnings. Relatedly, future studies need to empirically address which criteria for SCR transformation and exclusions are more or less sensitive to baseline differences (for an example from startle responding see **Bradford et al., 2015; Grillon and Baas, 2002**).

In summary, in light of the many (potential) problems associated with performance-based exclusion of participants, we forcefully echo Marin et al.'s conclusion that one needs "to be cautious when excluding SCR non-learners and to consider the potential implications of such exclusion when interpreting the findings from studies of conditioned fear" (**Marin et al., 2019**, abstract). Routinely, excluding participants who are intentionally or unintentionally characterized by specific individual differences represents a major threat to generalizability, replicability and potentially clinical translation of findings, as results might be contingent on a specific sub-sample and specific sample characteristics. This is also true when researchers are interested in the study of general processes. Furthermore, by excluding these individuals from further analyses, we may miss the opportunity to understand why some individuals do not show discrimination between the CS+ and the CS- in SCRs (or other outcome measures) or whether this lack of discrimination is maintained across subsequent experimental phases. It can be speculated that this lack of discrimination may carry meaningful information – at least for a subsample.

'Non-responders': conclusions, caveats and considerations

In addition to 'non-learners', 'non-responders' are also often excluded during fear conditioning research. We showed that the definition of 'non-responders', like that of 'non-learners', varies widely across studies. Heterogeneity in definitions manifests in different cutoff criteria for what is considered a valid response, the number of trials and the stimulus type(s) considered (**Appendix 1—table 2, Figure 6**). Surprisingly, most definitions are based on CS responses (i.e., SCRs to the CS+ and/or CS-) and only few are based on US responses. This highlights a potentially problematic overlap between 'non-learners' and 'non-responders': 'non-responding' to the CS (i.e., CS+ and CS- or CS+ only) is not necessarily indicative of physiological 'non-responding' – especially if high cutoffs are used. In fact, 'non-responding' to the CS may, or at least in some cases, reflect the absence of learning-based patterns in physiological responding – which may carry important information. Having observed the striking differences in percentages of 'non-responses' to the US (10%) and CS (48%) observed in our data (see **Figure 7** and **Appendix 4—table 1**), we suggest that physiological

'non-responding' cannot and should not be determined on the basis of the absence of responding to the CS.

More globally, the group of 'non-responders', as defined by the criteria identified here, probably lumps together several sub-groups: individuals (1) for whom technical problems resulted in no valid SCRs, (2) who fell asleep or did not pay attention, (3) who cognitively learned the CS+/US contingencies but did not express the expected corresponding responses in SCRs, and (4) who were attentive to the experiment but did not learn the contingencies (i.e., unaware participants) and hence did not show the expected SCR patterns (*Tabbert et al., 2011*).

In summary, although excluding physiological 'non-responders' makes sense (in terms of a manipulation check and independent of the hypothesis), we consider defining 'non-responders' on the basis of the absence of SCRs to the CS as problematic (dependent on the hypothesis). We suggest that physiological SCR 'non-responders' should be defined on the basis of US responses during fear acquisition training or to strong stimuli during pre-conditioning phases such as US calibration, startle habituation or forced breathing (reliably eliciting strong SCRs). If 'non-responding' to the US (during fear acquisition training) is used, it is difficult to suggest a universally valid cutoff with respect to the number or percentage of required valid US responses, because this critically depends on a number of variables such as hardware and sampling rate used. It remains an open question for future work whether data quality of novel mobile devices (e.g., smartwatches) for the acquisition of SCRs differs from traditional, robust lab-based recordings and how this would impact on the frequency of exclusions based on SCRs. Appendix 4 suggests that the cutoff may typically range between 1/3 and 2/3 of valid responses but may be data-set specific. US-based criteria are of course not trivial in multiple-day experiments, in which certain experimental days do not involve the presentation of US or involve few temporally clustered US presentations (i.e., reinstatement), or in paradigms not involving direct exposure to the US (i.e., observational or instructional learning; *Haaker et al., 2017*). In these cases, the other options listed above are strongly preferred to CS based criteria.

Where do we go from here?

In this work, we have comprehensively illustrated and argued that most of the current definitions employed to define 'non-learners' and 'non-responders' have to be considered as theoretically and empirically problematic. It is not sufficient, however, to raise awareness to these problems and the practical question of 'Where do we go from here?' remains to be addressed. What can we do to avoid getting lost in the garden of forking paths of exclusion criteria? Here, we would like to offer several solutions to improve practices in the field, which we expect to foster robustness, replicability and potentially clinical translation of findings: (1) transparency in reporting, (2) adopting open science practices, (3) increasing the level and quality of reporting and (4) graphical data presentation, (5) manipulation checks, and (6) fostering critical evaluation. We refer to see **Box 1** for specific recommendations.

More precisely, **transparency** can be enhanced 'if observations are eliminated, authors must also report what the statistical results are if those observations are included', as suggested by Simmons and colleagues, nearly a decade ago (*Simmons et al., 2011*, Table 2). Here, we echo this call that this recommendation should be implemented routinely in data reporting pipelines when employing performance-based participant exclusions ('non-learners') in fear conditioning research. We also call for a transparent and adequate reporting in the results (in brief) and discussion section rather than providing this information exclusively in the appendix. This being said, it is important to point out that should a finding turn out to be contingent on the exclusion of 'non-learners', this does not necessarily invalidate this finding. On the contrary, it may further specify the finding or hint to possible mechanisms and/or boundary conditions – yet inferences on boundary conditions should be made carefully (*Hardwicke and Shanks, 2016*). Relatedly, adopting an **open science culture** will facilitate transparent reporting of exclusion criteria (*Nosek et al., 2015*) and will minimize the risk of exploiting heterogeneous definitions in the field. Registered reports (*Hardwicke and Ioannidis, 2018*), publicly available data including those from excluded participants and pre-registration (*Munafò et al., 2017*) of definitions and analysis pipelines (*Ioannidis, 2014*), as well as openly accessible lab-specific standard operational protocols (SOPs), may also be helpful.

We acknowledge, however, that transparent reporting and particularly pre-registration of exclusion criteria is not trivial in light of the unsatisfactory **quality and level of detail in reporting** in the field of fear conditioning research. It was striking that the compilation of exclusion criteria ('non-

learners' and 'non-responders', see **Appendix 1—tables 1 and 2**) employed in the records included in our systematic literature search required extensive personal exchange with the authors because the definitions provided were often insufficient, ambiguous or incorrect. It is our responsibility as authors, reviewers and editors to improve these reporting standards to an acceptable level. As a guidance, **Box 1** provides a compilation of reporting details that we consider important to include in both pre-registered protocols and publications (an editable online version of **Box 1** is available to allow for further development, see Box caption).

Our recommendations to improve the level of reporting details and transparency extends to the **graphical illustration of results**, which should optimally allow for a complete presentation of data (**Weissgerber et al., 2015**) without risking obscuring important patterns, providing detailed distributional information rather than merely presenting summary statistics (see **Weissgerber et al., 2015** for a discussion). Such visualization options include, for instance, scatterplots, box plots, histograms, violin plots as well as their combination (see also **Figure 5**) in so called 'rain cloud plots' (see **Allen et al., 2018** for a tutorial in R, Matlab and Python) and utilizing colors or color gradients to visualize different groups of individuals (for instance 'learners' and 'non-learners') or discrimination scores. This will provide readers with the opportunity to evaluate the presented results and conclusions independently and comprehensively.

Finally, if criteria for 'non-learners' or 'non-responders' are employed to exclude participants from data analyses (or continuation of the experiment), we recommend that a **sanity or manipulation check** should be performed to determine whether – for instance - 'non-learners' really did not learn (i.e., really do not show significant CS+/CS– discrimination). We have empirically illustrated that most definitions of 'non-learners' fail this manipulation check (**Appendix 2—table 1**). Yet, it may not be feasible in all cases to determine such statistics, as these may not be appropriate for small samples and correspondingly small sub-groups of 'non-learners'. Relatedly, we urge authors to justify adequately all details of the exclusion criteria (if applied) – both theoretically and practically. Furthermore, we encourage authors, reviewers and editors alike to **critically evaluate** whether exclusions and applied criteria are warranted in the first place and appropriate in the specific context (vs. mere adopting published or previously employed criteria) and whether these exclusion criteria are transparently reported and discussed if results hinge on them (**Steege et al., 2016**).

Furthermore, future work should empirically address the question of how to best define 'non-learning' in particular in light of different outcome measures in fear conditioning studies, which capture different aspects of defensive responding (**Jentsch et al., 2020; Lonsdorf et al., 2017**).

Final remarks

In closing, the field of fear conditioning has been plagued with a lack of consensus on how to define and treat 'non-learners' and 'non-responders', which not seldomly impacts review processes and generates unnecessary lengthy discussions for editors, reviewers and authors. We argue that it is neither ethical (due to an excessive waste of tax money and human resources) nor scientifically meaningful to exclude up to two thirds of a sample. If only one third of the population performs 'as expected' in the experiment, experimental designs, data recording and processing techniques as well as definitions need to be reconsidered. We have shown that findings derived from such highly selective sub-samples may not generalize to other samples or to the general population, and as a consequence might be a threat to clinical translation. Most problematically, however, findings derived from such highly selective samples have been routinely and invariantly generalized to reflect 'general principles' and 'processes' in the past. Not surprisingly, such findings have also suffered replication failures. As such, exclusions of 'non-learners' can in fact be dangerous if not handled transparently (as suggested above), because they may bias and confuse a whole research field and may push research along a misleading path. Thus, we suggest recommendations and consensus suggestions, and recommend that common practices should be critically evaluated before we adopt them in future work, so that the field follows a path towards more robust and replicable research findings.

Materials and methods

This project has been pre-registered on the Open Science Framework (OSF) (Lonsdorf et al., 2019, March 22; retrieved from <https://osf.io/vjse4>).

A systematic literature search was performed according to PRISMA guidelines (Moher et al., 2009) covering all publications (including e-pubs ahead of print) in PubMed during the six months prior to the 22nd March 2019, using the following search terms: threat conditioning OR fear conditioning OR threat acquisition OR fear acquisition OR threat learning OR fear learning OR threat memory OR fear memory OR return of fear OR threat extinction OR fear extinction. In case of author corrections, we included the original study that the correction referred to unless this study itself was already included on the basis of the publication date.

From the identified 854 records listed in PubMed, 152 were included in stage 2 screening (abstract) and 86 were retained for stage 3 screening (full text). Finally, 50 records were included (see **Figure 1** for details) that reported results for (1) SCRs as an outcome measure from (2) the fear acquisition training phase (3) in human participants.

Extraction of criteria for ‘non-learners’ and ‘non-responders’

The 50 records were screened in-depth and information derived from each record was entered into a template file agreed on by the authors prior to literature screening (available from the OSF pre-registration <https://osf.io/vjse4>). We distinguished between ‘non-learners’ and ‘non-responders’. We considered an exclusion to be an exclusion of ‘non-learners’ if it was based on the key task performance –that is, CS+/CS– discrimination in SCRs. Exclusions were considered as exclusion of ‘non-responders’ when based on general (physiological) responding (i.e., not based on CS+/CS– discrimination). Participants who were explicitly excluded because of clear-cut and well-described technical problems, such as abortion of data recording or electrode disattachment, were not included in any definition. Criteria for defining ‘non-learners’ (see **Appendix 1—table 1**) and ‘non-responders’ (see **Appendix 1—table 2**) were extracted if applicable for the respective study. In case information in the publication was insufficient or ambiguous, the corresponding authors were contacted and asked for clarification.

Re-analysis of existing data applying the identified exclusion criteria

One aim of this work was to illustrate empirically the impact of different exclusion criteria on the study outcome and interpretation. To achieve this aim, we initially planned to re-analyze existing data sets and to exclude participants on the basis of the identified definitions, which was expected to demonstrate that results are not robust across the various definitions of ‘non-learners’ and ‘non-responders’ employed. More precisely, we planned to calculate CS+/CS– discrimination across different data sets for all definitions identified by the systematic literature search and to generate corresponding correlation matrices as well as the percentages of zero and non-responses (see pre-registration: <https://osf.io/vjse4>). Because the exclusion criteria identified through the systematic literature search were even more heterogeneous than expected, and as it was difficult to agree on a key outcome to quantify the impact of exclusion criteria, we eventually concluded that such extensive re-analyses would not add much to the tabular and graphical illustration of this heterogeneity. Instead, we provide illustrative case examples for: (i) the proportion of individuals excluded on the basis of the identified exclusion criteria for ‘non-learners’ (**Figure 3**) and (ii) the potential sample bias with respect to individual differences (exploratory aim) induced by employing different exclusion criteria features (i.e., discrimination cutoff; **Figure 5**). As planned, (iii) we provide the percentage of non-responses to the CS+, CS–, CS+ and CS– combined, and the US across different studies, as well as empirical information on the association between CS and US based non-responding as a base to guide empirical recommendations.

Data processing, statistical analyses and figures were generated with R version 3.6.0 (2019-04-26) using the following packages: cowplot, dplyr, ggplot2 (Wickham, 2009), ggridges, car, ez, lsr, psychReport, lubridate, RColorBrewer and flipPlot packages. Sanky plots were generated with help of <https://app.displayr.com>.

Data sets

Data set 1

Data set 1 is part of the baseline measurement of an ongoing longitudinal fear conditioning study. Here, fear ratings and SCR data from the first test-timepoint (T_0) were included ($N = 119$, 79 females, mean \pm SD age of 25 ± 4 years) whereas fMRI data were not used. All participants gave written informed consent to the protocol which was approved by the local ethics committee (PV 5157, Ethics Committee of the General Medical Council Hamburg).

Data set 1 is employed to illustrate a case example for the proportion of participants excluded when employing different CS+/CS- discrimination cutoffs ('non-learners', **Figure 3**) as well as the number of zero-responses across different stimulus types ('non-responders') and their association (**Figure 7**). Furthermore, we aimed to test exploratively whether even in groups defined as 'non-learners' a significant CS+/CS- discrimination on SCR and fear ratings can be detected (all results presented in the Appendix are based on Data set 1).

Paradigm and stimuli

The two-day paradigm consisted of habituation and acquisition training (day 1) and extinction training and recall testing (day 2) without any contingency instructions provided. Here, only data from the acquisition training phase (100% reinforcement rate) were used. CS were two light grey fractals, presented 14 times each in a pseudo-randomized order for 6–8 s (mean: 7 s). Visual stimuli were identical for all participants, but allocation to CS+ and CS- was counterbalanced between participants. During inter-trial intervals (ITIs), a white fixation cross was shown for 10–16 s (mean: 13 s). All stimuli were presented on a light gray background and controlled by Presentation software (Version 14.8, Neurobehavioral Systems, Inc, Albany California, USA).

The electro-tactile stimulus, serving as US, consisted of three 10 ms electro-tactile rectangular pulses with an interpulse interval of 50 ms (onset: 200 ms before CS+ offset) and was administered to the back of the right hand of the participants. It was generated by a Digitimer DS7A constant current stimulator (Welwyn Garden City, Hertfordshire, UK) and delivered through a 1 cm diameter platinum pin surface electrode (Speciality Developments, Bexley, UK). The electrode was attached between the metacarpal bones of the index and middle finger. US intensity was individually calibrated in a standardized step-wise procedure aiming at an unpleasant, but still tolerable level.

SCRs

SCRs were semi-manually scored by using a custom-made computer program (EDA View) as the first response from trough to peak 0.9–3.5 s after CS onset (0.9–2.5 s after US onset) as recommended (**Boucsein et al., 2012; Sjouwerman and Lonsdorf, 2019**). The maximum rise time was set to 5 s. Data were down-sampled to 10 Hz. Each scored SCR was checked visually, and the scoring suggested by EDA View was corrected if necessary (e.g., the foot or trough was misclassified by the algorithm). Data with recording artifacts or excessive baseline activity (i.e., more than half of the response amplitudes) were treated as missing data points and excluded from the analyses. SCRs below $0.01 \mu\text{S}$ or the absence of any SCR within the defined time window were classified as non-responses and set to 0. The threshold of $0.01 \mu\text{S}$ for this data set was determined empirically by visually inspecting response specifically above and below this cutoff, which suggested that in this data set, responses $> 0.01 \mu\text{S}$ can be reliably identified. 'Non-responders' ($N = 3$) were defined as individuals who showed more than two thirds of non-responses to the US (10 or more non-responses out of 14 US trials, see **Appendix 4—table 2**). Three individuals were classified as 'non-responders' and these individuals did not show any responses to the CS either. The three participants classified as 'non-responders' (see above) were only excluded for the analyses of 'non-learners'. Raw SCR amplitudes were normalized by taking the natural logarithm and range-corrected by dividing each logarithmized SCR by the maximum amplitude (maximum SCR to a CS or a US) per participant and day.

Fear ratings

Fear ratings were provided by participants through ratings on a visual analog scale (VAS) on the screen asking 'how much stress, fear, and tension' they experienced when they last saw the CS+ and CS-. The fear ratings used for the purpose of this manuscript are those obtained after fear acquisition training (no ratings were acquired during this phase). Answers were given within 5 s on the VAS,

which ranged from 0 (answer = none) to 25 (answer = maximum) by using a button box. Pressing the buttons moved a bar on the VAS to the aimed value and answers were logged in by pressing another button. Non-registered ratings were considered as missing values (8.4%).

Statistical analysis

To test whether exclusion groups differ in CS+/CS– discrimination, a mixed ANOVA with CS+/CS– discrimination in SCR or fear ratings as the dependent variable and the between-subjects factor ‘Exclusion group’ and the within-subject factor ‘CS-type’ was performed. Note that it is circular to test for differences in SCR CS+/CS– discrimination between groups that were selected on the basis of different SCR CS+/CS– discrimination cutoffs in the first place. Still, it is relevant to test whether all groups classified as ‘non-learners’ in the literature do indeed fail to show evidence of learning, which would be indicated by a lack of significant CS+/CS– discrimination in SCRs in this case. In essence, this is a test to evaluate whether the exclusion criteria used in the literature do indeed achieve what they purport to do, that is, classify a group of participants that do not show evidence of learning. To test whether these exclusion groups discriminated in SCRs and fear ratings, exclusion groups were cumulated, and t-tests were performed for each cumulative group (see Appendices 2 and 3, respectively). We acknowledge, however, that the absence of a statistically significant CS+/CS– discrimination effect in a group on average cannot be taken to imply that all individuals in this group do not show meaningful CS+/CS– discrimination. As such, this is a rather conservative test. To correct for multiple testing, all p-values deriving from t-tests were adjusted using the Bonferroni procedure. As effect size, Cohen’s *d* was reported for t-tests and partial eta-squared for ANOVAs. To illustrate the association between the non-responses to the US and the non-responses to the CS, a Spearman rank correlation test was computed.

Data set 2

For the purpose of this manuscript, a final sample of 268 individuals (195 female, mean \pm SD age of 25 ± 4 years) was re-analyzed. This sample is reported in a recent pre-print ([Sjouwerman et al., 2018](#)) in which we observed an association between trait anxiety and CS+/CS– discrimination in SCRs. Here, the re-analysis and graphical illustration of these data serve the purpose of a case example to illustrate the potential sample bias that may be induced by employing performance-based exclusion ([Figure 5](#)).

Paradigm and stimuli

A detailed experimental description is included in the preprint [Sjouwerman et al. \(2018\)](#). In brief, participants underwent a 100% reinforcement fear acquisition training phase in a behavioral laboratory setting, including 9 CS+ and 9 CS– trials, presented for 6 s each. Consequently, 9 US presentations were included that coincided 100 ms prior to CS+ offset. Trials were interleaved by 10–13 s ITIs with a white fixation cross presented on a black background. Black geometrical shapes served as CS, and electrical stimulation delivered by a DS7A electrical stimulator (Digitimer, Welwyn Garden City, UK) onto the outer surface of the right hand served as US. The intensity of the US was individually calibrated with a stair-case procedure in order to reach an unpleasant but tolerable level. Not of interest to the current case example were the acoustic startle probes (95 dB(A) burst of white noise) presented to elicit a startle response in two thirds of all acquisition trials, as well as three fear-rating blocks probed intermittently during fear acquisition training. Startle probes were presented 4 or 5 s post-CS onset, and 5 or 7 s post-ITI onset. No contingency instructions were given.

SCRs

SCRs were quantified as the first SCR within 0.9–4.0 s after stimulus onset (CS or US) and were scored semi-manually from trough-to-peak using a custom-made program. Signal increases smaller than $0.02 \mu\text{S}$ were treated as non-responses, that is set to 0. (Please note that this cut-off was not empirically determined as in Data set 1 but adopted from the previous publication of Data set 2. As we present re-analyses here, we decided not to change the cut-off to maintain comparability.) Responses confounded by recording artifacts, such as responses moving beyond the sampling window, excessive baseline activity, or electrode detachment were treated as missing values. Raw response amplitudes per trial were log-transformed and range-corrected for the maximum CS or US response per participant. Individuals not showing any valid SCR (i.e., missing or zero responses) in

more or equal than two thirds (≥ 6 out of 9, see **Appendix 4—table 2**) of US trials were treated as physiological ‘non-responders’ ($n = 19$) and were consequently excluded from graphical illustration and the statistical analysis. In addition, 31 participants were excluded prior to physiological processing, either because of abortion of the experiment or due to technical failures during data acquisition (e.g. errors during saving, overwritten logfile, or missing markers), leaving 307 out of 357 individuals with valid SCR data for fear acquisition training. Of these 307 participants, 39 had incomplete STAI-T data (*Spielberger et al., 1983*) resulting in a final sample size for this case example of 268 individuals (195 female, mean \pm SD age of 25 ± 4 years).

Statistical analysis

To test whether different exclusion groups differ in their mean trait anxiety levels, a univariate ANOVA with STAI-T score as the dependent variable and exclusion group as the independent variable was carried out. Post hoc pairwise *t*-tests were conducted to compare trait anxiety scores between the different exclusion group levels. The post hoc tests were corrected for multiple testing, using the Bonferroni correction method. 95% family wise confidence levels were determined using TukeyHSD tests.

Acknowledgements

The project is part of a European wide network (EIFEL-ROF network) of researchers working on meta-research topics in the field of fear conditioning research, which is funded by the German Research foundation (DFG; Grant ID LO1980/2-1) to TBL and CJM. Data used for re-analyses in the main manuscript paper are derived from projects funded by the DFG to TBL (Data set 1: Collaborative Research Center project number 44541416 TRR 58, sub-project B07. Data set 2: LO 1980/1-1). Data reported in **Appendix 1—table 1** are also funded by the German Research Foundation to CJM (Collaborative Research Center SFB 1280 project number 316803389, sub-project A09) and JW (WE 5873/1–1 and WE 2762/5–1). The authors represent eight different research groups across three European countries (Germany, the Netherlands, and Belgium).

We thank Prof. Dr. Matthias Gamer, University of Würzburg, for providing EDA View for SCR response quantification.

The authors thank Manuel Kuhn, Jan Haaker, Tanja Jovanovic and Dean Mobbs for helpful suggestions during discussions on this project. We also thank Jan Haaker for help with initial literature screening.

Additional information

Funding

Funder	Grant reference number	Author
Deutsche Forschungsgemeinschaft	LO 1980/2-1	Tina B Lonsdorf Christian J Merz
Deutsche Forschungsgemeinschaft	LO 1980/1-1	Tina B Lonsdorf
Deutsche Forschungsgemeinschaft	Project B07 44541416	Tina B Lonsdorf
Deutsche Forschungsgemeinschaft	316803389 - SFB1280	Christian J Merz
Deutsche Forschungsgemeinschaft	WE 5873/1-1	Julia Wendt
Deutsche Forschungsgemeinschaft	WE 5873/5-1	Julia Wendt

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Tina B Lonsdorf, Conceptualization, Data curation, Formal analysis, Supervision, Funding acquisition, Visualization, Methodology, Writing - original draft, Project administration, Writing - review and editing, Conceived the study, Organized the project group, Contributed to data acquisition, Data analysis (systematic literature search, Dataset 1, Dataset 2, Data visualization, Drafting the manuscript, Revised the draft critically and approved the final version to be published; Maren Klingelhöfer-Jens, Data curation, Formal analysis, Visualization, Writing - original draft, Writing - review and editing, Contributed to data analysis and data visualization (data set 1), Drafting the manuscript, Revised the draft critically and approved the final version to be published; Marta Andreatta, Tom Beckers, Anastasia Chalkia, Anna Gerlicher, Valerie L Jentsch, Shira Meir Drexler, Gaetan Mertens, Jan Richter, Julia Wendt, Formal analysis, Writing - original draft, Writing - review and editing, Contributed to data acquisition and data analysis (systematic literature search), Drafting the manuscript, Revised the draft critically and approved the final version to be published; Rachel Sjouwerman, Data curation, Formal analysis, Visualization, Writing - original draft, Writing - review and editing, Contributed to data acquisition, Data analysis (dataset 2), Data visualization, Drafting the manuscript, Revised the draft critically and approved the final version to be published; Christian J Merz, Conceptualization, Formal analysis, Supervision, Funding acquisition, Writing - original draft, Writing - review and editing, Conceived the study, Contributed to data acquisition and data analysis (systematic literature search), Drafting the manuscript, Revised the draft critically and approved the final version to be published

Author ORCIDs

Tina B Lonsdorf  <https://orcid.org/0000-0003-1501-4846>

Marta Andreatta  <https://orcid.org/0000-0002-1217-8266>

Tom Beckers  <https://orcid.org/0000-0002-9581-1505>

Anastasia Chalkia  <https://orcid.org/0000-0002-1613-2281>

Valerie L Jentsch  <https://orcid.org/0000-0001-9318-9540>

Shira Meir Drexler  <https://orcid.org/0000-0001-8797-6900>

Jan Richter  <https://orcid.org/0000-0002-7127-6990>

Julia Wendt  <https://orcid.org/0000-0003-2299-5881>

Christian J Merz  <https://orcid.org/0000-0001-5679-6595>

Ethics

Human subjects: Study 1: All participants gave written informed consent to the protocol which was approved by the local ethics committee (PV 5157, Ethics Committee of the General Medical Council Hamburg). Study 2: All participants gave written informed consent to the protocol which was approved by the Ethical Review Board of the German Psychological Association (TL072015).

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.52465.sa1>

Author response <https://doi.org/10.7554/eLife.52465.sa2>

Additional files

Supplementary files

- Transparent reporting form

Data availability

The minimal data sets (data set 1 and data set 2, both represent re-analysis of existing data), which were analysed during the current study, as well as code for figure production are available at OSF under <https://osf.io/mkxqe/> and DOI: <https://doi.org/10.17605/OSF.IO/MKXQE>.

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Tina B Lonsdorf, Maren Klingelhöfer-Jens	2019	Data_and_code_dataset1	https://osf.io/w9y8z/	Open Science Framework, w9y8z
Tina B Lonsdorf, Rachel Sjouwerman	2019	Data_and_code_dataset2	https://osf.io/7c5ag/	Open Science Framework, 7c5ag

References

- Ahmed O, Lovibond PF. 2019. Rule-based processes in generalisation and peak shift in human fear conditioning. *Quarterly Journal of Experimental Psychology* **72**:118–131. DOI: <https://doi.org/10.1177/1747021818766461>
- Alexandra Kredlow M, Pineles SL, Inslicht SS, Marin MF, Milad MR, Otto MW, Orr SP. 2017. Assessment of skin conductance in african american and Non-African american participants in studies of conditioned fear. *Psychophysiology* **54**:1741–1754. DOI: <https://doi.org/10.1111/psyp.12909>, PMID: 28675471
- Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit R. 2018. Raincloud plots: a multi-platform tool for robust data visualization. *PeerJ* **4**:63. DOI: <https://doi.org/10.7287/peerj.preprints.27137v1>
- Baeuchi C, Hoppstädter M, Meyer P, Flor H. 2019. Contingency awareness as a prerequisite for differential contextual fear conditioning. *Cognitive, Affective, & Behavioral Neuroscience* **19**:811–828. DOI: <https://doi.org/10.3758/s13415-018-00666-z>
- Belleau EL, Pedersen WS, Miskovich TA, Helmstetter FJ, Larson CL. 2018. Cortico-limbic connectivity changes following fear extinction and relationships with trait anxiety. *Social Cognitive and Affective Neuroscience* **13**:1037–1046. DOI: <https://doi.org/10.1093/scan/nsy073>, PMID: 30137604
- Boucein W, Fowles DC, Grimnes S, Ben-Shakhar G, Roth WT, Dawson ME, Filion DL, Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. 2012. Publication recommendations for electrodermal measurements. *Psychophysiology* **49**:1017–1034. DOI: <https://doi.org/10.1111/j.1469-8986.2012.01384.x>, PMID: 22680988
- Bradford DE, Starr MJ, Shackman AJ, Curtin JJ. 2015. Empirically based comparisons of the reliability and validity of common quantification approaches for eyeblink startle potentiation in humans. *Psychophysiology* **52**:1669–1681. DOI: <https://doi.org/10.1111/psyp.12545>
- Chauret M, La Buissonnière-Ariza V, Lamoureux Tremblay V, Suffren S, Servonnet A, Pine DS, Maheu FS. 2014. The conditioning and extinction of fear in youths: what's sex got to do with it? *Biological Psychology* **100**:97–105. DOI: <https://doi.org/10.1016/j.biopsycho.2014.06.001>, PMID: 24929048
- Craske MG, Kircanski K, Zelikowsky M, Mystkowski J, Chowdhury N, Baker A. 2008. Optimizing inhibitory learning during exposure therapy. *Behaviour Research and Therapy* **46**:5–27. DOI: <https://doi.org/10.1016/j.brat.2007.10.003>
- Drexler SM, Merz CJ, Hamacher-Dang TC, Tegenthoff M, Wolf OT. 2015. Effects of cortisol on reconsolidation of reactivated fear memories. *Neuropsychopharmacology* **40**:3036–3043. DOI: <https://doi.org/10.1038/npp.2015.160>, PMID: 26058664
- Drexler SM, Merz CJ, Wolf OT. 2018. Preextinction stress prevents Context-Related renewal of fear. *Behavior Therapy* **49**:1008–1019. DOI: <https://doi.org/10.1016/j.beth.2018.03.001>, PMID: 30316481
- Duits P, Cath DC, Lissek S, Hox JJ, Hamm AO, Engelhard IM, van den Hout MA, Baas JMP. 2015. Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and Anxiety* **32**:239–253. DOI: <https://doi.org/10.1002/da.22353>
- Flake JK, Fried E. 2019. Measurement schmeasurement: questionable measurement practices and how to avoid them. *PsyArXiv*. DOI: <https://doi.org/10.31234/osf.io/hs7wm>
- Garpenstrand H, Annas P, Ekblom J, Oreland L, Fredrikson M. 2001. Human fear conditioning is related to dopaminergic and serotonergic biological markers. *Behavioral Neuroscience* **115**:358–364. DOI: <https://doi.org/10.1037/0735-7044.115.2.358>, PMID: 11345960
- 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*: Columbia Statistics. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.
- Gerlicher AMV, Tüscher O, Kalisch R. 2018. Dopamine-dependent prefrontal reactivations explain long-term benefit of fear extinction. *Nature Communications* **9**:1–9. DOI: <https://doi.org/10.1038/s41467-018-06785-y>
- Gerlicher AMV, Tüscher O, Kalisch R. 2019. L-DOPA improves extinction memory retrieval after successful fear extinction. *Psychopharmacology* **236**:3401–3412. DOI: <https://doi.org/10.1007/s00213-019-05301-4>, PMID: 31243481
- Grégoire L, Greening SG. 2019. Opening the reconsolidation window using the mind's eye: Extinction training during reconsolidation disrupts fear memory expression following mental imagery reactivation. *Cognition* **183**:277–281. DOI: <https://doi.org/10.1016/j.cognition.2018.12.001>
- Grillon C, Baas JM. 2002. Comments on the use of the startle reflex in psychopharmacological challenges: impact of baseline startle on measurement of fear-potentiated startle. *Psychopharmacology* **164**:236–238. DOI: <https://doi.org/10.1007/s00213-002-1164-5>, PMID: 12481758

- Gruss LF, Keil A. 2019. Sympathetic responding to unconditioned stimuli predicts subsequent threat expectancy, orienting, and visuocortical Bias in human aversive pavlovian conditioning. *Biological Psychology* **140**:64–74. DOI: <https://doi.org/10.1016/j.biopsycho.2018.11.009>, PMID: 30476520
- Haaker J, Golkar A, Selbing I, Olsson A. 2017. Assessment of social transmission of threats in humans using observational fear conditioning. *Nature Protocols* **12**:1378–1386. DOI: <https://doi.org/10.1038/nprot.2017.027>, PMID: 28617449
- Hamacher-Dang TC, Merz CJ, Wolf OT. 2015. Stress following extinction learning leads to a context-dependent return of fear. *Psychophysiology* **52**:489–498. DOI: <https://doi.org/10.1111/psyp.12384>, PMID: 25410416
- Hamm AO, Greenwald MK, Bradley MM, Lang PJ. 1993. Emotional learning, hedonic change, and the startle probe. *Journal of Abnormal Psychology* **102**:453–465. DOI: <https://doi.org/10.1037/0021-843X.102.3.453>, PMID: 8408958
- Hamm AO, Weike AI. 2005. The neuropsychology of fear learning and fear regulation. *International Journal of Psychophysiology* **57**:5–14. DOI: <https://doi.org/10.1016/j.ijpsycho.2005.01.006>
- Hardwicke TE, Ioannidis JPA. 2018. Mapping the universe of registered reports. *Nature Human Behaviour* **2**:793–796. DOI: <https://doi.org/10.1038/s41562-018-0444-y>, PMID: 31558810
- Hardwicke TE, Shanks DR. 2016. Reply to walker and Stickgold: proposed boundary conditions on memory reconsolidation will require empirical verification. *PNAS* **113**:E3993–E3994. DOI: <https://doi.org/10.1073/pnas.1608235113>
- Hartley CA, Coelho CAO, Boeke E, Ramirez F, Phelps EA. 2019. Individual differences in blink rate modulate the effect of instrumental control on subsequent pavlovian responding. *Psychopharmacology* **236**:87–97. DOI: <https://doi.org/10.1007/s00213-018-5082-6>, PMID: 30386862
- Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. 2015. The extent and consequences of p-hacking in science. *PLOS Biology* **13**:e1002106. DOI: <https://doi.org/10.1371/journal.pbio.1002106>, PMID: 25768323
- Hermann A, Stark R, Milad MR, Merz CJ. 2016. Renewal of conditioned fear in a novel context is associated with hippocampal activation and connectivity. *Social Cognitive and Affective Neuroscience* **11**:1411–1421. DOI: <https://doi.org/10.1093/scan/nsw047>, PMID: 27053767
- Hu J, Wang W, Homan P, Wang P, Zheng X, Schiller D. 2018. Reminder duration determines threat memory modification in humans. *Scientific Reports* **8**:8848. DOI: <https://doi.org/10.1038/s41598-018-27252-0>, PMID: 29891856
- Hu J, Wang Z, Feng X, Long C, Schiller D. 2019. Post-retrieval oxytocin facilitates next day extinction of threat memory in humans. *Psychopharmacology* **236**:293–301. DOI: <https://doi.org/10.1007/s00213-018-5074-6>, PMID: 30370450
- Ioannidis JP. 2014. How to make more published research true. *PLOS Medicine* **11**:e1001747. DOI: <https://doi.org/10.1371/journal.pmed.1001747>, PMID: 25334033
- Jentsch VL, Wolf OT, Merz CJ. 2020. Temporal dynamics of conditioned skin conductance and pupillary responses during fear acquisition and extinction. *International Journal of Psychophysiology* **147**:93–99. DOI: <https://doi.org/10.1016/j.ijpsycho.2019.11.006>
- John LK, Loewenstein G, Prelec D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* **23**:524–532. DOI: <https://doi.org/10.1177/0956797611430953>, PMID: 22508865
- LeDoux J. 2012. Rethinking the emotional brain. *Neuron* **73**:653–676. DOI: <https://doi.org/10.1016/j.neuron.2012.02.004>, PMID: 22365542
- LeDoux JE. 2014. Coming to terms with fear. *PNAS* **111**:2871–2878. DOI: <https://doi.org/10.1073/pnas.1400335111>
- Leuchs L, Schneider M, Spoomaker VI. 2019. Measuring the conditioned response: a comparison of pupillometry, skin conductance, and startle electromyography. *Psychophysiology* **56**:e13283. DOI: <https://doi.org/10.1111/psyp.13283>, PMID: 30259985
- Lonsdorf TB, Menz MM, Andreatta M, Fullana MA, Golkar A, Haaker J, Heitland I, Hermann A, Kuhn M, Kruse O, Meir Drexler S, Meulders A, Nees F, Pittig A, Richter J, Römer S, Shiban Y, Schmitz A, Straube B, Vervliet B, et al. 2017. Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience & Biobehavioral Reviews* **77**:247–285. DOI: <https://doi.org/10.1016/j.neubiorev.2017.02.026>, PMID: 28263758
- Lonsdorf TB, Merz CJ, Fullana MA. 2019. Fear extinction retention: is it what we think it is? *Biological Psychiatry* **85**:1074–1082. DOI: <https://doi.org/10.1016/j.biopsych.2019.02.011>, PMID: 31005240
- Lonsdorf TB, Merz CJ. 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. *Neuroscience & Biobehavioral Reviews* **80**:703–728. DOI: <https://doi.org/10.1016/j.neubiorev.2017.07.007>
- Marin MF, Barbey F, Rosenbaum BL, Hammoud MZ, Orr SP, Milad MR. 2019. Absence of conditioned responding in humans: a bad measure or individual differences? *Psychophysiology* **57**:e13350. DOI: <https://doi.org/10.1111/psyp.13350>, PMID: 30758048
- Meir Drexler S, Merz CJ, Hamacher-Dang TC, Wolf OT. 2016. Cortisol effects on fear memory reconsolidation in women. *Psychopharmacology* **233**:2687–2697. DOI: <https://doi.org/10.1007/s00213-016-4314-x>, PMID: 27137198
- Meir Drexler S, Merz CJ, Lissek S, Tegenthoff M, Wolf OT. 2019. Reactivation of the unconditioned stimulus inhibits the return of fear independent of cortisol. *Frontiers in Behavioral Neuroscience* **13**:254. DOI: <https://doi.org/10.3389/fnbeh.2019.00254>, PMID: 31780910

- Meir Drexler S, Wolf OT. 2017. Stress disrupts the reconsolidation of fear memories in men. *Psychoneuroendocrinology* **77**:95–104. DOI: <https://doi.org/10.1016/j.psyneuen.2016.11.027>, PMID: 28024275
- Mertens G, Leer A, van Dis EAM, Vermeer L, Steenhuizen A, van der Veen L, Engelhard IM. 2019. Secondary extinction reduces reinstatement of threat expectancy and conditioned skin conductance responses in human fear conditioning. *Journal of Behavior Therapy and Experimental Psychiatry* **62**:103–111. DOI: <https://doi.org/10.1016/j.jbtep.2018.09.007>, PMID: 30296630
- Merz CJ, Hamacher-Dang TC, Wolf OT. 2014. Exposure to stress attenuates fear retrieval in healthy men. *Psychoneuroendocrinology* **41**:89–96. DOI: <https://doi.org/10.1016/j.psyneuen.2013.12.009>, PMID: 24495610
- Merz CJ, Hamacher-Dang TC, Stark R, Wolf OT, Hermann A. 2018a. Neural underpinnings of cortisol effects on fear extinction. *Neuropsychopharmacology* **43**:384–392. DOI: <https://doi.org/10.1038/npp.2017.227>, PMID: 28948980
- Merz CJ, Kinner VL, Wolf OT. 2018b. Let's talk about sex . . . differences in human fear conditioning. *Current Opinion in Behavioral Sciences* **23**:7–12. DOI: <https://doi.org/10.1016/j.cobeha.2018.01.021>
- Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLOS Medicine* **6**:e1000097. DOI: <https://doi.org/10.1371/journal.pmed.1000097>, PMID: 19621072
- Morriss J, Chapman C, Tomlinson S, van Reekum CM. 2018. Escape the bear and fall to the lion: the impact of avoidance availability on threat acquisition and extinction. *Biological Psychology* **138**:73–80. DOI: <https://doi.org/10.1016/j.biopsycho.2018.08.017>, PMID: 30144498
- Morriss J, van Reekum CM. 2019. I feel safe when i know: contingency instruction promotes threat extinction in high intolerance of uncertainty individuals. *Behaviour Research and Therapy* **116**:111–118. DOI: <https://doi.org/10.1016/j.brat.2019.03.004>, PMID: 30878772
- Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers E-J, Ware JJ, Ioannidis JPA. 2017. A manifesto for reproducible science. *Nature Human Behaviour* **1**:0021. DOI: <https://doi.org/10.1038/s41562-016-0021>
- Naveteur J, Freixa I Baque E. 1987. Individual differences in electrodermal activity as a function of subjects' anxiety. *Personality and Individual Differences* **8**:615–626. DOI: [https://doi.org/10.1016/0191-8869\(87\)90059-6](https://doi.org/10.1016/0191-8869(87)90059-6)
- Nitta Y, Takahashi T, Haitani T, Sugimori E, Kumano H. 2018. Avoidance behavior prevents modification of fear memory during reconsolidation. *Psychological Reports* **15**:33294118811116. DOI: <https://doi.org/10.1177/0033294118811116>
- Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, et al. 2015. Promoting an open research culture. *Science* **348**:1422–1425. DOI: <https://doi.org/10.1126/science.aab2374>
- Oyarzún JP, Lopez-Barroso D, Fuentemilla L, Cucurell D, Pedraza C, Rodriguez-Fornells A, de Diego-Balaguer R. 2012. Updating fearful memories with extinction training during reconsolidation: a human study using auditory aversive stimuli. *PLOS ONE* **7**:e38849. DOI: <https://doi.org/10.1371/journal.pone.0038849>, PMID: 22768048
- Oyarzún JP, Càmara E, Kouider S, Fuentemilla L, de Diego-Balaguer R. 2019. Implicit but not explicit extinction to threat-conditioned stimulus prevents spontaneous recovery of threat-potentiated startle responses in humans. *Brain and Behavior* **9**:e01157. DOI: <https://doi.org/10.1002/brb3.1157>, PMID: 30516021
- Pavlov I. 1927. *Conditioned Reflexes*. Oxford University Press.
- Pleidl W, Wotjak CT. 2010. Dissociation of within- and between-session extinction of conditioned fear. *Journal of Neuroscience* **30**:4990–4998. DOI: <https://doi.org/10.1523/JNEUROSCI.6038-09.2010>, PMID: 20371819
- Prenoveau JM, Craske MG, Liao B, Ornitz EM. 2013. Human fear conditioning and extinction: timing is everything. . . or is it? *Biological Psychology* **92**:59–68. DOI: <https://doi.org/10.1016/j.biopsycho.2012.02.005>, PMID: 22349998
- Reddan MC, Wager TD, Schiller D. 2018. Attenuating neural threat expression with imagination. *Neuron* **100**:994–1005. DOI: <https://doi.org/10.1016/j.neuron.2018.10.047>, PMID: 30465766
- Schiller D, Monfils MH, Raio CM, Johnson DC, LeDoux JE, Phelps EA. 2010. Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* **463**:49–53. DOI: <https://doi.org/10.1038/nature08637>, PMID: 20010606
- Schiller D, Monfils MH, Raio CM, Johnson DC, LeDoux JE, Phelps EA. 2018. Addendum: preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* **562**:E21. DOI: <https://doi.org/10.1038/s41586-018-0405-7>, PMID: 30050064
- Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, Bahnik Š, Bai F, Bannard C, Bonnier E, Carlsson R, Cheung F, Christensen G, Clay R, Craig MA, Dalla Rosa A, Dam L, Evans MH, Flores Cervantes I, Fong N, et al. 2018. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science* **1**:337–356. DOI: <https://doi.org/10.1177/2515245917747646>
- Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**:1359–1366. DOI: <https://doi.org/10.1177/0956797611417632>, PMID: 22006061
- Sjouwerman R, Niehaus J, Kuhn M, Lonsdorf TB. 2016. Don't startle me-Interference of startle probe presentations and intermittent ratings with fear acquisition. *Psychophysiology* **53**:1889–1899. DOI: <https://doi.org/10.1111/psyp.12761>, PMID: 27628268

- Sjouwerman R**, Scharfenort R, Lonsdorf TB. 2018. Individual differences in fear learning: specificity to trait-anxiety beyond other measures of negative affect and mediation via amygdala activation. *bioRxiv*. DOI: <https://doi.org/10.1101/233528>
- Sjouwerman R**, Lonsdorf TB. 2019. Latency of skin conductance responses across stimulus modalities. *Psychophysiology* **56**:e13307. DOI: <https://doi.org/10.1111/psyp.13307>, PMID: 30461024
- Spielberger CD**, Gorsuch RL, Lushene RE. 1983. *Manual for the State-Trait Anxiety Inventory*. Consulting Psychologists Press.
- Steege S**, Tuerlinckx F, Gelman A, Vanpaemel W. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**:702–712. DOI: <https://doi.org/10.1177/1745691616658637>, PMID: 27694465
- Tabbert K**, Merz CJ, Klucken T, Schweckendiek J, Vaitl D, Wolf OT, Stark R. 2011. Influence of contingency awareness on neural, electrodermal and evaluative responses during fear conditioning. *Social Cognitive and Affective Neuroscience* **6**:495–506. DOI: <https://doi.org/10.1093/scan/nsq070>, PMID: 20693389
- Tani H**, Tada M, Maeda T, Konishi M, Umeda S, Terasawa Y, Mimura M, Takahashi T, Uchida H. 2019. Comparison of emotional processing assessed with fear conditioning by interpersonal conflicts in patients with depression and schizophrenia. *Psychiatry and Clinical Neurosciences* **73**:116–125. DOI: <https://doi.org/10.1111/pcn.12805>, PMID: 30499148
- Taylor VA**, Roy M, Chang L, Gill LN, Mueller C, Rainville P. 2018. Reduced Fear-Conditioned pain modulation in experienced meditators: a preliminary study. *Psychosomatic Medicine* **80**:799–806. DOI: <https://doi.org/10.1097/PSY.0000000000000634>, PMID: 30134359
- Tuominen L**, Boeke E, DeCross S, Wolthusen RP, Nasr S, Milad M, Vangel M, Tootell R, Holt D. 2019. The relationship of perceptual discrimination to neural mechanisms of fear generalization. *NeuroImage* **188**:445–455. DOI: <https://doi.org/10.1016/j.neuroimage.2018.12.034>, PMID: 30572112
- Vervliet B**, Craske MG, Hermans D. 2013. Fear extinction and relapse: state of the art. *Annual Review of Clinical Psychology* **9**:215–248. DOI: <https://doi.org/10.1146/annurev-clinpsy-050212-185542>, PMID: 23537484
- Weissgerber TL**, Milic NM, Winham SJ, Garovic VD. 2015. Beyond bar and line graphs: time for a new data presentation paradigm. *PLOS Biology* **13**:e1002128. DOI: <https://doi.org/10.1371/journal.pbio.1002128>, PMID: 25901488
- Wendt J**, Neubert J, Lindner K, Ernst FD, Homuth G, Weike AI, Hamm AO. 2015. Genetic influences on the acquisition and inhibition of fear. *International Journal of Psychophysiology* **98**:499–505. DOI: <https://doi.org/10.1016/j.ijpsycho.2014.10.007>, PMID: 25455425
- Wendt J**, Hufenbach MC, König J, Hamm AO. 2020. Effects of verbal instructions and physical threat removal prior to extinction training on the return of conditioned fear. *Scientific Reports* **10**:1202. DOI: <https://doi.org/10.1038/s41598-020-57934-7>
- Wickham H**. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. DOI: <https://doi.org/10.1007/978-0-387-98141-3>

Appendix 1—table 1. Summary of criteria used to define ‘non-learners’ across records included in the systematic literature search. Criteria used to define ‘non-learners’ were identified in eleven records reported in a total of 14 individual studies.

Reference	% excluded participants ('non-learners')	CS+/CS- cut-off (in μ S) for 'non-learners'	N trials _{acq} total CS+/CS-	N trials _{acq} considered CS+/CS-	Trials _{phase} (unless otherwise stated, this refers to fear acquisition training)	Additional criteria/notes
Ahmed and Lovibond, 2019 , Exp. 1	24%	<or = 0	3/3	2/2	last two thirds	only considered as 'non-learners' if applicable to both SCL and ratings
Ahmed and Lovibond, 2019 , Exp. 2	16%					
Reddan et al., 2018^a	35%	<or = 0	16/8	8 ^b /8	full phase	
Grégoire and Greening, 2019	16%	<0.1	13 ^c /8	4 ^c /4	last third	participants were also excluded if they did not show equivalent responding to both CS+s (difference > 0.1 μ S) or when not showing equal extinction to both CS+s or complete differential extinction to both CS+s vs. CS- (difference > 0.1 μ S)
Hu et al., 2018	6%	<or = 0	16/10	5/5 ^d OR 1/1	second half ^d OR last trial ^d	'non-learners' discontinued after day 1 of the experiment
Oyarzún et al., 2019 , Exp. 1	27%	<or = 0 ^e	eight ^c /8	4 ^c /4	second half	only considered as 'non-learners' if applicable to both SCR and fear-potentiated startle
Oyarzún et al., 2019 , Exp. 2	41%					
Belleau et al., 2018	2%	<0.05	5/5	5/5	full phase	only considered as 'non-learners' when also failing to show any differential ratings (i.e., <or = 0 in discrimination) ^f
Morriss et al., 2018	6% g	< or = 0 ^h	12/6	6 ^b /6	full phase	only considered as 'non-learners' if applicable across all phases (fear acquisition and extinction training, avoidance acquisition and extinction)

Appendix 1—table 1 continued on next page

Appendix 1—table 1 continued

Reference	% excluded participants ('non-learners')	CS+/CS- cut-off (in μ S) for 'non-learners'	N trials _{acq} total CS+/CS-	N trials _{acq} considered CS+/CS-	Trials _{phase} (unless otherwise stated, this refers to fear acquisition training)	Additional criteria/notes
Schiller et al., 2018; Schiller et al., 2010, Exp. 1	48% ⁱ	<0.1/mean SCR to the US	16 ^c /10 13 ^c /8	5 ^b /5 OR 5 ^b /5 OR 1 ^b /1 OR increase from first to last trial	first half of acquisition OR second half OR last trial of acquisition, OR the increase from the first to last trial of acquisition	
Schiller et al., 2018; Schiller et al., 2010, Exp. 2	74% ⁱ			4 ^b /4 OR 4 ^b /4 OR 1 ^b /1 OR increase from first to last trial		
Nitta et al., 2018	52%	<0	13 ^c /8	2 ^c /2	last two trials	one additional participant showed strong SCR during re-extinction phase to CS+ and was therefore excluded
Hartley et al., 2019	16%	<or = 0.05	6/6	3/3	last half	
Hu et al., 2019	16%	< 0 ^k	8 ^b /8	4 ^b /4	second half	

^aPersonal communication with D. Schiller (20.5.2019 and 30.8.2019) confirmed that individuals were classified as 'non-learners' when they 'did not demonstrate greater SCRs to the CS+ relative to the CS- on average across all acquisition trials (n = 24)' (see 'Materials and methods' section). The personal communication clarified that the statement included in the results section that defines 'non-learners' as individuals that "did not demonstrate a discriminatory SCR during acquisition, defined as greater SCR to the CS+ relative to the CS- during either the first or last half of threat-acquisition on average" was intended to refer to the same procedure (i.e., the full acquisition phase).

^bRefers to unreinforced CS+ trials (CS+ trials not followed by the US) only.

^cFor each CS+1 and CS+2.

^dPersonal communication with D. Schiller (1.5.2019): late acquisition as reported in the publication refers to the last half or last trial. Anyone that had a positive difference (>0.000 μ S) in either the second half or last trial of acquisition was kept.

^ePersonal communication with J. Oyarzun (21.5.2019): all difference scores > 0 μ S were considered as CS+/CS- discrimination.

^fPersonal communication with E. Balleau, PhD (5.5.2019): 'differential ratings' means that CS+>CS- is equal to or below 0 μ S was non-differentiation.

^gThese were not excluded as results did not change.

^hPersonal communication with J. Morriss (15.4.2019): no positive differential response is defined as any number <0 μ S.

ⁱPercentages for 'non-learners', 'non-extinguishers' and 'non-responders' reported together.

^kPersonal communication with D. Schiller (21.5.2019): zero differences were kept.

Appendix 1

Definition of performance-based exclusion of participants ('non-learners') and numbers of participants excluded across studies

Appendix 1—table 2. Summary of criteria used to define 'non-responders' across records included in the systematic literature search. Fifteen records, reporting a total of 17 studies, were identified.

Record	% excluded participants ('non-responders')	Cut-off (in μ S) for a valid SCR	Valid responses in at least % of trials	Stimulus type (also referred to as 'trial') on which the exclusion is based	Additional criteria/notes
<i>Baeuchi et al., 2019</i>	10%	>0.01	≥66%	CXT+	
<i>Tuominen et al., 2019</i>	12%	>0.05	≥13%	CS+ and CS-	
<i>Gruss and Keil, 2019</i>	11%	visual inspection ^a		CS+, CS- and US	
<i>Sjouwerman and Lonsdorf, 2019</i>	14%	≥0.02	US: ≥67% CS: no valid response in each CS modality	CS+, CS- and US	
<i>Grégoire and Greening, 2019</i>	8%	>0.02	≥25% ^b	CS+ and CS- ^c	
<i>Hu et al., 2018</i>	3%	≥0.02	100%	CS+ and CS- ^c	'non-responders' discontinued after day 1 of the experiment
<i>Oyarzún et al., 2019, Exp. 1</i>	0%	≥0.02	≥25%	CS+ ^d and CS- ^c	
<i>Oyarzún et al., 2019, Exp. 2</i>	9%				
<i>Tani et al., 2019</i>	10%	>0.03 ^e	100%	CS+	
<i>Marin et al., 2019</i>	0% ^f	≥0.03	≥10%	US	
<i>Taylor et al., 2018</i>	5%	NA	100%	motor test ^g	
<i>Morriss et al., 2018</i>	6%	>0.03	≥90%	CS+ ^d and CS- ^c	only applicable if true across all phases/days of the experiment
<i>Schiller et al., 2018; Schiller et al., 2010, Exp. 1</i>	48% ^h	≥0.02	≥25%	CS+ ^d and CS-	
<i>Schiller et al., 2018; Schiller et al., 2010, Exp. 2</i>	74% ^h				

Appendix 1—table 2 continued on next page

Appendix 1—table 2 continued

Record	% excluded participants ('non-responders')	Cut-off (in μ S) for a valid SCR	Valid responses in at least % of trials	Stimulus type (also referred to as 'trial') on which the exclusion is based	Additional criteria/notes
Morriss and van Reekum, 2019 , Exp. 1	2%	>0.03	>90%	CS+ ^d and CS-	
Morriss and van Reekum, 2019 , Exp. 2	2%				
Hartley et al., 2019	6%	<0.05 ⁱ	$\geq 33\%$ ⁱ	CS+ and CS- ^c	
Hu et al., 2019	4% ^k	≥ 0.02	100% ^k	US	
Leuchs et al., 2019	4%	NA	$\geq 33\%$	CS+ and CS- ^c	only applicable if true across both days of the experiment

^aPersonal communication with L. Forest Gruss (29.4.2019): "the determination of non-responders was done, this was done on visual inspection by me through all trials of all individuals. I verified after determining who the lowest, i.e. non-responders were, in the same fashion as the startle non-responders in summing responding over the entire experiment, and this responding falling below a threshold of overall response ($\sim 10\%$) AND one individual due to lack of response at the end of the trial to the UCS specifically".

^bPersonal communication with S.G. Greening (24.4.2019): "non-responders if more than 75% of data were missing (i.e., SCR <0.02 μ S) during the training phase. So, that means, if a participant had at least six trials (out of 24) with measurable SCRs (whatever the condition), we kept them (if the other acquisition criteria were OK, see below). If they had five trials or fewer with measurable GSR, we considered them a non-responder and removed them".

^cPersonal communications that 'trial' or this statement refers to CS+ and CS- trials: S. Greening (24.4.2019), D. Schiller (1.5.2019), J. Oyarzun (21.5.2019), J. Morriss (15.4.2019), C. Hartley (2.5.2019), V. Spoomaker (18.4.2019).

^d CS+ unpaired.

^ePersonal communication with H. Tani (2.5.2019): only CS+ trials were considered (here as response to the sound or the intrapersonal stimulus).

^fPersonal communication with M.-F. Marin (23.4.2019): exclusion criteria were defined, but no participant met these criteria and hence none was excluded.

^gPersonal communication with V. Taylor (6.6.2019): clarified that "non-responders' were identified in a "motor test of SCR responding during the preliminary session. Essentially, they had to compress a ball with the right hand with maximal physical force for a few seconds on about 10 trials, which typically elicits quite large SCRs in subjects. Failure to respond to an SCR to all of these trials was considered a non-responder".

^hPercentages for 'non-learners', 'non-extinguishers' and 'non-responders' reported together.

ⁱPersonal communication with C. Hartley (2.5.2019): clarified that "participants were considered non-responder if they had SCR values of 0 for more than 8 of the 12 trials in acquisition (<4 responsive trials)".

^k The percentage of 'non-responders' and 'non-learners' was reported together without percentages for each category; personal communication with D. Schiller (21.5.2019): in the paper, it is reported that five individuals 'were excluded due to equipment malfunction (N = 2) or had non-measurable skin conductance response (SCR) to the shock (N = 3)". It was confirmed that these individuals excluded for non-measurable SCR did not show any responses to any stimulus.

Appendix 2

Applying the identified performance-based exclusion criteria to existing data: a case example

In this case example based on Data set 1 (see main manuscript), we tested whether CS+/CS– discrimination in SCRs does indeed differ between the different exclusion groups as defined by the cut-offs retrieved from the literature (see **Figure 2B**). Note that this is somewhat circular as exclusion groups are defined by different SCR CS+/CS– cutoffs, which then are used in an analysis in which differential SCRs are the dependent measure. However, that this is exactly what is sometimes done in the literature (see main manuscript).

Still, this is an important manipulation check to test empirically whether those classified in a group of ‘non-learners’ in the literature do indeed show no evidence of learning, which would be indicated by comparable SCRs to the CS+ and the CS– (i.e., no significant discrimination). Here, we test this for cumulative exclusion groups. Note that this is only a rough manipulation check, as a non-significant CS+/CS– discrimination effect in the whole group (e.g., those showing a CS+/CS– discrimination $<0.05 \mu\text{S}$ based on raw scores) cannot be taken as evidence that all individuals in this group do not display meaningful or statistically significant CS+/CS– discrimination. More precisely, half of this group who did not meet the cut-off of $0.05 \mu\text{S}$ in CS+/CS– discrimination do show a negative or zero discrimination score, which may bias the group average score towards non-discrimination. Yet, statistically testing for discrimination within each exclusion group (e.g. specifically in the group showing a discrimination between >0 and $< 0.05 \mu\text{S}$) is not unproblematic.

Appendix 2—table 1. Results of two-tailed t-tests for differences in SCR CS+/CS– discrimination in Data set 1 for the different cumulative exclusion groups (indicated by the + in the table) based on the criteria identified in the literature with respect to CS+/CS– discrimination cutoffs (in μS). For completeness sake and as it is not always clear whether CS+/CS– discrimination is based on raw or transformed values, we report results based on analyses of both raw (A) and transformed values (B). P-values for these post-hoc tests are Bonferroni corrected.

A) t-tests: CS+/CS– discrimination based on raw values

Exclusion group (cumulative)	CS+ M (SD)	CS– M (SD)	df	t	P _{bonf_corr}	d
<0	0.04 (0.04)	0.07 (0.07)	10	–2.67	.140	0.81
+ = 0	0.02 (0.04)	0.03 (0.05)	33	–2.24	.193	0.38
+ > 0 and < 0.05	0.04 (0.05)	0.03 (0.05)	66	2.14	.219	0.26
+ = 0.05	0.04 (0.05)	0.03 (0.05)	70	2.88	.031	0.34
+ > 0.05 and < 0.1	0.06 (0.06)	0.04 (0.05)	88	5.87	.0000005	0.62
+ \geq 0.1	0.10 (0.10)	0.04 (0.06)	115	7.87	<0.000000001	0.73

B) t-tests: CS+/CS– discrimination based on log-transformed and range-corrected values

Exclusion group (cumulative)	CS+ M (SD)	CS– M (SD)	df	t	P _{bonf_corr}	d
<0	0.09 (0.10)	0.13 (0.11)	13	–3.46	0.025	0.93
+ = 0	0.04 (0.08)	0.06 (0.10)	28	–2.90	0.043	0.54
+ > 0 and < 0.05	0.06 (0.10)	0.07 (0.11)	42	–0.88	>0.999	0.13
+ = 0.05	0.07 (0.10)	0.07 (0.11)	46	–0.06	>0.999	0.01
+ > 0.05 and < 0.1	0.09 (0.11)	0.07 (0.11)	60	2.81	.040	0.36
+ \geq 0.1	0.21 (0.19)	0.10 (0.11)	115	9.56	<0.000000001	0.89

Appendix 3

Exploratory analyses on consistency of classification ('learners' vs. 'non-learners') across outcome measures and criteria employed

Throughout the main manuscript and particularly in the discussion, we highlight that differential (CS+>CS-) SCRs alone cannot be taken to infer 'learning' (**Figure 4—figure supplement 1**).

Appendix 3—table 1 provides statistical information on CS+/CS- discrimination in fear ratings in (cumulative) exclusion groups as defined by CS+/CS- discrimination in SCRs.

Appendix 3—table 1. CS+/CS- discrimination in fear ratings in (cumulative) exclusion groups (indicated by the + in the table) as defined by CS+/CS- discrimination in SCRs (based on raw scores).

Exclusion group (cumulative)	CS+ M (SD)	CS- M (SD)	df	t	P _{bonf_corr}	d
<0	15.8 (8.94)	2.45 (4.70)	10	5.37	0.002	1.62
+ = 0	16.6 (7.73)	3.15 (5.82)	31	9.69	<0.000000001	1.71
+ > 0 and < 0.05	16.2 (7.37)	3.06 (5.86)	64	12.8	<0.000000001	1.59
+ = 0.05	16.3 (7.26)	2.96 (5.75)	67	13.4	<0.000000001	1.62
+ > 0.05 and < 0.1	16.5 (6.97)	2.94 (5.47)	84	16.0	<0.000000001	1.74
+ >= 0.1	17.3 (6.64)	3.08 (5.04)	110	20.2	<0.000000001	1.92

Appendix 4—table 1. Overview of SCR response quantification specifications (i.e., min. amplitude, scoring approach) and procedural details during fear acquisition training (i.e., number of CS and US presentations) as well as number (mean and range) and percentage of SCR non-responses towards the different stimuli (US, CS+, CS-, CS).

TTP: trough-to-peak; CS+E: CS+ extinguished; CS+U: CS+ unextinguished, CS: for both the CS+ and CS-.

Reference	N	Minimum amplitude cutoff (in μ S) for valid SCRs	Scoring details	Number of ...		'Non-responses' towards ...							
				US	CS (CS+/CS-)	US ($M \pm SD$, range)	US (%)	CS+ ($M \pm SD$, range)	CS+ (%)	CS- ($M \pm SD$, range)	CS- (%)	CS ($M \pm SD$, range)	CS (%)
Jentsch et al., 2020	41	≥ 0.02	TTP (max peak), latency 0.5–4 s / 1–80.5 s (US/CS)	10	16/16	1.12 \pm 1.66 (0–10)	11.22	2.22 \pm 3.31 (0–16)	13.87	4.49 \pm 3.92 (0–16)	28.05	6.71 \pm 6.68 (0–32)	20.96
Hermann et al., 2016	45	≥ 0.02	TTP (max peak), latency 0.5–6 s / 1–60.5 s (US/CS)	10 (5 for CS+E, 5 for CS+U)	8 CS+E/8 CS+U/16 CS-	0.24 \pm 0.88 (0–5)	2.44	2.64 \pm 3.49 (0–13); CS+E: 1.47 \pm 2.19 (0–8); CS+U: 1.18 \pm 1.80 (0–7)	16.53 CS+E: 18.33; CS+U: 14.72	8.07 \pm 4.14 (0–16)	50.42	10.71 \pm 6.65 (0–26)	33.47
Merz et al., 2018a	39	≥ 0.02	TTP (max peak), latency 0.5–6 s / 1–60.5 s (US/CS)	10 (5 for CS+E, 5 for CS+U)	8 CS+E/8 CS+U/8 CS-	2.08 \pm 1.98 (0–8)	20.77	3.36 \pm 4.55 (0–16); CS+E: 1.59 \pm 2.35 (0–8); CS+U: 1.77 \pm 2.32 (0–8)	21.00; 19.87; CS+U: 22.12	2.41 \pm 2.27 (0–8)	30.13	5.77 \pm 6.49 (0–24)	24.04
Merz et al., 2014	40	≥ 0.02	TTP (max peak), latency 0.5–6 s / 1–60.5 s (US/CS)	10 (5 for CS+E, 5 for CS+U)	8 CS+E/8 CS+U/16 CS-	0.13 \pm 0.33 (0–1)	1.25	1.08 \pm 2.04 (0–11); CS+E: 0.58 \pm 1.08 (0–5); CS+U: 0.50 \pm 1.11 (0–6)	6.72; CS+E: 7.19 CS+U: 6.25	3.13 \pm 2.96 (0–11)	19.53	4.20 \pm 4.39 (0–21)	13.13
Hamacher-Dang et al., 2015	39	≥ 0.02	TTP (max peak), latency 0.5–6 s / 1–60.5 s (US/CS)	10 (5 for CS+E, 5 for CS+U)	8 CS+E/8 CS+U/16 CS-	0.23 \pm 0.48 (0–2)	2.31	2.33 \pm 3.77 (0–12); CS+E: 1.31 \pm 2.21 (0–8); CS+U: 1.03 \pm 1.81 (0–7)	14.58; CS+E: 16.35; CS+U: 12.82	3.77 \pm 4.20 (0–14)	23.56	6.10 \pm 7.71 (0–26)	19.07
Mertens et al., 2019	59	≥ 0.02	TTP (max peak), latency 1–8 s (baseline 0–2 s)	10	10/5	0.78 \pm 1.69 (0–6)	7.8	4.75 \pm 2.97 (0–10)	47.5	2.93 \pm 1.66 (0–5)	58.6	7.68 \pm 4.30 (0–15)	51.2
Klingel-höfer-Jens et al., unpublished	119	≥ 0.01	TTP (first peak), latency 0.9–2.5 s / 3.5 s (US/CS)	14	14/14	1.40 \pm 2.47 (0–14)	10.0	5.30 \pm 4.42 (0–14)	37.9	8.20 \pm 3.99 (0–14)	58.6	6.75 \pm 4.44 (0–14)	48.2
Gerlicher et al. unpublished	52	≥ 0.02	TTP (first peak), latency 0.9–4 s	6	6/6	0.73 \pm 1.39 (0–6)	12.18	2.73 \pm 2.06 (0–6)	45.5	3.54 \pm 1.82 (0–6)	59.0	6.27 \pm 3.54 (0–12)	52.24

Appendix 4—table 1 continued on next page

Appendix 4—table 1 continued

Reference	N	Minimum amplitude cutoff (in μ S) for valid SCRs	Scoring details	Number of ...		'Non-responses' towards ...							
				US	CS (CS+/CS-)	US ($M \pm SD$, range)	US (%)	CS+ ($M \pm SD$, range)	CS+ (%)	CS- ($M \pm SD$, range)	CS- (%)	CS ($M \pm SD$, range)	CS (%)
Gerlicher et al., 2018	39	≥ 0.02	TTP (first peak) latency 0.9–4 s	5	10/10	0.33 \pm 0.93 (0–5)	6.67	1.05 \pm 2.21 (0–10)	10.51	2.36 \pm 2.49 (0–10)	23.59	3.41 \pm 4.48 (0–20)	17.05
Andreatta et al. unpublished	76	≥ 0.02	TTP (first peak) latency 0.8–4 s	16 (8 in analysis due to startle probes)	16/16 (8/8 in analysis due to startle probes)	1.34 \pm 1.69 (0–8)	16.78	4.17 \pm 2.30 (0–8)	52.14	5.00 \pm 1.98 (0–8)	62.50	9.17 \pm 3.77 (0–16)	57.32
Wendt et al., 2020	112	≥ 0.04	TTP (first peak), latency 0.9–4 s	9	12/12	0.46 \pm 1.15 (0–7)	5.06	5.88 \pm 3.63 (0–12)	48.96	7.06 \pm 3.19 (0–12)	58.85	12.94 \pm 6.39 (0–24)	53.91
Wendt et al., 2015	108	≥ 0.04	TTP (first peak), latency 0.9–4 s	12	12/12	0.27 \pm 0.99 (0–8)	2.24	6.44 \pm 3.81 (0–12)	53.63	8.53 \pm 2.65 (0–12)	71.06	14.96 \pm 6.04 (0–24)	62.35
Drexler et al., 2015	46	≥ 0.02	TTP (max peak), latency 1–4.5 s	18	13 CS1+/13 CS2+/13 CS-	2.8 \pm 4.18 (0–16)	15.57	9.67 \pm 7.64 (0–26); CS1+: 4.87 \pm 4.07 (0–13); CS2+: 4.80 \pm 3.78 (0–13)	37.20; CS1+:37.45; CS2+: 36.95	5.26 \pm 3.95 (0–13)	40.46	14.93 \pm 11.37 (0–39)	38.29
Meir Drexler et al., 2016	73	≥ 0.02	TTP (max peak), latency 1–4.5 s	18	13 CS1+/13 CS2+/13 CS-	3.37 \pm 4.72 (0–18)	18.72	11.51 \pm 7.96 (0–25); CS1+: 5.78 \pm 3.97 (0–13); CS2+: 5.73 \pm 4.22 (0–13)	44.25; CS1+:44.67; CS2+:44.04	6.29 \pm 3.94 (0–13)	48.36	17.79 \pm 11.67 (0–37)	45.62
Meir Drexler and Wolf, 2017	72	≥ 0.02	TTP (max peak), latency 1–4.5 s	18	13 CS1+/13 CS2+/13 CS-	1.92 \pm 2.96 (0–11)	10.64	9.65 \pm 7.21 (0–25); CS1+: 4.78 \pm 3.72 (0–12); CS2+: 4.88 \pm 3.85 (0–13)	37.12; CS1 +: 36.75; CS2+: 37.50	5.42 \pm 3.54 (0–12)	41.66	15.07 \pm 10.40 (0–36)	38.63
Drexler et al., 2018	40	≥ 0.02	TTP (max peak), latency 1–4.5 s	10 (5 for CS+E, 5 for CS+U)	8 CS+E/8 CS+U/16 CS-	0.32 \pm 0.69 (0–3)	3.25	4.17 \pm 4.45 (0–16); CS+E: 2.02 \pm 2.47 (0–8); CS+U: 2.15 \pm 2.38 (0–8)	26.09; CS +E: 25.31; CS+U: 26.87	6.07 \pm 4.37 (0–16)	37.96	10.25 \pm 8.24 (1–27)	32.03
Meir Drexler et al., 2019	75	≥ 0.02	TTP (max peak), latency 0.5–6 s/ 1–80.5 s (US/CS)	6	10/10	0.89 \pm 01.57 (0–6)	14.88	4.07 \pm 3.40 (0–10)	40.66	4.68 \pm 3.23 (0–10)	46.8	8.75 \pm 6.41 (0–20)	43.73

Appendix 4—table 1 continued on next page

Appendix 4—table 1 continued

Reference	N	Minimum amplitude cutoff (in μ S) for valid SCRs	Scoring details	Number of...		'Non-responses' towards...							
				US	CS (CS+/CS-)	US ($M \pm SD$, range)	US (%)	CS+ ($M \pm SD$, range)	CS+ (%)	CS- ($M \pm SD$, range)	CS- (%)	CS ($M \pm SD$, range)	CS (%)
Chalkia et al., unpublished	238	≥ 0.02	TTP (first peak), latency 0.5–4.5 s	6	16/10 (10/10 in analysis, only unrein-forced trials)	0 (0–6)	0	0.03 \pm 0.19 (0–10)	0.29	0.05 \pm 0.29 (0–10)	0.50	0.08 \pm 0.42 (0–20)	0.40
Hollandt et al., unpublished	30	> 0.04	TTP (first peak), latency 0.9–4 s	6	10/10	0	0	2.97 \pm 2.81 (0–10)	29.67	7.23 \pm 2.61 (0–10)	72.33	10.20 \pm 4.72	51.0
Sjouwerman et al., 2018	326	≥ 0.02	TTP (first peak), latency 0.9–4.5 s	9	9/9	1.38 \pm 1.73 (0–9)	15.37	3.11 \pm 2.69 (0–9)	34.59	3.77 \pm 2.68 (0–9)	41.92	6.87 \pm 5.01 (0–18)	38.26

Appendix 4

Definition of ‘non-responders’ and amount of participants excluded across studies

In the main manuscript, we discuss different frequencies of ‘non-responding’ to different experimental stimuli (e.g., US, CS+ and CS– in isolation or in combination), which inherently lead to different exclusion frequencies when classifying ‘non-responders’ on the basis of different types of stimuli. As there is little empirical work on the frequency of ‘non-responses’ to the US, CSs (i.e., CS+ and CS–) and CS+ only to base recommendations on, we compiled this information across 20 different data sets (see **Appendix 4—table 1**), including information on SCR response quantification specifications (i.e., minimum amplitude, scoring approach) and procedural details during fear acquisition training (i.e., number of CS and US presentations). These data sets were provided by different co-authors involved in this manuscript.

In addition, **Appendix 4—table 2** provides information on the number and percentage of individuals in a sample showing SCR ‘non-responses’ to a certain number of US presentations during fear acquisition training as well as mean number and percentage of CS responses (CS refers to the CS+ and CS– combined) in these individuals to guide the development of empirically based criteria to define SCR ‘non-responders’.

Appendix 4—table 2. Number and percentage of individuals in a sample showing SCR non-responses to a certain number of US presentations during fear acquisition training (exemplarily for one to eight USs[#]), as well as mean number of and percentage of CS responses (CS refers to the CS+ and CS– combined) in these individuals. [#]Here only up to eight USs are included as eight is half of the maximum number of US presentations in the samples included here.

Reference	a) n (%) of individuals with 0, 1, 2, 3, 4, 5, 6, 7, and 8 SCRs towards the US. b) M (%) of valid CS responses for these individuals.								
	0 US	1 US	2 US	3 US	4 US	5 US	6 US	7 US	8 US
<i>Jentsch et al., 2020</i>	a) 1 (2.4%) b) 0 (0%)	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 2 (4.9%) b) 27.5 (85.9%)	a) 7 (17.1%) b) 25.4 (79.5%)
<i>Hermann et al., 2016</i>	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 1 (2%) b) 12 (37.5%)	a) 0 (0%) b) NA	a) 1 (2%) b) 14 (43.7%)	a) 0 (0%) b) NA
<i>Merz et al., 2018a</i>	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 1 (2.6%) b) 23.0 (95.8%)	a) 0 (0%) b) NA	a) 2 (5.1%) b) 20.0 (83.3%)	a) 3 (7.7%) b) 23.0 (95.8%)	a) 1 (2.6%) b) 21.0 (87.5%)	a) 5 (12.8%) b) 21.4 (89.1%)	a) 9 (23.1%) b) 21.6 (85.6%)
<i>Merz et al., 2014</i>	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA
<i>Hamacher-Dang et al., 2015</i>	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 1 (3%) b) 24 (75.0%)
<i>Mertens et al., 2019</i>	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 4 (6.78%) b) 1.75 (11.67%)	a) 0 (0%) b) NA	a) 2 (3.39%) b) 3.5 (23.33%)	a) 2 (3.39%) b) 9 (60%)	a) 0 (0%) b) NA
<i>Klingelhöfer-Jens et al., unpublished</i>	a) 2 (1.68%) b) 0 (0%)	a) 0 (0%) b) NA	a) 1 (0.84%) b) 10 (35.7%)	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 1 (0.84%) b) 1 (3.57%)	a) 2 (1.68%) b) 2 (7.14%)	a) 1 (0.84%) b) 0 (0%)

Appendix 4—table 2 continued on next page

Appendix 4—table 2 continued

a) n (%) of individuals with 0, 1, 2, 3, 4, 5, 6, 7, and 8 SCRs towards the US. b) M (%) of valid CS responses for these individuals.

Reference	0 US	1 US	2 US	3 US	4 US	5 US	6 US	7 US	8 US
Gerlicher et al., unpublished	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 3 (5.77%) b) 4 (33.33%)	a) 5 (9.62%) b) 4.8 (40%)	a) 7 (13.46%) b) 6.7 (55.91%)	a) 35 (67.31%) b) 6.15 (51.25%)	NA	NA
Gerlicher et al., 2018	a) 1 (2.56%) b) 0 (0%)	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 2 (5.13%) b) 19.5 (97.5%)	a) 4 (10.26%) b) 17.5 (87.50%)	a) 32 (82.05%) b) 16.81 (84.05%)	NA	NA	NA
Wendt et al., 2020	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 1 (0.9%) b) 18 (75%)	a) 1 (0.9%) b) 24 (100%)	a) 1 (0.9%) b) 0 (0%)	a) 0 (0%) b) NA	a) 2 (1.8%) b) 12 (50%)	a) 8 (7.1%) b) 13.13 (54.69%)	a) 11 (9.9%) b) 11.09 (46.21%)
Wendt et al., 2015	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 1 (0.9%) b) 18 (75%)	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 1 (0.9%) b) 17 (70.83%)	a) 0 (0%) b) NA
Drexler et al., 2015	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 2 (4.3%) b) 0.5 (1.28%)	a) 1 (2.2%) b) 0.0 (0.0%)	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 1 (2.2%) b) 7.0 (17.94%)	a) 0 (0%) b) NA
Meir Drexler et al., 2016	a) 1 (1.4%) b) 29.00 (74.35%)	a) 0 (0%) b) NA	a) 2 (2.7%) b) 2.0 (5.12%)	a) 1 (1.4%) b) 9.0 (23.07%)	a) 1 (1.4%) b) 2.0 (5.12%)	a) 1 (1.4%) b) 3.0 (7.69%)	a) 0 (0%) b) NA	a) 4 (5.5%) b) 5.0 (12.82%)	a) 2 (2.7%) b) 6.50 (16.66%)
Meir Drexler and Wolf, 2017	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 1 (1.4%) b) 5.0 (12.82%)	a) 1 (1.4%) b) 5.0 (12.82%)
Drexler et al., 2018	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 1 (2.5%) b) 8 (25%)	a) 2 (5.0%) b) 12.5 (39.06%)
Meir Drexler et al., 2019	a) 3 (4.0%) b) 0.33 (1.66%)	a) 1 (1.3%) b) 1 (5.0%)	a) 4 (5.3%) b) 4.25 (21.25%)	a) 2 (2.7%) b) 3.0 (15.0%)	a) 3 (4.0%) b) 1.33 (6.66%)	a) 19 (25.3%) b) 12.63 (63.15%)	a) 43 (57.3%) b) 13.21 (66.04%)	a) 0 (0%) b) NA	a) 0 (0%) b) NA
Chalkia et al., unpublished	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 238 (100%) b) 19.92 (99.6%)	a) 0 (0%) b) NA	a) 0 (0%) b) NA
Hollandt et al., unpublished	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	NA	NA
Sjouwerman et al., 2018	a) 4 (1.23%) b) 0.5 (2.78%)	a) 2 (0.61%) b) 2.5 (13.89%)	a) 4 (1.23%) b) 4.13 (22.92%)	a) 2 (0.61%) b) 7.25 (40.28%)	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA	a) 0 (0%) b) NA



Figures and figure supplements

Navigating the garden of forking paths for data exclusions in fear conditioning research

Tina B Lonsdorf et al

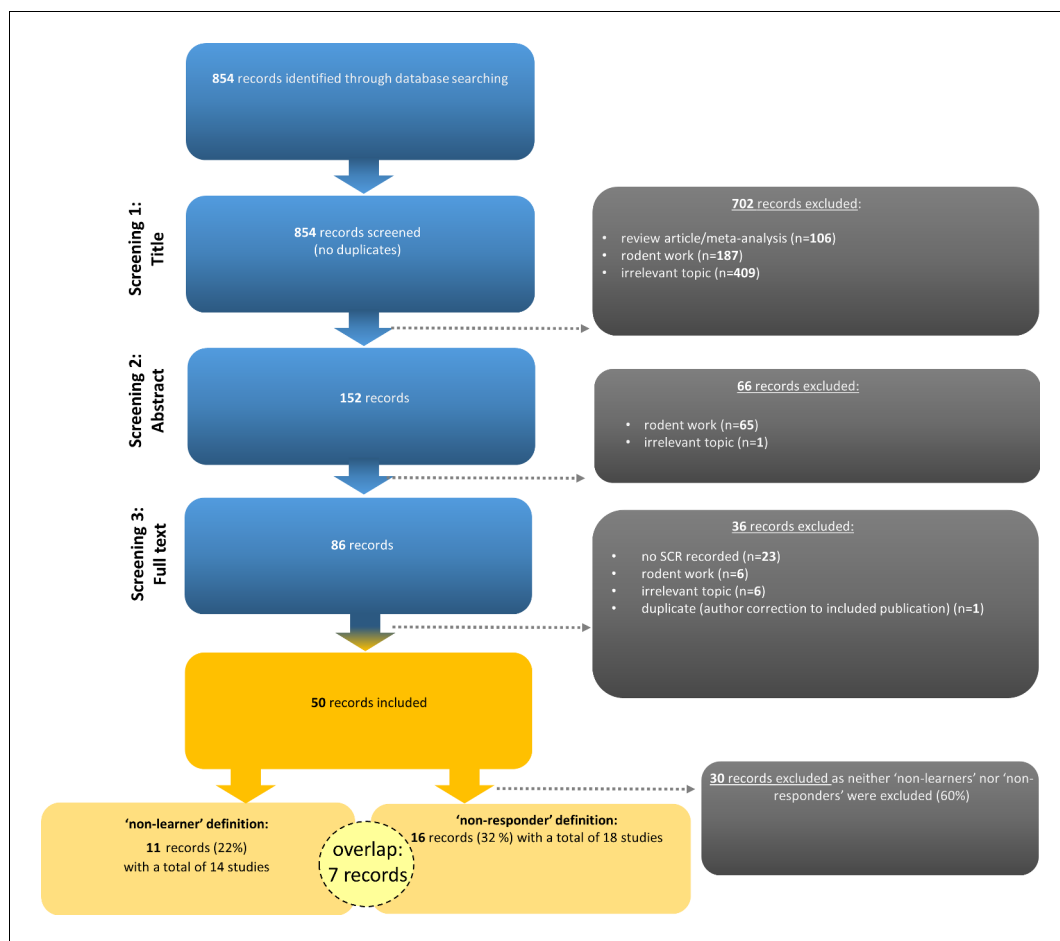


Figure 1. Flow chart illustrating the selection of records according to PRISMA guidelines (Moher et al., 2009). Note that seven records (14%) employed the definition and exclusion of both 'non-learners' and 'non-responders'. Examples of irrelevant topics included studies that did not use fear conditioning paradigms (see <https://osf.io/uxdhk/> for a documentation of excluded publications).

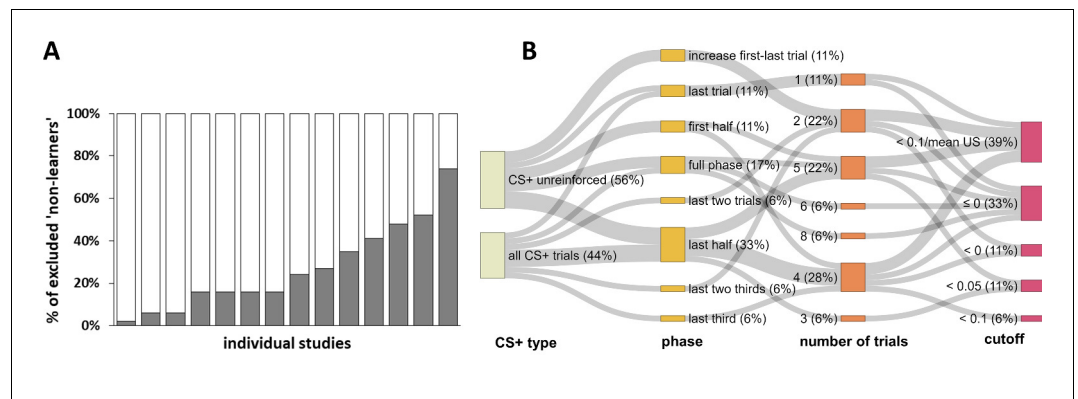


Figure 2. Graphical illustration of the percentage of 'non-learners' and forking path analysis across studies. (A) Illustration of the percentage of participants excluded ('non-learners') based on SCR CS+/CS− discrimination scores across studies included in the systematic literature search (note that these 14 individual studies are derived from 11 different records, as three records reported two individual studies each). Please note that some studies excluded participants on the basis of 'non-learning' as well as 'non-responding' (cf. **Figure 1**), and hence the percentages displayed here do not necessarily map onto the percentage of total participants excluded per study. Also note that the study with the highest percentage of excluded participants (i.e., 74%) reported the percentage of excluded participants as a single value that included 'non-learners' and 'non-responders'. This study is only included here because the largest proportion of exclusions can be expected to result from 'non-learning'. (B) Sankey plot showing the 'forking paths' of performance-based exclusion of participants as 'non-learners', illustrating differences in the experimental phase, number of trials, the SCR CS+/CS− discrimination score in μS used to define a 'non-learner', the CS+ type considered (illustrated as the nodes in graded colors) and their combinations used to define 'non-learners' across studies. Path width was scaled in relation to frequency of the combinations. Note that for some 'nodes' the percentages do not add up to 100% because of rounding.

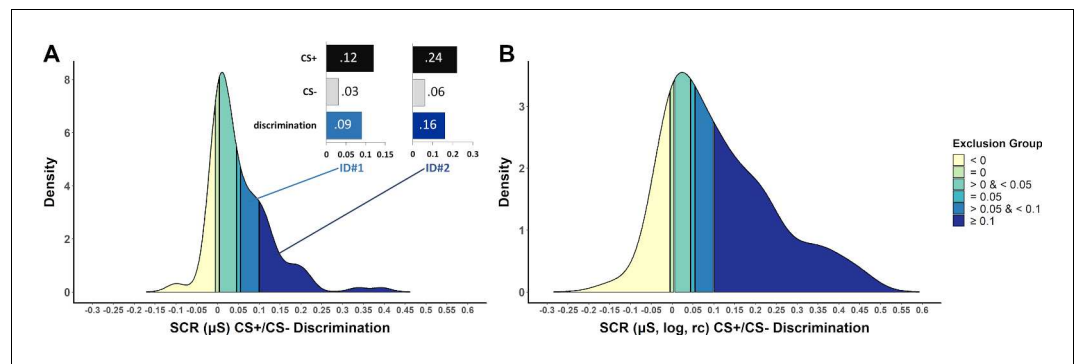


Figure 3. Density plots illustrating the frequency of CS+/CS– discrimination scores in a sample of $N = 116$ (Data set 1) based on the last half of the acquisition phase (including 7 CS+ and 7CS–, 100% reinforcement rate) for (A) SCR raw data and (B) logarithmized and range-corrected (rc; individual trial $SCR/SCR_{max_across_all_trials}$) SCR data (as it is typically not reported to which data exclusion criteria are applied). Color coding (yellow to blue) illustrates which part of the sample would be excluded when applying the performance-based exclusion criteria (i.e. CS+/CS– discrimination) as identified by the systematic literature search. Panel (A) also illustrates two case examples (ID#1 and ID#2) that differ in SCR amplitudes but importantly show the same discrimination ratio between CS+ and CS– (4:1). These two case examples illustrate that high CS+/CS– discrimination cutoffs favor individuals with high SCR amplitudes to remain in the final sub-sample. Data are based on a re-analysis of an unpublished data set recorded in the fMRI environment (Klingelhöfer-Jens M., Kuhn, M. and Lonsdorf, T.B.; unpublished).

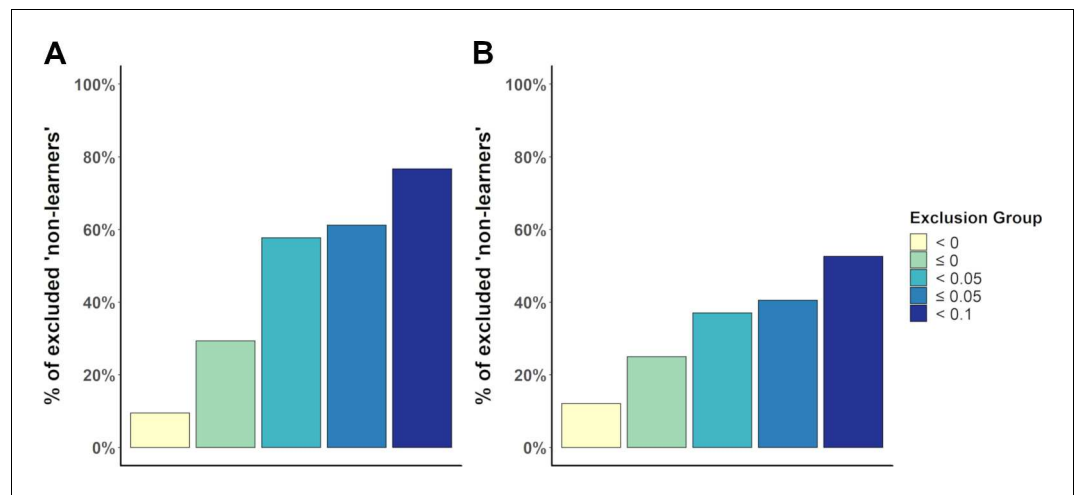


Figure 3—figure supplement 1. Percentages of participants excluded (Data set 1) when employing the different CS+/CS- discrimination cutoffs (as identified by the systematic literature search and graphically shown in **Figure 3B**) which are illustrated as density plots in **Figure 3**. Percentages are calculated on the basis of (A) raw SCR scores or (B) logarithmized and range-corrected scores in Data set 1. Note that the different groups are cumulative (i.e., the darker colored groups also comprise the lighter colored groups).

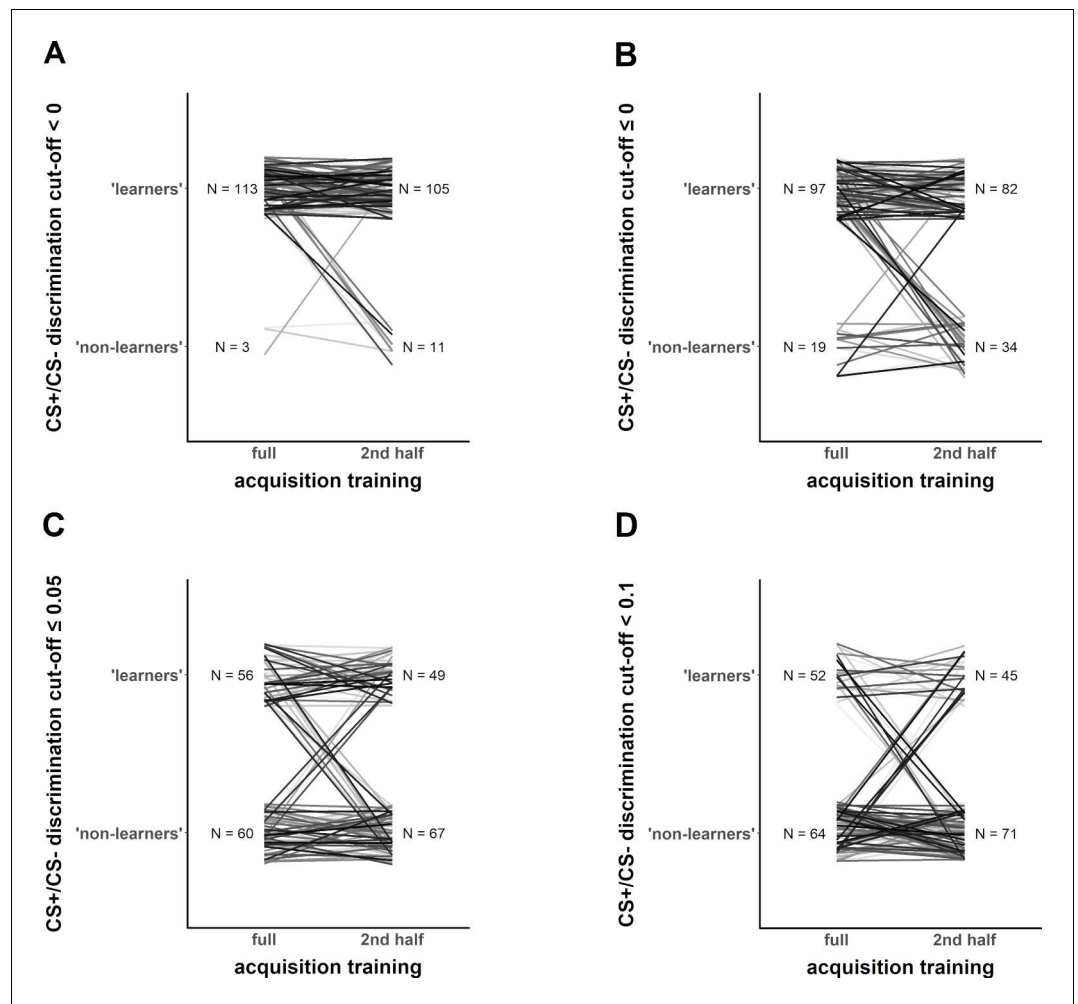


Figure 4. Exemplary illustration of individuals (Data set 1) that switch from being classified as 'learners' vs. 'non-learners' depending on the different CS+/CS- discrimination cutoff level (panels A–D), when calculation of CS+/CS- discrimination is based on either the full fear acquisition phase or the second half of the fear acquisition training (left and right part of each panel, respectively).

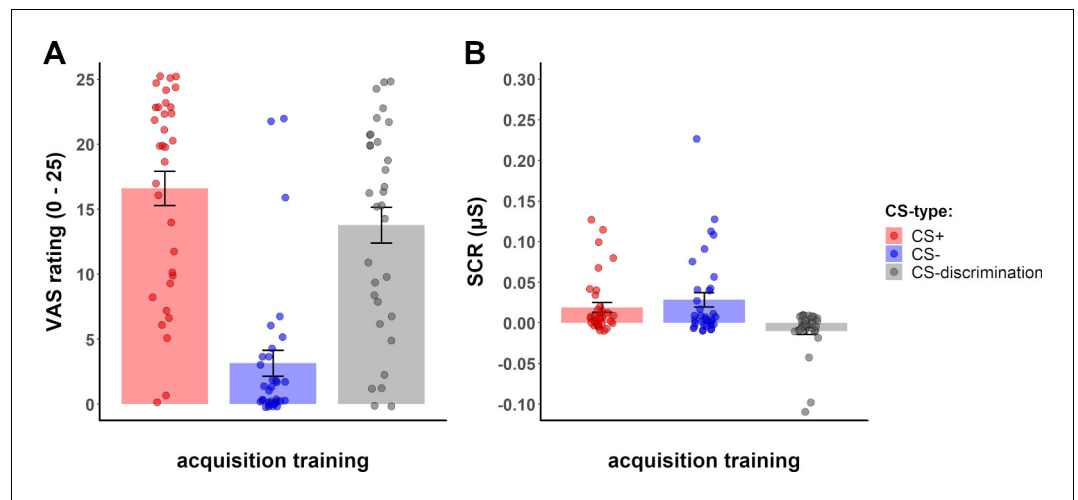


Figure 4—figure supplement 1. Bar plots (mean \pm SE) on which the superimposed individual data points show CS+ and CS- amplitudes (of raw SCR values) and CS+/CS- discrimination in (A) fear ratings and (B) SCRs raw values in the group of 'non-learners', as exemplarily defined for this example as a group consisting of individuals in the two lowest SCR CS+/CS- discrimination cutoff groups (i.e., ≤ 0) in Data set 1. This illustrates that individuals who fail to show CS+/CS- discrimination in SCRs (B) may in fact show substantial CS+/CS- discrimination (as an indicator for successful learning) in other outcome measures, as exemplarily illustrated here for fear ratings (A).

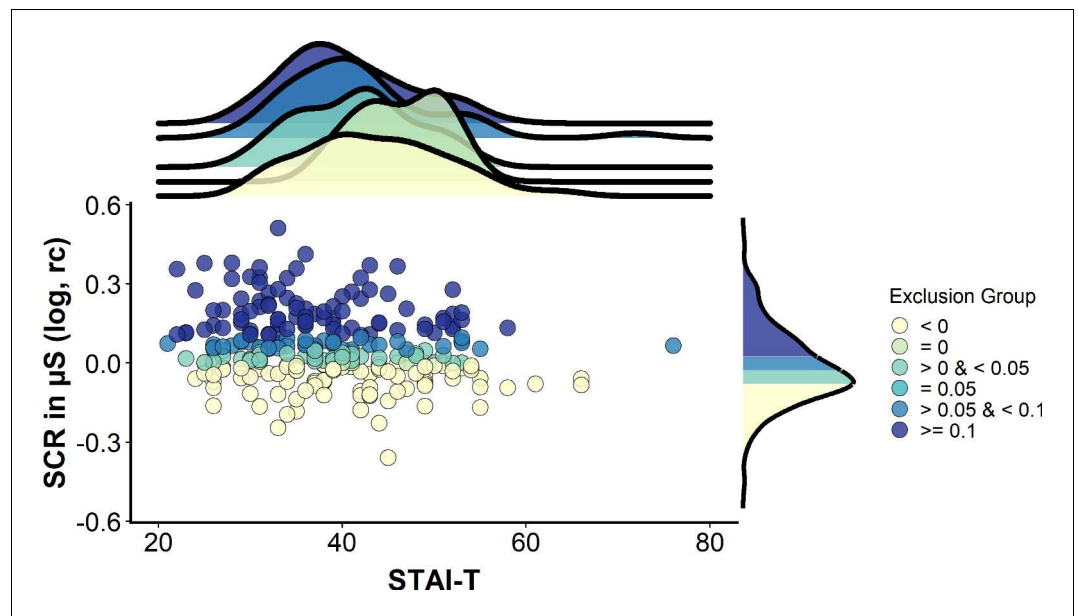


Figure 5. A case example illustrating potential sample bias induced by excluding individuals on the basis of CS+/CS– discrimination scores (based on logarithmized, range-corrected (rc) SCR data). Scatterplot illustrating the association between trait anxiety (measured via the trait version of the State-Trait Anxiety Inventory, STAI-T) and CS+/CS– discrimination scores in a sample of $N = 268$ (Data set 2). Color coding (yellow to blue) illustrates which part of the sample would be excluded when applying the performance-based exclusion criteria (i.e. CS+/CS– discrimination) as identified by the systematic literature search. Note that within this sample, no individuals were identified with CS+/CS– discrimination equaling $0.05 \mu\text{S}$. The upper panel illustrates densities for trait anxiety for the different CS+/CS– discrimination groups. The rightmost panel illustrates the density for CS+/CS– discrimination in the full sample. Data are based on a re-analysis of a data set recorded in the behavioral environment (Schiller et al., 2010). Note that despite the different color coding, which serves illustrative purposes only, the groups are in practice cumulative. More precisely, the groups illustrated by lighter colors are always contained in the darker colored groups when applying the respective cutoffs. For example, the group excluded when employing a cutoff of $<0.1 \mu\text{S}$ (mid blue) also comprises the groups already excluded for the lower cutoffs of $= 0.05 \mu\text{S}$ (light blue), $<0.05 \mu\text{S}$ (turquoise), $= 0 \mu\text{S}$ (light green) and $<0 \mu\text{S}$ (yellow). For illustrative purposes, the different groups are treated as separate groups in this figure.

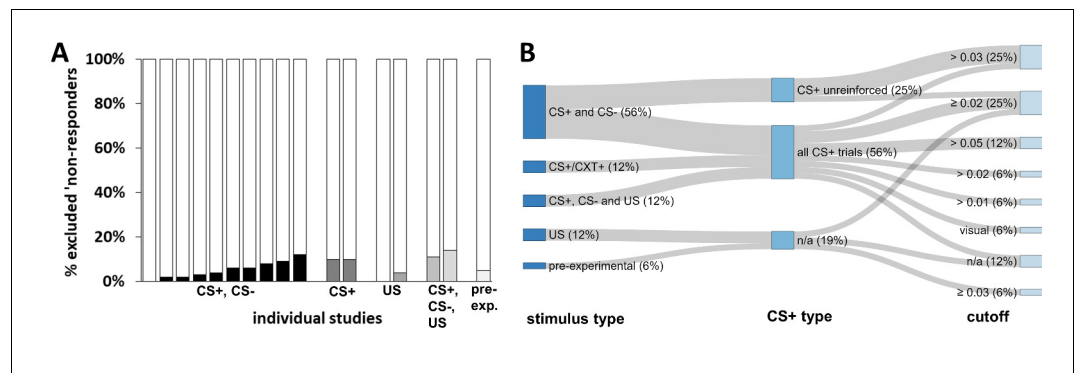


Figure 6. Graphical illustration of the percentage of 'non-responders' and forking path analysis across studies. (A) Illustration of the percentage of participants excluded from each study as a result of 'SCR non-responding' to (i) the conditioned stimuli (i.e., CS+ and CS-), (ii) the US, (iii) the CS+ (which also comprises a study that used the CXT+, i.e. context), (iv) the CS+, CS- and US or (v) a pre-experimental test. Note that these 18 individual studies are derived from 16 different records, two of which included two different studies that used the same criteria. Note that some studies excluded participants on the basis of 'non-learning' as well as 'non-responding', and hence the percentages displayed here do not necessarily map onto the percentage of total participants excluded from each study. Also note that a single study (*Schiller et al., 2018*) is not included in this visualization because it reported % 'non-learners' and % 'non-responders' as a single value. This value has been included in the visualization of 'non-learners' (*Figure 2*) as these are expected to represent the largest proportion. (B) Sanky plot illustrating the stimulus type (pre-experiment refers to determination of 'responding' in an unrelated phase prior to the experiment), the minimally required response amplitude in μ S (note that 'visual' refers to visual inspection of the data without a clear-cut amplitude cutoff, NA refers to no criterion applied) illustrated as the nodes in graded colors and their combinations that lead to classification as a 'non-responder'. Path width was scaled in relation to frequency of the combinations. Note that for some 'nodes' the percentages do not add up to 100% because of rounding.

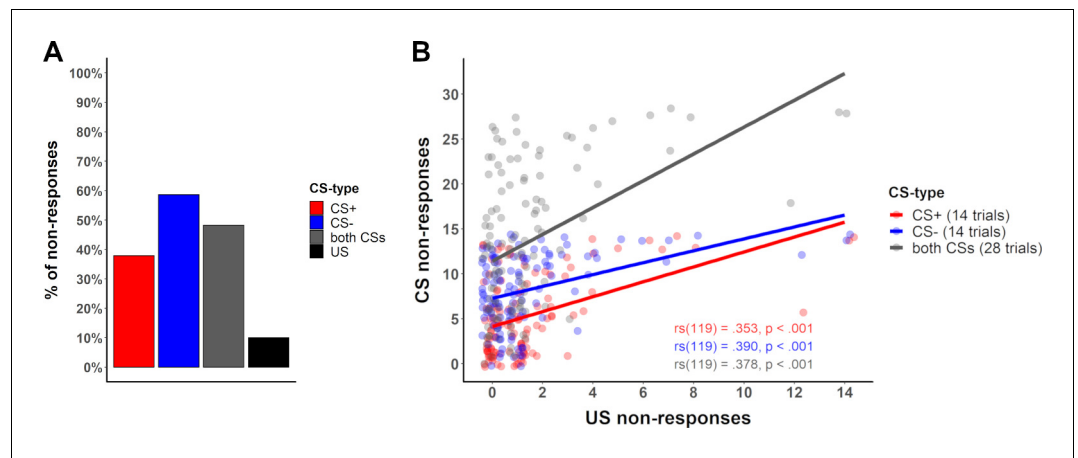


Figure 7. Percentage of no-responses across stimuli and correlation between CS and US non-responses. **(A)** Bar plot displaying the number of 'non-responses' to the CS+, CS-, across both CS and to the US across all participants in Data set 1 (see **Appendix 4—table 1** for percentages across different data sets). **(B)** Scatterplot illustrating the number of 'non-responses' (i.e., zero-responses, here defined by an amplitude $<0.01 \mu\text{S}$) to the US presentations (total of 14 presentations) and the CS+ (red) and CS- (blue) responses (14 presentations each) for each participant in Data set 1. For completeness sake, 'non-responses' across CS types are illustrated in gray (CS+ and CS- combined, total of 28 presentations). Lines illustrate the Spearman correlation (r_s) between 'non-responses' to the US and 'non-responses' to the CS+, CS- and both CS, with corresponding correlation coefficients (font color corresponds to CS type) included in the figure.

10 Study IV

This article was published in *Behaviour Research and Therapy*, 153, Lonsdorf, T. B., Gerlicher, A., Klingelhöfer-Jens, M., & Krypotos, A.-M., Multiverse analyses in fear conditioning research, 104072, no changes have been implemented, Copyright Elsevier (2022).



Multiverse analyses in fear conditioning research

Tina B. Lonsdorf^{a,*}, Anna Gerlicher^b, Maren Klingelhöfer-Jens^a, Angelos-Miltiadis Kryptos^{c,d}

^a Department of Systems Neuroscience, University Medical Center Hamburg Eppendorf, Germany

^b Department of Clinical Psychology, University of Amsterdam, the Netherlands

^c Department of Experimental Psychology, Utrecht University, the Netherlands

^d KU Leuven, Belgium

ARTICLE INFO

Keywords:

Anxiety disorders
Questionable research practices
Good research practices
p-hacking
Transparency

ABSTRACT

There is heterogeneity in and a lack of consensus on the preferred statistical analyses in light of a multitude of potentially equally justifiable approaches. Here, we introduce multiverse analysis for the field of experimental psychopathology research. We present a model multiverse approach tailored to fear conditioning research and, as a secondary aim, introduce the R package ‘multifear’ that allows to run all the models through a single line of code. Model specifications and data reduction approaches were identified through a systematic literature search. The heterogeneity of statistical models identified included Bayesian ANOVA and t-tests as well as frequentist ANOVA, t-test as well as mixed models with a variety of data reduction approaches. We illustrate the power of a multiverse analysis for fear conditioning data based on two pre-existing data sets with partial (data set 1) and 100% reinforcement rate (data set 2) by using CS discrimination in skin conductance responses (SCRs) during fear acquisition and extinction training as case examples. Both the effect size and the direction of effect was impacted by choice of the model and data reduction techniques. We anticipate that an increase in multiverse-type of studies will aid the development of formal theories through the accumulation of empirical evidence and ultimately aid clinical translation.

1. Introduction

Scientific work - also in experimental psychopathology - consists of multiple steps including data recording, measurement, processing, analysis, illustration, and interpretation. Yet, every single step during the scientific process inherently involves a plethora of decisions in light of a large pool of potentially equally justifiable options with respect to data recording, response quantification, data processing and statistical analysis. This has been referred to as “researchers degrees of freedom” (Simmons, Nelson, & Simonsohn, 2011) or the “garden of forking paths” (Gelman & Loken, 2014), the navigation of which can be challenging in absence of empirical evidence and/or precise (formal) theories providing a justification for one specific choice. As a result, many decisions are more or less arbitrary and potentially equally justifiable, even though it remains unclear if all different paths converge in the identical statistical result and interpretation or to what extent they diverge. This often results in extensive discussions both during data analyses as well as during peer-review and generally hampers the translation of basic research findings to the clinics.

The consequences and implications resulting from this plethora of

alternative choices at each step of the scientific process as well as potential remedies have been discussed intensively in psychology recently (Botvinik-Nezer et al., 2020; Sandre et al., 2020; Silberzahn et al., 2018; Simmons et al., 2011). These meta-scientific topics have been highlighted in the past years also for fear conditioning research in humans with a focus on *procedural* heterogeneity and construct operationalization: More precisely, the role of procedural heterogeneity has been discussed for the reinstatement-induced return of fear (Haaker, Golkar, Hermans, & Lonsdorf, 2014; Sjouwerman & Lonsdorf, 2020), the impact of inconsistent definitions of key learning indices such as “extinction retention” (Lonsdorf, Merz, & Fullana, 2019) as well as the definition of “learning” vs. “non-learning” and “responding” vs. “non-responding” (Lonsdorf et al., 2019) as well as skin conductance response quantification (Kuhn, Gerlicher, & Lonsdorf, 2022; Sjouwerman, Illius, Kuhn, & Lonsdorf, 2021).

A multiverse of statistical models. What has not yet been systematically addressed in the field of fear conditioning research are the many decisions required when planning statistical analyses of a fear conditioning study (Ney et al., 2020) which involves questions, such as: Shall I run a t-test or an Analysis of Variance (ANOVA)? Shall I use

* Corresponding author. 20246, Hamburg, Germany.

E-mail address: t.lonsdorf@uke.de (T.B. Lonsdorf).

<https://doi.org/10.1016/j.brat.2022.104072>

Received 28 September 2021; Received in revised form 4 February 2022; Accepted 7 March 2022

Available online 21 March 2022

0005-7967/© 2022 Elsevier Ltd. All rights reserved.

p-values or Bayes factors? Do I need to include covariates in my analyses? Shall I use aggregated scores across an experimental phase or should I consider each trial separately? Different decisions for each of these data analytical questions and their combinations lead to different, yet often equally justifiable data analytical pipelines which hampers comparability across studies and also leaves room for potential Questionable Research Practices (QRP) engaged in unintentionally or intentionally (Simmons et al., 2011). In absence of precisely formalized theories and hypotheses, different researchers are likely to pick different – often equally justifiable – analytical pipelines to answer the same research question. This has been impressively illustrated across a number of studies in different research fields in the past years that showed the different paths can lead to substantially different conclusions (e.g., Boehm et al., 2018; Botvinik-Nezer et al., 2020; Dutilh et al., 2019; Kuhn et al., 2022; Lonsdorf et al., 2019a, 2019b; Silberzahn et al., 2018).

In psychology, verbal theories dominate. With the term “verbal theories” we refer to the description of different latent constructs and their relationships in natural language only (Farrell & Lewandowsky, 2018; Lewandowsky & Farrell, 2010). This type of descriptions inherently gives room for statistical flexibility: For example, a theory may predict that after reliable pairing of a neutral stimulus (Conditioned Stimulus or CS+) with an unpleasant event (Unconditioned Stimulus, US) while a second neutral stimulus (CS-) is not paired with the US, the CS+ but not the CS- will elicit an anticipatory fear reaction (i.e., conditioned response, CR). This anticipatory fear reaction will manifest as larger skin conductance responses (SCRs) to the CS+ as compared to the CS-, referred to as CS discrimination. Yet, this verbal theory is ill-defined as it does not specify for instance a) how high those responses will be (e.g., 10, 20, 50 point differences), and b) how many pairings between the CS+ and the US are required for differential responses will be expressed (e.g., after 2, 3, or 10 trials). This imprecision in theory results in a multitude of different statistical models that may be used (Muthukrishna & Henrich, 2019), idiosyncratic criteria about how large CS discrimination needs to be (Lonsdorf et al., 2019), or to consider different amounts of trials in analyses (Lonsdorf et al., 2019, 2019; Ney et al., 2020). The decisions to choose a specific statistical analysis from a set of plausible analyses can be considered to occur mostly in good faith. Yet, even for models intended to test the same predictions it remains unclear if the statistical results derived from different statistical approaches or processing pipelines and the interpretation based on them are comparable and converge across data analytical pipelines. Recently, Ney et al. (2020) described inconsistent statistical strategies when analyzing skin conductance data in fear extinction training. Their results suggest unsatisfying correlations between the different analysis approaches as applied to the same data-set which were mainly attributable to the selection of trials from different stages of the experimental phases and employment of trial-by-trial analyses vs. averaged scores (Ney et al., 2020). This may not be particularly surprising as different analytic strategies may not test exactly the same underlying hypothesis but may - intentionally or unintentionally - test different hypotheses. This is true for models with and without covariates (Del Giudice & Gangestad, 2021) but also for models using different numbers of trials. Including only the first 2 trials of a (delayed) extinction phase tests for fear recall, while including only the last two trials tests for end-point extinction learning successes and trial-by-trial analyses test for temporal dynamics during extinction learning. In sum, model specification is a major issue and often characterised by uncertainty about which variables to include, how to operationalize them and their interrelations with associated variables. Hence, it is desirable to formalize the to date predominantly verbal theories. This, however, requires a deep understanding on the impact of individual specifications which must be considered a stopover on the path towards more formalized models.

How to navigate the multiverse of statistical analyses. A promising approach to systematically and comprehensively explore the impact of such methodological heterogeneity in the data processing or statistical analyses, is a multiverse-type analysis (Steege, Tuerlinckx,

Gelman, & Vanpaemel, 2016) or the related specification curve analyses (Simonsohn, Simmons, & Nelson, 2020). Multiverse-type analyses consider the i) multiverse of justifiable data sets that can be generated from a single set of raw data through reasonable data processing decisions (i.e., “data multiverse”) or considers ii) the multiverse of different reasonable statistical models applied to a single data set to answer a single research question (i.e., “model multiverse”) or iii) their combinations. The multiverse approach thus systematically generates a set of universes for alternative data processing and/or statistical pipelines. This holds promise to achieve a better estimate of a given effect as well as its robustness as compared to the standard approach of analyzing and reporting results based on a single processing and analytical pipeline which is often selected based on more or less arbitrary decisions. Thus, in case the results of a multiverse analysis show convergence across the different processing and/or analytical choices (i.e., forking paths), the robustness of an effect independent of the used preprocessing pipeline (for a data multiverse) or statistical pipeline (for a model multiverse), can be assumed. However, if divergence is observed, this may inform us on potential boundary conditions (for instance inclusion of specific covariates or trial numbers) that may systematically impact the strength of the effect under study.

The main aims of the current work are: a) to introduce the readers to the idea of multiverse-type of studies by focusing on fear conditioning research and b) to showcase an illustrative application example on the (degree of) impact of different data analysis choices when applying different statistical models to the same data set (i.e., model multiverse analysis; Steege et al., 2016) based on two pre-existing datasets. Of note, the choice of the different statistical models included was guided by a systematic literature search covering a representative 6-month period which hence reflects which statistical analyses are typically performed in the field. A secondary aim of this work is to introduce, via a short tutorial, a new open software package, named ‘multifear,’ that allows researchers in the field of fear conditioning to employ this computationally demanding approach of running all the models (as identified in the literature) with ease and through a single line of code to their own data - for both Null Hypothesis Significance Testing (NHST) as well as Bayesian statistics using Bayes factors.

2. Methods

Systematic literature search: A systematic literature search was performed as suggested by the PRISMA guidelines (Moher et al., 2009). The search covered all publications (including e-pubs ahead of print) in PubMed in a six months period (22.9.2018 to 22.3.2019) and served the purpose to extract procedural and statistical specifications employed in the field of fear conditioning relevant for a number of planned research projects (e.g., Lonsdorf et al., 2019). As described previously (Lonsdorf et al., 2019), the following search terms were employed: threat conditioning OR fear conditioning OR threat acquisition OR fear acquisition OR threat learning OR fear learning OR threat memory OR fear memory OR return of fear OR threat extinction OR fear extinction. In case, the search included author corrections published within the search period, the original study was included unless already included. A total of 854 records as listed in PubMed were identified, stage 2 screening of the abstract yielded 152 records. Eighty-six publications were retained at stage 3 screening of the full text. The final set of publications consisted of 50 records which all reported Results for (1) SCRs as an outcome measure from (2) the fear acquisition training phase (3) in human participants. From those records we selected all the analyses that tested the hypothesis of differences between the CS+ and the CS-. This included also the statistical models that in addition to CS differences included also the factor time or a between group factor. As such, we excluded any analysis that included the use of a computational model. Also, in case covariates were included in a statistical model, the model was categorized without these covariates to increase the generalizability of the findings. A flow chart and more details are provided in our previous

publication (Lonsdorf et al., 2019).

2.1. Data-set 1

Participants Data from a previous publication of $N = 40$ male participants (Age: mean = 28.1 years; SD = 2.7 years) were re-analyzed (Gerlicher, Tüscher, & Kalisch, 2018). Written informed consent was provided by all participants and the protocol was approved by the local ethics committee (Ethikkommission der Landesärztekammer, Rheinland-Pfalz).

Stimuli In brief, two black geometric symbols (square, rhombus), presented for 4.5s, served as CS+ and CS- superimposed on two different background context pictures (A, B; kitchen or a living room). Assignment of the symbols to CS+ or CS- and the rooms to contexts A or B was randomized between participants. The US consisted of an electrical stimulus (three square-wave pulses of 2 ms, 50 ms interstimulus interval) generated by a DS7A electrical stimulator (Digitimer, Weybridge) and applied to the right dorsal hand via a surface electrode with platinum pin (Specialty Developments, Bexley, UK). US delivery terminated with CS+ presentation. Inter-trial intervals lasted 17, 18, or 19 s (mean of 18.5 s). Trial order was randomized with the restriction that not more than two trials of the same type (i.e., CS+, CS-) followed each other.

Procedure Data were recorded in a three-day fMRI paradigm comprising fear acquisition on day 1, extinction and subsequent drug administration on day 2, and a test of the effect of the drug manipulation on day 3. For the purpose of the present work, only SCR data recorded prior to drug intake during fear acquisition and extinction are re-analyzed. US intensity was calibrated to a level described as painful, but still tolerable by the participant prior to the experiment. During fear acquisition training on day 1, ten CS+ and ten CS- trials were presented in context A. Five out of ten CS+ presentations (i.e., 50%) were reinforced with an electric stimulus. During extinction training on day 2, fifteen CS+ and CS- trials were presented in context B. Stimulus presentation was controlled by Presentation software (Version 14.8, Neurobehavioral Systems, Inc, Albany California, USA).

Skin conductance recording Electrodermal activity was recorded from the thenar and hypothenar of the non-dominant hand using self-adhesive Ag/AgCl electrodes (EL-509, BIOPAC Systems Inc., Goleta, CA, USA) filled with an isotonic electrolyte medium. The signal was recorded using the Biopac MP150 with EDA100C. We low-pass filtered the raw signal offline with a second-order Butterworth filter with a cut-off frequency of 1 Hz in Matlab (Mathworks, Natick, Massachusetts, USA).

2.2. Data-set 2

Participants Participants from the baseline-time point (T0) of a longitudinal fear conditioning study in 120 participants were included whereof data from four participants were excluded due to protocol deviations leaving 116 participants for analyses (77 females; age: mean = 24.38 years; SD = 0.34 years). These data have been included as a case example in a previous publication focusing on the methodological question of defining 'no-responder' and 'non-learner' (Lonsdorf et al., 2019), the impact of different SCR quantification approaches (Kuhn et al., 2022, Sjouwerman et al., 2021), have been analyzed with respect to temporal stability (i.e., test-retest for a six month period, Klingelhöfer-Jens, Ehlers, Kuhn, Keyaniyan, & Lonsdorf, 2022), and associations between conditioned responding and brain morphology (Ehlers, Nold, & Kuhn, 2020). All participants gave written informed consent to the protocol which was approved by the local ethics committee (PV 5157, Ethics Committee of the General Medical Council Hamburg).

Stimuli The US was an electrostatic stimulus consisting of three 2 ms electrostatic rectangular pulses with an interpulse interval of 50 ms (onset: 200 ms before CS+ offset) and was administered to the back of the right hand of the participants. It was generated by a Digitimer DS7A constant current stimulator (Welwyn Garden City, Hertfordshire, UK)

and delivered through a 1 cm diameter platinum pin surface electrode (Specialty Developments, Bexley, UK). The electrode was attached between the metacarpal bones of the index and middle finger. US intensity was individually calibrated in a standardized step-wise procedure aiming at an unpleasant, but still tolerable level.

Two light grey fractals served as conditioned stimuli which were presented 14 times in a pseudo-randomized order for 6–8 s (mean: 7 s). Allocation to CS+ and CS- was counterbalanced between participants and the CS+ was followed by the US in all cases during fear acquisition training. A white fixation cross was shown for 10–16 s (mean: 13 s) which served as the inter-trial intervals (ITIs). All stimuli were presented on a dark grey background and controlled by Presentation software (Version 14.8, Neurobehavioral Systems, Inc, Albany California, USA).

Procedure The paradigm (for details see Lonsdorf, Klingelhöfer-Jens et al., 2019) consisted of a two-day un instructed fear conditioning paradigm with habituation and acquisition training (100% reinforcement rate) taking place on day 1 and extinction training and reinstatement test taking place on day 2. The study included a baseline measurement (T0) and a follow-up measurement (T1) six month later when the identical paradigm was conducted again. Only data from T0 are included here. During all experimental phases, BOLD fMRI, fear ratings (prior to and after each experimental phase) and skin conductance responses were acquired. BOLD fMRI as well as fear ratings are, however, not included in the present work, as it focuses on different statistical models using skin conductance as a case exemplary outcome measure.

Skin conductance recording Skin conductance response was measured via self-adhesive Ag/AgCl electrodes placed on the palmar side of the left hand on the distal and proximal hypothenar. Data were recorded with a skin conductance unit together with a Biopac MP100-amplifier system (BIOPAC® Systems Inc., Goleta, CA, USA) and converted from analog to digital using a CED2502-SA with Spike 2 software (Cambridge Electronic Design, Cambridge, UK).

Skin conductance response quantification and processing (data set 1 and 2) SCRs were scored computer-assisted by using a custom-made computer program (EDA View, developed by Prof. Dr. Matthias Gamer, University of Würzburg) according to published guidelines (Boucsein et al., 2012) and while being blind to stimulus type associated with a given SCR. More precisely, the trough was identified in a post stimulus onset latency window (OLW) of 0.9–4s for data-set 1 (Boucsein et al., 2012) and 0.9–3.5s for data set 2 (Sjouwerman & Lonsdorf, 2019). The peak was identified in a peak detection window (PDW) of maximally 5s post SCR onset. In case of multiple peaks in the PDW, the first peak was considered.

Data were down-sampled to 10 Hz. Each scored SCR was checked visually, and the scoring suggested by the algorithm was corrected if necessary (e.g., the foot or trough was misclassified by the algorithm). Data with recording artifacts or excessive baseline activity (i.e., more than half of the response amplitudes) were treated as missing data points and excluded from the analyses. For data set 2, SCRs below 0.01 μS or the absence of any SCR (i.e., flat line or habituation drift) within the defined time window were classified as non-responses and set to 0. The threshold of 0.01 μS for this data set was determined empirically by visually inspecting responses specifically above and below this cutoff (Lonsdorf et al., 2019), which suggested that in this data set, responses $>0.01 \mu\text{S}$ can be reliably identified. For data set 1, a minimum amplitude criterion of 0.02 μS was used.

In contrast to the original analysis for data set 1 (Gerlicher et al., 2018) where data was excluded when more than 75% of CS-evoked SCR were scored as zero, we here only excluded trials when it was affected by recording artifacts. This led to the exclusion of data of four participants during fear conditioning and two participants during extinction, leaving data of $N = 38$ participants for statistical analysis, respectively. Raw data were log transformed using the formula $\log(1 + \text{raw value})$.

2.3. Statistical analyses

Multiverse analyses can be run in any statistics software. Given the volume of analyses, though, a scripting language seems less time consuming and error prone than click-based statistical softwares. Here, we used the R software language (R Core Team, 2013). To enable researchers in fear conditioning research to easily adopt a multiverse approach, we present the freely available R package named ‘multifear’ available at <https://github.com/AngelosPsy/multifear>. The R package is able to run all the analyses described in the manuscript in a single line of code, with the researcher having to only load their data in R, name the columns names for each CS, and the column name for the groups (if different groups were tested). The package is also able to generate plots as well as a summary of results (see main results for examples). For NHST analyses, we computed the mean and median of p -values across all tests, proportion of p values below the chosen alpha level (using an alpha level of 0.05 as it is common in the literature), as well as plotted a histogram of all p -values. We did the same separately for Bayes factors, with Bayes factors above 1 indicating that there is relatively more evidence that the data came from the alternative compared to the null hypothesis, and the reversed for values below 1. We also plotted a histogram for Bayes factors. Lastly, we have created different forest plots separately for the acquisition and extinction phase, plotting the Cohen’s d effect size for each test.¹ Note that the computed effect sizes are based on the collected data and they cannot answer the question as to whether the observed effects are substantial or not. This is something that is purely based on a study’s research questions, as, for example, when evaluating the effectiveness for a drug a larger effect may be thought to be substantial compared to when comparing two conditions in a fear conditioning study. A detailed vignette about how to install and use the R package is available at <https://angelopsy.github.io/multifear/>. We have also created a vignette, available within the package as well as <https://htmlpreview.github.io/?https://github.com/AngelosPsy/multifear/blob/master/doc/internals.html>, that describes in plain words the internals of the package. As the major aim of the present work is to showcase the idea and value of multiverse-type of analyses for the field of experimental psychopathology, we refrain from providing specific details on the steps from entering data to getting results in the ‘multifear’ package and refer to the online vignette for these details.

3. Results

Results of the systematic literature search. Table 1, shows the frequencies with which each statistical model was used in the publications included in the systematic literature review. The most common statistical analysis employed in the field is a repeated measures ANOVA

Table 1

Number of studies that used any one of the statistical models (i.e., repeated measures analysis of variance with different factors, t -test, mixed models). Note that the sum of studies is higher than 50 (i.e., the number of records of our review), because some publications reported multiple experiments or analyses.

	Acquisition	Extinction
Repeated Measures ANOVA of CS (+group)	11	6
Repeated Measures ANOVA of CS x Trial/(Block) (+group)	29	21
Paired t -test	5	1
Mixed Models (including Multilevel Models)	4	2

¹ Please note that for some effects the whiskers were too small to plot and they are hidden by the size of the box. Also, it is not uncommon for η^2 to have asymmetric confidence intervals, given that by definition the effect cannot be lower than zero.

with a test of CS \times Trial interaction or without the Trial factor being included. In case between group differences were tested, an extra between group factor was included. Mixed models and paired t -tests were also used in the literature, although sparingly.

Importantly, the different statistical models described above include data processed through different data reduction procedures as identified from the systematic literature search. Specifically the identified statistical models for the repeated measures ANOVA and the mixed models included (a) single trial SCRs to CS+ and CS-, or (b) SCRs evoked by the first and last CS+ and CS- trials (i.e. first vs. last trial), or (c), the SCR averaged across the first minus the last two CS+ and CS- trials (i.e., first 2 vs. last 2 trials), or (d) SCRs averaged across two succeeding CS+ and CS- trials (i.e., averages per 2 trials), respectively. Similarly, SCRs were averaged across succeeding blocks of (e) 10%, (f) 20%, (g) 33%, or (h) 50% (i.e., half of the trials) of CS+ and CS- trials, respectively, and the SCR averages of all 10%, 20%, 33% and 50% trial-blocks per CS type were subjected to the analysis. Lastly (i), SCRs were averaged across all trials except for the first CS+ and CS- trial (as no learning could possibly have taken place), respectively, and the CS+ and CS- averages were entered into the analysis.² For the repeated measures ANOVAs the CS and trial were included as repeated measures factors. For the present analyses we did not include group as a factor in any of our analyses. For the t -tests analyses we used the same data reduction procedures as described above (a - i) but we averaged across the CS+ and CS- trials. This means, for example, that in case we had averaged across succeeding blocks of 20% of the trials, those blocks were then averaged again separately for CS+ and CS-.

We now turn to showcasing a model multiverse analysis based on the specifications derived from the systematic literature search by using two pre-existing data sets as case examples. Based on this principled approach we offer and showcase a tool (the ‘multifear’ R package) that allows to run this model multiverse covering the typically used statistical models with as little as a single line of code.

Multiverse Results. The top panel of Fig. 1 depicts log-transformed SCRs (+se), averaged across participants per trial, for the acquisition training and (delayed) extinction training phases for data set 1 (50% reinforcement rate), and the bottom panel shows the same for data set 2 (100% reinforcement rate). In both data sets, we observe the expected pattern indicating successful fear acquisition and extinction training: participants exhibit stronger SCRs to the CS + than to the CS- in the acquisition training phase. In the delayed extinction training phase, we see a pattern of incomplete extinction for data set 1, with responses to the CS + remaining higher than the responses to the CS- even after 15 trials (data set 1). For data set 2, we observe a different pattern with comparable response SCR amplitudes to both CS types throughout delayed extinction which is already evident from the very first trial of the extinction training phase. Note, SCRs were relatively larger in data set 1 than data set 2. While the reason for this is unclear, a potential explanation might be the usage of a more aversive US in data set 1: US intensity was calibrated to a level perceived as ‘maximally painful, but still tolerable’ in data set 1 compared to ‘maximally uncomfortable, but not painful’ in data set 2. Empirical and theoretical work suggests that stronger US intensity is associated with larger conditioned responses (e.g., Morris & Bouton, 2006; Rescorla & Wagner, 1972). An alternative explanation might be by the different reinforcement rates employed in data set 1 (partial) and 2 (100%). That is, SCRs have been suggested to reflect the associability of a stimulus (e.g., Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; Seymour et al., 2005; Tzovara, Korn, & Bach, 2018; Zhang, Mano, Ganesh, Robbins, & Seymour, 2016). In a paradigm

² Based on the number of trials included in each phase, there could be overlap between the different data reduction methods. To illustrate, in case 10 trials are used and a repeated measures ANOVA is used with cs as the main effect, then methods (d) and (f) will return identical Results (see results of acquisition phase for data set 1).

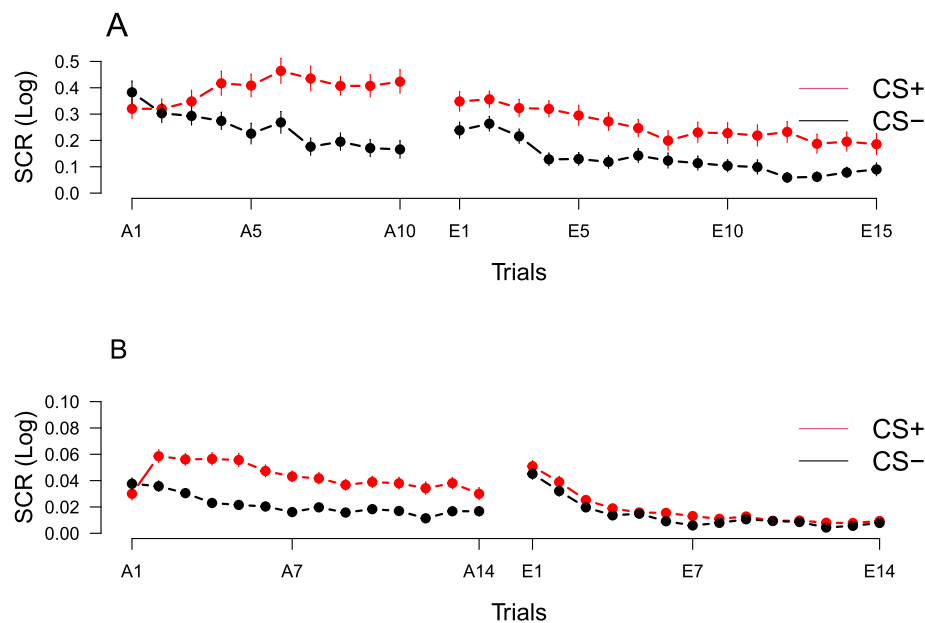


Fig. 1. Depiction of log transformed SCRs per CS (i.e., CS+, CS-) and trial for the Acquisition (i.e., A) and Extinction (i.e., E) training phase for study 1 (A) and study 2 (B).

with 100% reinforcement rate (data set 2) the associability of the CS rapidly decreases over the course of acquisition, whereas the associability of the CS, and with it SCRs in general, may stay comparably higher in paradigms with 50% reinforcement rate (data set 1).

We then performed the full multiverse (i.e., all different combinations of models and procedures) separately for the acquisition and the extinction training phases. The *multifeat* package allows such extensive analyses in a single line of code (see below for an illustrative example). [footnote] Please note that although here we present for illustrative purposes an example with a single group, the *multifeat* package can also accommodate group analyses with just specifying the name of the column that includes the group data. We point to our github page for more examples. The function runs all the relevant models as derived from the literature using both Null Hypothesis Significance Testing (NHST) as well as Bayesian statistics using Bayes factors. The output is a data frame, with each line including the Results of the different models (e.g., *t*-test, ANOVA), the different data reduction procedures employed (e.g., means per whole block), as well as the relevant inferential statistics (e.g., *p*-values, Bayes factors). In the code line below we see that the *multifeat* package is able to generate a data frame with all test by simply defining the data set (here named 'my_data'), the column names for the CS+ (here 'csp'), the column names for CS- (here 'csm'), and the name of the column including the participant number (here 'id').

```
multifeat::multiverse_cs(cs1 = csp, cs2 = csm, data = my_data, subj = "id")
```

Fig. 2 includes a histogram of *p*-values and Bayes factors for the acquisition (data set 1: panel A, data set 2: panel C) and the extinction data (data set 1: panel B, data set 2: panel D), for each model and data reduction procedure used in the multiverse. Our analyses returned 116 lines for the results from the acquisition training data and 116 lines for the results from the extinction training data. Regarding the acquisition training data of data set 1, the mean *p*-value was smaller than 0.001, with the 100% of the values falling below the alpha level of 0.05. For the Bayes factors, the mean Bayes factor was above 1000 and the proportion of Bayes factors above 1 was equal to 100%. Note that we abstain from evaluating whether Bayes factors provide evidence that is weak or strong or even anecdotal. We refer researchers to commonly used categories of the interpretation of Bayes factors (Wethzels, 2011). For data set 2, the mean *p*-value was equal to 0.06, with the 73.53% of the values falling below the alpha level of 0.05. For the Bayes factors, the mean

Bayes factor was above 1000 and the proportion of Bayes factors above 1 was equal to 70.59%. Fig. 2 shows which models results in non-significant results and detailed information can be returned from the data frame returned with the results.

For the extinction training data of the first data set, the mean *p*-value was equal to 0.41, with the 50% of the values falling below the alpha level of 0.05. For the Bayes factors, the mean Bayes factor was above 1000 and the proportion of Bayes factors above 1 was equal to 50%. Similarly, for the second data set, the mean *p*-value was equal to 0.36, with the 38.24% of the values falling below the alpha level of 0.05. For the Bayes factors, the mean Bayes factor was equal to 8.47 and the proportion of Bayes factors above 1 was equal to 29.41%.

Apart from inferential statistics, researchers may be interested in the size of the effect. Although the package provides Cohen's *d* for the *t*-tests and omega squared for the repeated measures ANOVA, we strived to provide a common effect measure so that we can readily compare the results with each other. As such, we transformed the effect sizes of the ANOVAs and the *t*-tests, and their confidence intervals, to η^2 effect size.³ The left panel of Fig. 3 plots η^2 (and corresponding .90 confidence intervals indicated by the whiskers)⁴ for the acquisition training (data set 1: Panel A, data set 2: Panel C) and extinction training (data set 1: Panel B, data set 2: Panel D) phases. Each square represents the mean estimate of the effect, and the whiskers the 90% confidence intervals around that effect (for the data that were used for each test see review analysis section). For acquisition training data in data set 1, the effect sizes for CS discrimination ("CS" effect; CS + vs. CS-) are medium to large and the CS \times time interaction small to medium. In data set 2, the effect sizes for CS discrimination vary between effects close to 0 and large effects. For the CS \times time interaction, effects are either close to 0 or small.

³ For the *t*-test, we transformed the *t*-values to η^2 values using the formula: $\eta^2 = t^2/(t^2 + df)$, and bootstrapped the confidence intervals using the same function. We did not report the effect sizes for the multilevel models, as, to our knowledge, there is not a consensus as to the report of effect sizes for the individual terms of each model.

⁴ Please note that for some effects the whiskers were too small to plot and they are hidden by the size of the box. Also, it is not uncommon for η^2 to have asymmetric confidence intervals, given that by definition the effect cannot be lower than zero.

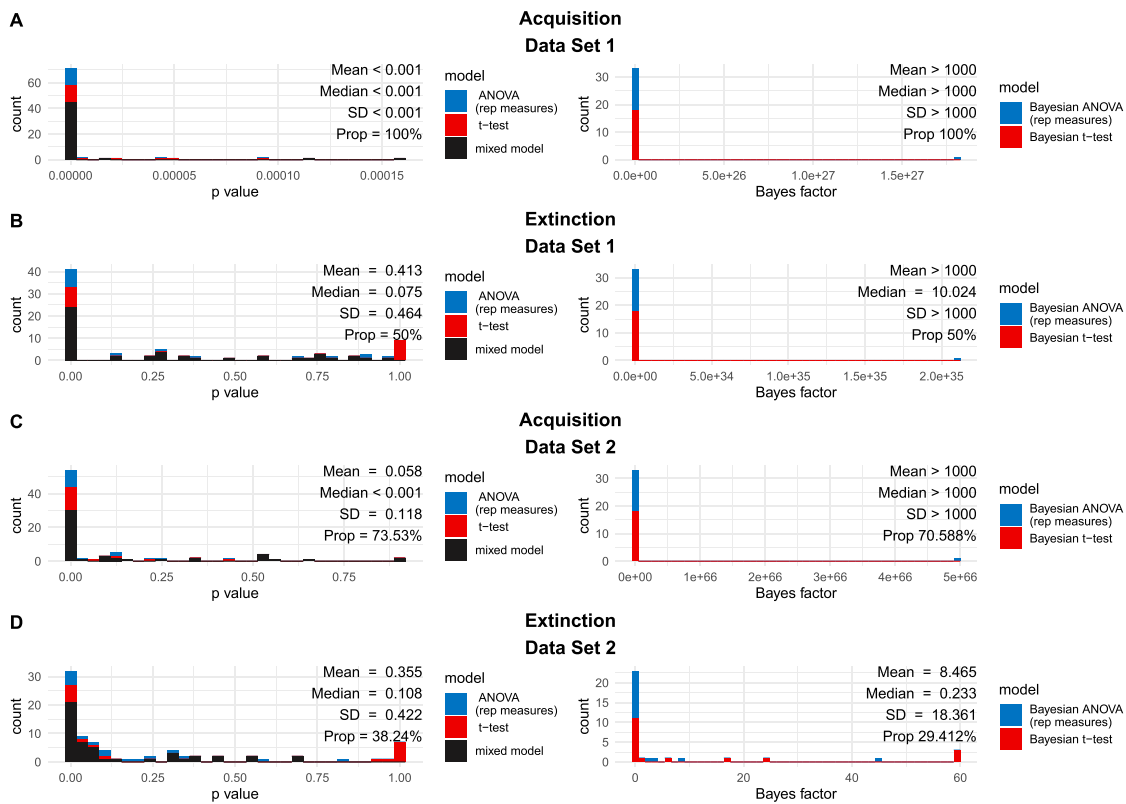


Fig. 2. Histogram of p-values (left panel) and Bayes factors (right panel) of the multiverse analyses for the acquisition training (panel A) and extinction training (panel B) phases of the data set 1, as well as for acquisition training (panel C) and extinction training (panel D) data set 2.

While, for simplicity, we here highlight CS+/CS- discrimination effects only, the R package we introduce also allows for the integration of a group factor, as this is relevant to many research questions. For simplicity, we refrain from showcasing additional analyses including a group factor but refer the interested reader to <https://github.com/AngelosPsy/multifear> for more details.

4. Discussion

In light of a multitude of potentially equally justifiable approaches, there is heterogeneity in and a lack of consensus on the preferred statistical analyses for fear conditioning effects. Typically, researchers select one of these approaches which - in absence of strong empirical and theoretical justifications - result in ambiguity with respect to the robustness of results. Questions like “Would the employment of different exclusion criteria still yield a comparable result” often come to the researcher’s own mind and not seldomly lead to lengthy discussions at the level of peer-review. In this context, “exclusion criteria” can be replaced by “statistical models” (which is the focus of this work), “covariates,” “number of trials” and many other decision nodes a researcher is facing during the scientific process from designing a study, processing the data and selecting a statistical model. Multiverse-type of approaches (Steegen et al., 2016) or specification curve approaches (Simonsohn et al., 2020) meet this challenge by including all (or many) reasonable or equally justifiable decisions in a massive set of tailored robustness analyses.

Model multiverse analyses reveal heterogeneity in results and precision of results: Where to go from here. Here, we present a *model multiverse approach* specifically tailored to fear conditioning research and as a secondary aim introduce the novel and easy to use R package ‘multifear’ that allows to run the multiverse of plausible models (as derived from a systematic literature search) through a single line of code in R. We showcase the idea and value of multiverse-type of studies for the field based on two pre-existing data sets with partial (data set 1) and

100% reinforcement rate (data set 2) by using CS discrimination in skin conductance responses (SCRs) during fear acquisition and extinction training as a case example. Model specifications and data reduction approaches were identified through a representative systematic literature search, which revealed substantial heterogeneity in statistical models employed which we hope to tackle through the ‘multifear’ package in the future. Model multiverse results for both fear acquisition and extinction training showed that a) both the size of the effect as well as the direction of effects (i.e., statistically significant or not) is based on the model that is used, b) that the choice of trials used in the analyses influenced the direction of the results. Even though these results themselves are not utterly surprising, they demonstrate empirically and systematically that indeed analytic flexibility in the analysis of conditioning results influences the direction of the results. This is valuable information that aids fine-tune for future work in the field. To this end, multiverse-type of analyses can be seen as a stopover on the way to develop a formal model that will by consequence result in less heterogeneous approaches for the research field. More precisely, we propose that the results of large-scale multiverse type of work can serve as an optimal starting point for experimental measurement calibration (Bach, Meliščák, Fleming, & Voelkle, 2020), the development of more refined (formal) theoretical frameworks (Oberauer & Lewandowsky, 2019) and the development of formalized computational models (Krypotos, Crombez, Meulders, Claes, & Vlaeyen, 2020). To this end, multiverse-type of analyses are more a means to the end than an end in itself because we need a principled approach that allows us to extract and deliver the information we need to develop a) better theories, b) formal models and identify c) the “best” measure for a given application. The approach we followed here is related to what is referred to as “many-analysts” (Botvinik-Nezer et al., 2020; Silberzahn et al., 2018) approaches which relies on many (teams of) analysts analyzing the same data which typically resulted in a heterogeneous collection of approaches that do not necessarily converge. Here, we have used a related

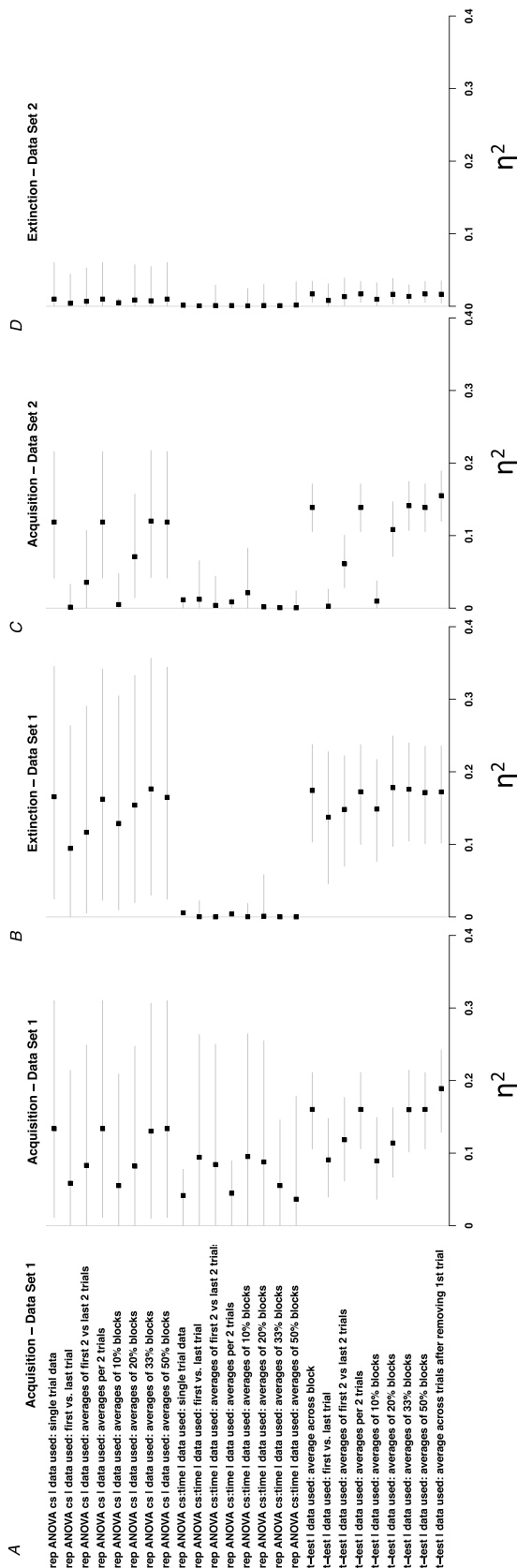


Fig. 3. Forest plots for the multiverse for the acquisition and extinction phases for data set 1 and data set 2.

approach and extracted the approaches typically chosen in the field from the literature which also results in a set of heterogeneous approaches that we then combined into a multiverse analysis and related R package to allow to run all these models with ease. At the first glance, it may seem counterintuitive how adding heterogeneity at the single-study level may be helpful to solve the problem of between-study heterogeneity. Before going into detail on the answers to this question, we first provide some thoughts on how to interpret the results of a multiverse analysis which is a precondition to make use of its results.

How to interpret the outcome of a multiverse-type of analysis?

More precisely, the results of the multiverse of model robustness analyses presented here provide information to what degree different justifiable analytical pipelines yield comparable results - yet it needs to be defined how *comparability* is defined and consequently evaluated. To our knowledge, such a framework for the evaluation of robustness analyses does not exist yet, but, we may borrow some criteria from a framework suggested for the evaluation of replicability (LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018) to the interpretation of the outcome of robustness analyses: LeBel et al. (2018) suggests several criteria for the evaluation of a *replication outcome* (i.e., applying identical methods and analyses to different data derived from a closely identical experiment). The analyses included in the model multiverse presented here may be viewed as different replication attempts by using the same data but applying slightly different procedures (i.e., *robustness* analyses). LeBel suggested to evaluate replication outcomes in terms of there being a ‘signal’ (i.e., effect defined as 90% of the CI includes zero) or not, whether this signal is consistent (i.e., whether the replication’s CI includes the original effect size point estimate) across analyses and its precision (i.e., the width of the CI of the different effects across analyses). From these criteria we can borrow and apply the criteria of precision and consistency⁵ to evaluate the robustness analyses in the multiverse approach presented here.

Evaluating the data presented in Fig. 3 with respect to precision and consistency, we can conclude that the main effect of CS type for acquisition and extinction training (less so the CS × trial effect) provides a rather consistent effect in both data sets. When using the results from the t-test with full data as a reference, only results generated by t-tests (and rmANOVAs in data set 2) with trials divided by 10% and with first vs. last trial (as well as with first 2 vs. last 2 trials in data set 2) during fear acquisition training would not fulfill the criteria of being consistent with the latter providing a less precise estimate as the reference model. This, in principle, is good news for the field, as this would mean that the results are rather *comparable* despite heterogeneity in statistical models applied. Still we want to bring the low precision of the estimates to the reader’s attention - despite both samples having relatively high sample sizes (N = 42 and N = 116) given the standard in the field. Larger sample sizes are expected to generate more precise estimates that could lead to different conclusions with respect to the conclusion of “comparability.”

Our Results come to underline the need for better developing common statistical techniques for our field. Indeed, given the strong translational importance of fear conditioning procedures in guiding future intervention and prevention programs in clinical populations, there is an urgent need to establish procedures for better determining common analytic techniques across studies. This would facilitate or even allow comparisons between studies’ results and thereby potentially promote replicability and a faster translation of fear conditioning research to the clinic.

Deflating the multiverse and towards better theories and formal models

Going back to the question on how adding heterogeneity in analysis pipelines at the single study level can help to tackle the consequences of heterogeneity on the between-study level. We suggest that

⁵ Note that consistency can only be evaluated pair-wise as there is no *original effect* in a multiverse-type of study given that all paths are assumed to be equally justifiable.

multiverse analyses as employed here are only one of several promising ways to battle the lack of consensus in statistical analyses in fear conditioning. Yet, and importantly, multiverse analyses only battle the *consequences* of this lack of consensus by providing a comprehensive overview covering all potentially justifiable models (i.e., robustness analyses). At the core of this lack of consensus and the resulting heterogeneity and uncertainty, however, is a lack of and underdevelopment of formal models for fear conditioning effects. To date, most psychology research is based on verbal rather than formal accounts of theories. This results in flexibility in statistical analyses, as different analyses could be argued to better serve the (often ill-defined) underlying theory. Relatably, it has been highlighted that researchers degrees of freedom mostly do not derive from malicious intent but are mostly due to 'ambiguity in how to best make the decision in question' (cf. Simmons et al., 2011). The development of formal models would get to the roots of data processing and analytical heterogeneity and could present a sustainable approach for battling analytic heterogeneity. Yet, formal models in psychology are used sparingly. As such, multiverse analyses are a pragmatic approach for current research until we have accumulated the necessary empirical evidence to generate formal models. In fact, better theories (Oberauer & Lewandowsky, 2019) about the construct and its measurement (Bach et al., 2020) would serve to deflate the multiverse. This can be achieved through systematic (cross-study) multiverse analyses which may aid the development of formal theories as they may reveal specific models or operationalizations that may consistently impact on variability of the results. Even though we here mainly focus on statistical models, this is related to the idea of calibration experiments that evaluate a measurement method under controlled circumstances and allow choosing the method that yields the highest effect size in independent benchmark experiments (Bach et al., 2020; Bach & Melinscak, 2020) which also may serve the aim to deflate the multiverse.

Critical considerations for multiverse-type of studies. A common point of confusion with multiverse analyses is that they are sensitive to multiple comparisons. However, multiple comparison problems arise when multiple tests are run and only the significant results are highlighted. For example, a researcher runs 20 tests, and only reports the single test that turned out to be significant at $p < 0.05$. However, a multiverse approach is not sensitive to this as all tests run are taken into account when summarizing the results (e.g., computing proportions of values below an alpha threshold). In the above example (i.e., reporting a single significant result from a set of 20 tests), then, the proportion of significant p-values would be 1/20, showing extremely weak evidence for a true effect.

Yet, the multiverse approach employed here has limitations: First, we decided on the statistical models based on a systematic literature search in the field of fear conditioning research. This revealed a heterogeneous set of models employed in the field with some models used very frequently while others are used sparingly. Still, our approach (i.e., average p-values and proportion of studies passing a criterion) gives an equal weight to approaches that are frequently used (e.g., rmANOVA) as well as approaches that are used more sparingly (*t*-tests) without evaluating the individual approaches further. Thus, the inclusion of unjustified specifications may result in *analytical black holes* (cf. Del Giudice & Gangestad, 2021) in which genuine effects might be swallowed in massive analyses that include unjustified or inappropriate decision nodes which then may dilute the effect of the justified or appropriate nodes. Relatedly, selecting statistical models from the literature (as done here) may be susceptible to the impact of publication bias as the published analyses may just represent the set of analyses that are likely to show an effect and consequently made it into a publication.

Second, for conciseness, none of the analyses included here took into account covariates that may have been relevant (e.g., sex or age) but as the package is open source, any models could be added to the multiverse and we explicitly welcome such contributions. Yet, we highlight that analyses with and without covariates do - in a strict sense - not provide answers to the same but to different questions. As a consequence, they

may not be considered *equal* and may not be part of the same multiverse (Simonsohn et al., 2020, Del Giudice and Gangestad, 2021). In a strict sense, however, also the different trial numbers included in the models as employed here may implicitly test different hypotheses such as end point extinction performance or fear recall when using the last or first trial(s) of extinction training respectively. Furthermore, different numbers of trials in a statistical model have consequences for reliability, statistical power of the effect, and the precision of the estimates. The same applies to different sample sizes due to different exclusion criteria (e.g., compare the results of the first and second data set with $N = 38$ and $N = 116$, respectively). This highlights, that it is inherently challenging to define *reasonable* or *equally justifiable* options for a multiverse approach which requires careful consideration (Del Giudice et al., 2020) and which is hampered by the lack of precise theories to guide what can be considered *equally justifiable*. Yet, as discussed above, these problems are not inherent to the multiverse approach but originate from the *researchers degrees of freedom* allowed for by ill specified (verbal) theories. We propose that multiverse-type of analyses (also within a single data-set) can be helpful in deflating the multiverse in providing insights into which paths converge (i.e., are comparable) and which diverge.

Third, we exemplify only strong main effects during fear acquisition and extinction training and it is plausible that more subtle effects (e.g., individual differences, group effects) may hinge more strongly on the selection of the statistical model and may thus yield less comparable results across the multiverse of models. While the accompanying R package 'multifear' allows for the integration of a group-level effect, we have refrained from providing an example there for simplicity and refer to the online tutorial for this (<https://github.com/AngelosPsy/multifear>).

Finally, we provide a minimal attempt to establish a model multiverse that could be derived from aiming to test a single hypothesis. Of note, this does not take into account the multiverse of different data-sets that can be generated from a single set of observations through different data processing decisions such as different ways to quantify SCRs (Kuhn et al., 2022) as well as different transformations or filter settings (Privatsky et al., 2020). The most complete, but also most challenging approach, would be to cross the data- and model multiverse approach to reveal a comprehensive set of p-values, BF's, and/or effect sizes.

Introducing multiverse analyses enabled by the easy-to-use R-package 'multifear' A secondary aim of this work is to introduce the open source 'multifear' package which provides a first step in the direction of enabling computationally demanding multiverse-style analyses in an easy-to-use way. The analyses presented here are can be seen as an illustrative example on how to and why to use the 'multifear' package (see section on deflating the multiverse and towards better theories and formal models).

In our view, the most pressing further extension include the extension of the package to other fear conditioning procedures/phases (e.g., fear generalization), inclusion of covariates, data multiverse analyses based on different transformations or exclusion criteria as well as the inclusion of other outcome measures beyond skin conductance (e.g., startle reflex, ratings). Furthermore, a multiverse of data-collection methods or experimental designs has been recently suggested which also provides an interesting future perspective (Harder, 2020), which is, however, much more demanding with respect to resources as it involves new data-collections and can hence not easily be implemented in 'multifear.' Lastly, our package could be further extended by including continuous predictor effects.

In closing, with the 'multifear' package, we present an easy-to-use tool that allows the easy running of (model) multiverse analyses for fear conditioning studies based on statistical models and data reduction techniques derived from a systematic literature review. We hope that this approach and the 'multifear' package will be used widely in the fear conditioning community and enhance our understanding of the robustness of different analytical approaches employed and ultimately help to enhance comparability between studies and in the long run aid

the development of better theories and formal models.

How to navigate the multiverse. Here, we have showcased the idea, application and value of multiverse type of studies for experimental psychopathology - more precisely the field of fear conditioning research. Of note, the multiverse approach has to be seen as only one way to battle analytic heterogeneity (here: in statistical analyses) which extends beyond other remedies suggested to enhance transparency and robustness of research. More precisely, while pre-registration of a study protocol as well as registered reports enhance transparency of the scientific process, neither of them does counteract the (often) arbitrariness of deciding for one specific statistical model and one specific type of variable operationalization or processing pipeline (Krypotos, Klugkist, Mertens, & Engelhard, 2019). To this end, even though pre-registration and registered reports are certainly useful tools, they provide no information to what extent specific findings hinge on the specific choices made or can be generalized to other processing and analysis paths. Indeed, the pre-registered specifications may neither generate robust, representative or generalizable results. To this end, different remedies and tools proposed to enhance transparency, replicability and/or robustness of research may serve completely different and potentially synergistic purposes.

In closing, we suggest that multiverse type-of analyses can either be run as the major analysis or may be included as an additional supplementary analyses to inform on the robustness of a reported finding. Most importantly, we anticipate that an increase in multiverse-type of studies will guide and aid the development of formal theories (Del Giudice & Gangestad, 2021) through the accumulation of empirical evidence guiding their development which we anticipate to ultimately contribute to a more successful and faster translation of fear conditioning research to clinical applications.

Author note

TBL was funded through grants awarded by the German Research Foundation (DFG) DFG LO1980/7-1, DFG LO1980/4-1, DFG LO1980/2-1, and DFG CRC TRR 58 INST 211/633-2. AMK is supported by a senior post-doctoral grant from FWO (Reg. # 12X5320N) and a replication grant from NWO (Reg. # 401.18.056).

CRedit authorship contribution statement

Tina B. Lonsdorf: Conceptualization, Methodology, Writing – original draft, Project administration, Resources. **Anna Gerlicher:** Resources, Writing – review & editing, Validation, Verification. **Maren Klingelhöfer-Jens:** Resources, Writing – review & editing, Validation, Verification. **Angelos-Miltiadis Krypotos:** Conceptualization, Methodology, Writing – original draft, Project administration, Formal analysis, Visualization, Software.

Acknowledgments

We would like to thank Manuel Kuhn and for data acquisition, processing, and curation, Claudia Immisch for data acquisition, and Irene Klugkist for statistical advice as well Raffael Kalisch and Oliver Tüscher for funding acquisition for data set 1 (CRC1193, subproject C01 to RK and C04 to OT)

References

- Bach, D. R., & Melinscak, F. (2020). Psychophysiological modelling and the measurement of fear conditioning. *Behaviour Research and Therapy*, 127, 103576. <https://doi.org/10.1016/j.brat.2020.103576>
- Bach, D. R., Melinscak, F., Fleming, S. M., & Voelkle, M. C. (2020). Calibrating the experimental measurement of psychological attributes. *Nature Human Behaviour*, 4(12), 1229–1235. <https://doi.org/10.1038/s41562-020-00976-8>
- Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., et al. (2018). Estimating across-trial variability parameters of the diffusion decision model: Expert

- advice and recommendations. *Journal of Mathematical Psychology*, 87, 46–75. <https://doi.org/10.1016/j.jmp.2018.09.004>
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810), 84–88. <https://doi.org/10.1038/s41586-020-2314-9>
- Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E., et al., Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, 49, 1017–1034. <https://doi.org/10.1111/j.1469-8986.2012.01384.x>
- Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, 4(1). <https://doi.org/10.1177/2515245920954925>, 2515245920954925.
- Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P., et al. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, 26(4), 1051–1069. <https://doi.org/10.3758/s13423-017-1417-2>
- Ehlers, M. R., Nold, J., Kuhn, M., et al. (2020). Revisiting potential associations between brain morphology, fear acquisition and extinction through new data and a literature review. *Sci Rep*, 10, 19894. <https://doi.org/10.1038/s41598-020-76683-1>
- Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.
- Gelman, A., & Loken, E. (2014). 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. *Psychological Bulletin*, 140, 1272–1280. <https://doi.org/10.1037/a0037714>
- Gerlicher, A., Tüscher, O., & Kalisch, R. (2018). Dopamine-dependent prefrontal reactivations explain long-term benefit of fear extinction. *Nature Communications*, 9(1), 1–9. <https://doi.org/10.1038/s41467-018-06785-y>
- Haaker, J., Golkar, A., Hermans, D., & Lonsdorf, T. B. (2014). A review on human reinstatement studies: An overview and methodological challenges. *Learning & Memory*, 21(9), 424–440. <https://doi.org/10.1101/lm.036053.114>
- Harder, J. A. (2020). The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, 15(5), 1158–1177. <https://doi.org/10.1177/1745691620917678>
- Klingelhöfer-Jens, M., Ehlers, M. R., Kuhn, M., Keyaniyan, V., & Lonsdorf, T. B. (2022). Robust group- but limited individual-level (longitudinal) reliability and insights into cross-phases response prediction of conditioned fear. *bioRxiv*, reprint. <https://doi.org/10.1101/2022.03.15.484434>
- Krypotos, A.-M., Crombez, G., Meulders, A., Claes, N., & Vlaeyen, J. W. (2020). Decomposing conditioned avoidance performance with computational models. *Behaviour Research and Therapy*, 133, 103712. <https://doi.org/10.1016/j.brat.2020.103712>
- Krypotos, A.-M., Klugkist, I., Mertens, G., & Engelhard, I. M. (2019). A step-by-step guide on preregistration and effective data sharing for psychopathology research. *Journal of Abnormal Psychology*, 128(6), 517. <https://doi.org/10.1037/abn0000424>
- Kuhn, M., Gerlicher, A., & Lonsdorf, T. (2022). Navigating the manifold of skin conductance response quantification approaches – a direct comparison of trough-to-peak, baseline-correction and model-based approaches in LedaLab and PsPM, 2022. *Psychophysiology*, Article DOI:10.1111/psyp.14058. In press.
- LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, 1(3), 389–402. <https://doi.org/10.1177/2515245918787489>
- Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE publications.
- Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, 14(10), 1250–1252. <https://doi.org/10.1038/nn.2904>
- Lonsdorf, T. B., Klingelhöfer-Jens, M., Andreatta, M., Beckers, T., Chalkia, A., Gerlicher, A., et al. (2019a). Navigating the garden of forking paths for data exclusions in fear conditioning research. *Elife*, 8, e52465. <https://doi.org/10.7554/eLife.52465>
- Lonsdorf, T. B., Merz, C. J., & Fullana, M. A. (2019b). Fear extinction retention: Is it what we think it is? *Biological Psychiatry*, 85(12), 1074–1082. <https://doi.org/10.1016/j.biopsych.2019.02.011>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Group, P., & others. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Morris, R. W., & Bouton, M. E. (2006). Effect of unconditioned stimulus magnitude on the emergence of conditioned responding. *Journal of Experimental Psychology: Animal Behavior Processes*, 32(4), 371. <https://doi.org/10.1037/0097-7403.32.4.371>
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3(3), 221–229. <https://doi.org/10.1038/s41562-018-0522-1>
- Ney, L. J., Laing, P. A., Steward, T., Zuj, D. V., Dymond, S., & Felmingham, K. L. (2020). Inconsistent analytic strategies reduce robustness in fear extinction via skin conductance response. *Psychophysiology*, 57(11), e13650. <https://doi.org/10.1111/psyp.13650>
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5), 1596–1618. <https://doi.org/10.3758/s13423-019-01645-2>
- R Core Team. (2013). *R: A language and environment for statistical computing*.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black, &

- W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.
- Sandre, A., Banica, I., Riesel, A., Flake, J., Klawohn, J., & Weinberg, A. (2020). Comparing the effects of different methodological decisions on the error-related negativity and its association with behaviour and gender. *International Journal of Psychophysiology*, 156, 18–39. <https://doi.org/10.1016/j.ijpsycho.2020.06.016>
- Seymour, B., O'doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., et al. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, 8(9), 1234–1240. <https://doi.org/10.1038/nn1527Tzovara>
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect Results. *Advances in Methods and Practices in Psychological Science*, 1(3), 337–356. <https://doi.org/10.1177/2515245917747646>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214. <https://doi.org/10.1038/s41562-020-0912-z>
- Sjouwerman, R., Illius, S., Kuhn, M., & Lonsdorf, T. (2021). A data multiverse analysis investigating non-model based SCR quantification approaches. <https://doi.org/10.31234/osf.io/q24t8>
- Sjouwerman, R., & Lonsdorf, T. (2019). Latency of skin conductance responses across stimulus modalities. *Psychophysiology*, 56(4), e13307. <https://doi.org/10.1111/psyp.13307>
- Sjouwerman, R., & Lonsdorf, T. B. (2020). Experimental boundary conditions of reinstatement-induced return of fear in humans: Is reinstatement in humans what we think it is? *Psychophysiology*, 57(5), e13549. <https://doi.org/10.1111/psyp.13549>
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- Tzovara, A., Korn, C. W., & Bach, D. R. (2018). Human pavlovian fear conditioning conforms to probabilistic learning. *PLoS Computational Biology*, 14(8), e1006243. <https://doi.org/10.1371/journal.pcbi.1006243Zhang>
- Wethels, R., et al. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6(3), 291–298. <https://doi.org/10.1177/1745691611406923>
- Zhang, S., Mano, H., Ganesh, G., Robbins, T., & Seymour, B. (2016). Dissociable learning processes underlie human pain conditioning. *Current Biology*, 26(1), 52–58. <https://doi.org/10.1016/j.cub.2015.10.066>

11 Abstract

Meta-science or meta-research is the study of science's own methods and practices in order to improve research practices and ensure the integrity of scientific processes. Three key concepts in meta-science are reproducibility, robustness, and replication which are crucial for ensuring the credibility of scientific findings. Recently, various studies have consistently revealed that achieving reproducible, robust, and replicable results is still a challenge in psychological research. One aspect that accentuates this challenge is the significant heterogeneity in employed methods which is also referred to as the 'garden of forking paths'. This heterogeneity affects the robustness of research results and makes them difficult to compare, integrate, and generalize. These issues are also present in research addressing anxiety- and stress-related processes.

Research on these processes is of great importance, since anxiety disorders are highly prevalent, causing significant suffering for a large number of individuals and imposing a substantial financial burden on the healthcare system. In the laboratory, the acquisition, treatment, and relapse of fear- and stress-related disorders can be modeled within the classical fear conditioning paradigm. In this paradigm, individual differences in defensive responding – which was regarded as “noise” for decades – can significantly impact fear conditioning processes, and might also play a key role in clinical settings. However, there is considerable methodological heterogeneity in fear conditioning research, and the robustness of findings, particularly concerning individual differences, has been little studied so far. Thus, this thesis aims at bridging that gap by addressing methodological heterogeneity in fear conditioning research and contributing to the accumulation of knowledge on result robustness in a comprehensive series of four studies.

The majority of the data analyzed in these studies were derived from a longitudinal investigation that included a large sample and spanned six measurement time points. Two of these time points involved a two-day differential fear conditioning paradigm, and a variety of outcome measures, including physiological measures (e.g., fMRI and SCR) and self-report data (e.g., fear ratings and questionnaires). While **Study I** demonstrates, that previously reported associations between individual differences in brain structure and defensive responding could not be replicated, **Study II** reveals robust group- but limited individual-level longitudinal reliability of commonly used measures in fear conditioning research. **Study III** highlights the massive heterogeneity in participant exclusion due to 'non-learning' and 'non-responding' and its impact on results and their interpretation, whereas **Study IV** introduces an efficient method to explore methodological heterogeneity systematically by testing various analytic approaches simultaneously.

In summary, while fear conditioning research faces challenges such as robustness and methodological heterogeneity, the studies presented in this thesis offer potential remedies to enhance robustness, reduce heterogeneity, and improve comparability, integrability, generalizability, and replicability of research findings. These remedies involve improving reliability and precision, and, more broadly, promoting transparency in reporting as well as fostering a change in scientific culture. To achieve this, meta- and open science tools play an important role and were also used extensively in the studies included in this thesis. In conclusion, the fear conditioning paradigm has strong potential for clinical use, but more research on fear conditioning research is needed to expand the cumulative meta-scientific knowledge. Ultimately this should advance the progress in the field by obtaining reproducible, robust, and replicable findings that accelerate the translation of fear conditioning discoveries into successful clinical interventions.

12 Zusammenfassung

Meta-Wissenschaft oder Meta-Forschung ist die Untersuchung der Methoden und Praktiken der Wissenschaft selbst, um die Forschungspraktiken zu verbessern und die Integrität wissenschaftlicher Prozesse zu gewährleisten. Drei Schlüsselkonzepte der Meta-Forschung sind Reproduzierbarkeit, Robustheit und Replikation, die für die Glaubwürdigkeit wissenschaftlicher Ergebnisse entscheidend sind. In jüngster Zeit haben verschiedene Studien immer wieder gezeigt, dass es in der psychologischen Forschung immer noch eine Herausforderung darstellt, reproduzierbare, robuste und replizierbare Ergebnisse zu erzielen. Ein Aspekt, der diese Herausforderung noch verschärft, ist die erhebliche Heterogenität der verwendeten Methoden, die auch als "Garten der sich verzweigenden Wege" bezeichnet wird. Diese Heterogenität beeinträchtigt die Robustheit der Forschungsergebnisse und erschwert deren Vergleich, Integration und Verallgemeinerung. Diese Probleme treten auch in der Forschung auf, die sich mit angst- und stressbezogenen Prozessen befasst.

Die Erforschung dieser Prozesse ist von großer Bedeutung, da Angststörungen weit verbreitet sind, bei einer großen Zahl von Menschen erhebliches Leid verursachen und eine erhebliche finanzielle Belastung für das Gesundheitssystem darstellen. Im Labor kann der Erwerb, die Behandlung und der Rückfall von angst- und stressbedingten Störungen mit dem Paradigma der klassischen Angstkonditionierung modelliert werden. Individuelle Unterschiede in der Abwehrreaktion innerhalb des Paradigmas – die jahrzehntelang als Rauschen betrachtet wurden – können die Prozesse der Angstkonditionierung erheblich beeinflussen und könnten auch im klinischen Bereich eine wichtige Rolle spielen. Allerdings gibt es eine erhebliche methodische Heterogenität in der Furchtkonditionierungsforschung, und die Robustheit der Ergebnisse, insbesondere in Bezug auf individuelle Unterschiede, wurde bislang nur wenig untersucht. Ziel dieser Arbeit ist es daher, diese Lücke zu schließen, indem diese methodische Heterogenität in der Furchtkonditionierungsforschung thematisiert wird und in einer umfassenden Serie von vier Studien ein Beitrag zur Akkumulation von Wissen über die Robustheit von Ergebnissen geleistet wird.

Der Großteil der in diesen Studien analysierten Daten stammt aus einer Längsschnittuntersuchung, die eine große Stichprobe umfasste und sich über sechs Messzeitpunkte erstreckte. Zwei dieser Zeitpunkte umfassten ein zweitägiges Paradigma zur differenziellen Angstkonditionierung und eine Vielzahl von Ergebnismessungen, darunter physiologische Messungen (z. B. fMRI und SCR) und Selbstauskünfte (z. B. Angstbewertungen und Fragebögen). Während **Studie I** zeigt, dass zuvor berichtete Assoziationen zwischen individuellen Unterschieden in der Hirnstruktur und defensivem Verhalten nicht repliziert werden konnten, zeigt **Studie II** eine robuste gruppenbezogene, aber begrenzte individuelle longitudinale Reliabilität von häufig in der Furchtkonditionierungsforschung verwendeten Messgrößen. **Studie III** unterstreicht die massive Heterogenität beim Ausschluss von Teilnehmern aufgrund von "Nicht-Lernen" und "Nicht-Reagieren" und deren Auswirkungen auf die Ergebnisse und deren Interpretation, während **Studie IV** eine effiziente Methode zur systematischen Erforschung der methodischen Heterogenität einführt, die verschiedene analytische Ansätze gleichzeitig testet.

Zusammenfassend lässt sich sagen, dass die Furchtkonditionierungsforschung zwar mit Herausforderungen wie Robustheit und methodischer Heterogenität konfrontiert ist, die in dieser Arbeit vorgestellten Studien jedoch potenzielle Abhilfemaßnahmen bieten, um die Robustheit zu erhöhen, die Heterogenität zu verringern und die Vergleichbarkeit, Integrierbarkeit, Generalisierbarkeit und Replizierbarkeit der Forschungsergebnisse zu verbessern. Diese Abhilfemaßnahmen umfassen die

Verbesserung der Reliabilität und Präzision, die Förderung der Transparenz in der Berichterstattung sowie die Förderung eines Wandels der wissenschaftlichen Kultur. Um dies zu erreichen, spielen Meta- und Open-Science-Instrumente eine wichtige Rolle und wurden auch in den Studien, die dieser Arbeit zu Grunde liegen, intensiv verwendet. Zusammenfassend lässt sich sagen, dass das Paradigma der Furchtkonditionierung ein großes Potenzial für die klinische Anwendung hat, dass aber mehr Forschung zur Furchtkonditionierung erforderlich ist, um das kumulative metawissenschaftliche Wissen zu erweitern. Letztendlich sollte dies den Fortschritt auf diesem Gebiet vorantreiben, indem reproduzierbare, robuste und replizierbare Ergebnisse gewonnen werden, die die Umsetzung der Befunde zur Furchtkonditionierung in erfolgreiche klinische Interventionen beschleunigen.

13 Statement of author contribution

- Study I:** Data curation, Visualization, Writing – review and editing
- Study II:** Conceptualization, Data curation, Software, Formal analysis, Visualization, Methodology, Writing - original draft, Pre-registration of the study
- Study III:** Data curation, Formal analysis, Visualization, Writing - original draft, Writing - review and editing
- Study IV:** Data curation, Resources, Writing – review and editing, Validation, Verification

14 Danksagung

An erster Stelle möchte ich Professorin Tina Lonsdorf danken, meiner Doktormutter, die mich schon so lange auf meinem Forschungsweg begleitet. Ich danke Dir für Deine unermüdliche Unterstützung, Dein Verständnis und Deine Geduld während meines Promotionsstudiums. Deine permanente, zeitnahe und unkomplizierte Hilfe und Deine umfassende Expertise haben maßgeblich dazu beigetragen, dass ich diesen Weg (vor allem in diesen Krisenzeiten) meistern konnte. Durch Deine konstruktive Kritik und Dein wertvolles Feedback konnte ich immer weiter an mir arbeiten und meine Arbeit stetig verbessern. Du warst immer für mich da, fachlich und persönlich. All dies in seiner Kombination war und ist für mich sehr wertvoll.

Weiterhin möchte ich mich bei meiner Arbeitsgruppe bedanken – der jetzigen, aber auch der vorherigen, bestehend aus Dr. Manuel Kuhn, Dr. Rachel Sjouwerman und Dr. Robert Scharfenort und mittlerweile aus Dr. Mana Ehlers, Alina Koppold, Alexandros (Alex) Kastrinogiannis, Julia Ruge; Alena Russmann und Maria Bruntsch. Habt vielen Dank für die vielen anregenden Diskussionen und gemeinsamen Erfahrungen, die mir während meiner Promotionszeit sehr viel bedeutet haben. Eure Unterstützung und Eure Anregungen haben maßgeblich dazu beigetragen, dass ich meine Arbeit immer wieder neu reflektieren konnte und unsere Bürogemeinschaft, Alina und Alex (sowie Tatia Buidze), hat immer für viel Inspiration und Freude gesorgt. Ein besonderer Dank geht an Dr. Manuel Kuhn und Dr. Mana Ehlers, die ihre Rollen als Post Docs sehr ernst und mich so wunderbar an die Hand genommen haben. Vielen Dank auch an alle Hiwis und ehemaligen Kollegen, die den Großteil der Datenerhebung übernommen haben – darunter Claudia Immisch, Karoline Rosenkranz, Janne Nold, Kevin Rozario, Habiba Schiller, Hannes Carsten, Jonas Rauh und Linus Kluth.

Ebenso möchte ich mich bei Professor Christian Büchel, ohne den es dieses Institut nicht geben würde, für die Unterstützung und Förderung während meiner Promotionszeit bedanken. Ein herzliches Dankeschön auch an alle noch nicht genannten Kollegen, die das Institut zu einem inspirierenden Ort machen. Vielen Dank an die medizinisch-technischen Assistent:innen Katrin Bergholz, Kathrin Wendt und Waldemar Schwarz, die die Datenerhebung so hilfreich begleitet haben. Hervorheben möchte ich auch das Sekretariat mit Carolina Dlugosch und Heike Path sowie das (ehemalige) IT-Team mit Florian Werle, Mathias Pietsch und Peter Kammer: Vielen Dank für Eure unermüdliche Hilfe bei administrativen Angelegenheiten und für Eure stets freundliche Hilfsbereitschaft.

Ein ganz besonderer Dank gilt vor allem meiner Familie und meinen Freundinnen und Freunden (allen voran Anja, Sabine, Hilke und Simone, Lisa, Inke, Lissa, Pia, Judith, Rebecca und Renée), die mich während meiner gesamten Promotionszeit unermüdlich unterstützt haben, die meine Sorgen verstanden und die Teilerfolge mit mir gefeiert haben. Ohne Euer Verständnis, Eure Geduld, ohne die aufbauenden Gespräche und die Unterstützung in schwierigen Phasen hätte ich meine Arbeit nicht so erfolgreich abschließen können.

Herzlichen Dank auch an die Co-Autoren der Publikationen, die Teil meiner Dissertation sind. Ohne ihre Expertise und ihr Engagement wäre es nicht möglich gewesen, die Ergebnisse so umfassend darzustellen und zu diskutieren. Ich bin stolz darauf, ein Teil dieser vielen engagierten Teams gewesen zu sein.

Insgesamt möchte ich mich bei allen bedanken, die mich während meiner Promotionszeit unterstützt und begleitet haben. Ohne Eure Hilfe wäre diese Arbeit nicht möglich gewesen.

15 CURRICULUM VITAE

– Lebenslauf aus datenschutzrechtlichen Gründen nicht enthalten –

– Curriculum vitae not included for data privacy reasons –

16 Eidesstattliche Versicherung

Ich versichere ausdrücklich, dass ich die Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe.

Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig um Zulassung zur Promotion beworben habe.

Ich erkläre mich einverstanden, dass meine Dissertation vom Dekanat der Medizinischen Fakultät mit einer gängigen Software zur Erkennung von Plagiaten überprüft werden kann.

Unterschrift: