Behaviour Research and Therapy 153 (2022) 104072

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

Multiverse analyses in fear conditioning research

Tina B. Lonsdorf [a,*], Anna Gerlicher [b], Maren Klingelhöfer-Jens [a], Angelos-Miltiadis Krypotos [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 though 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 (Steegen, 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; Steegen 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/AgACl 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 quantication appraoches (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 electrotactile stimulus consisting of three 2 ms electrotactile 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.

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 uninstructed 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 μS 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 + 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://angelospsy.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 with a test of CS × 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.

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 ANOVAof 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

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

[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.

[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).

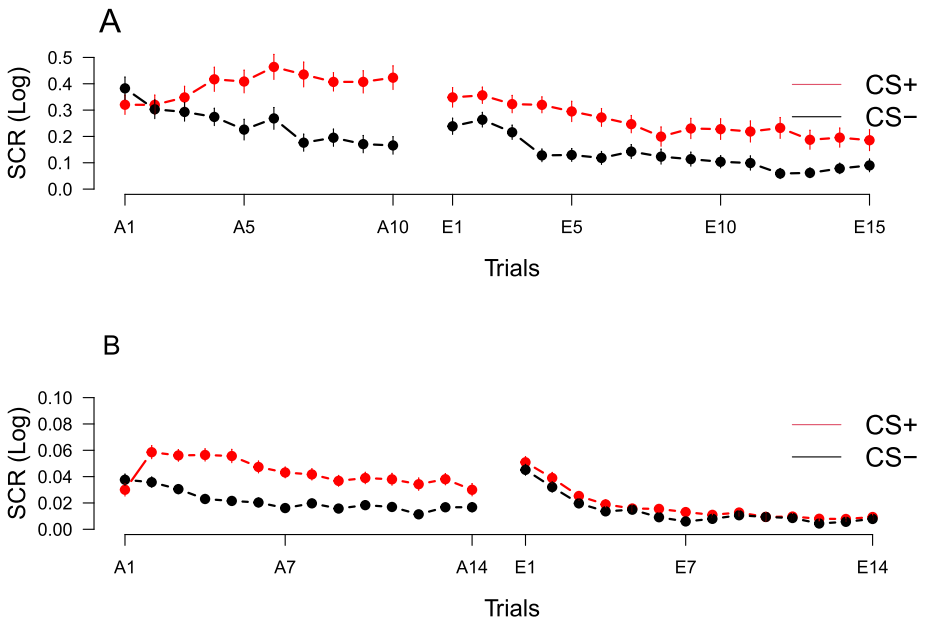

Fig. 1. Depiction of log transformed SCRs per CS (i.e., CS+, CS-) and trial for the Acquisition (i.e., A) and Extinction (i.e, E) training phase for study 1 (A) and study 2 (B).

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 *multifear* 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 multifear 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 *multifear* 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').

multifear::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 C), 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 × 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× 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.

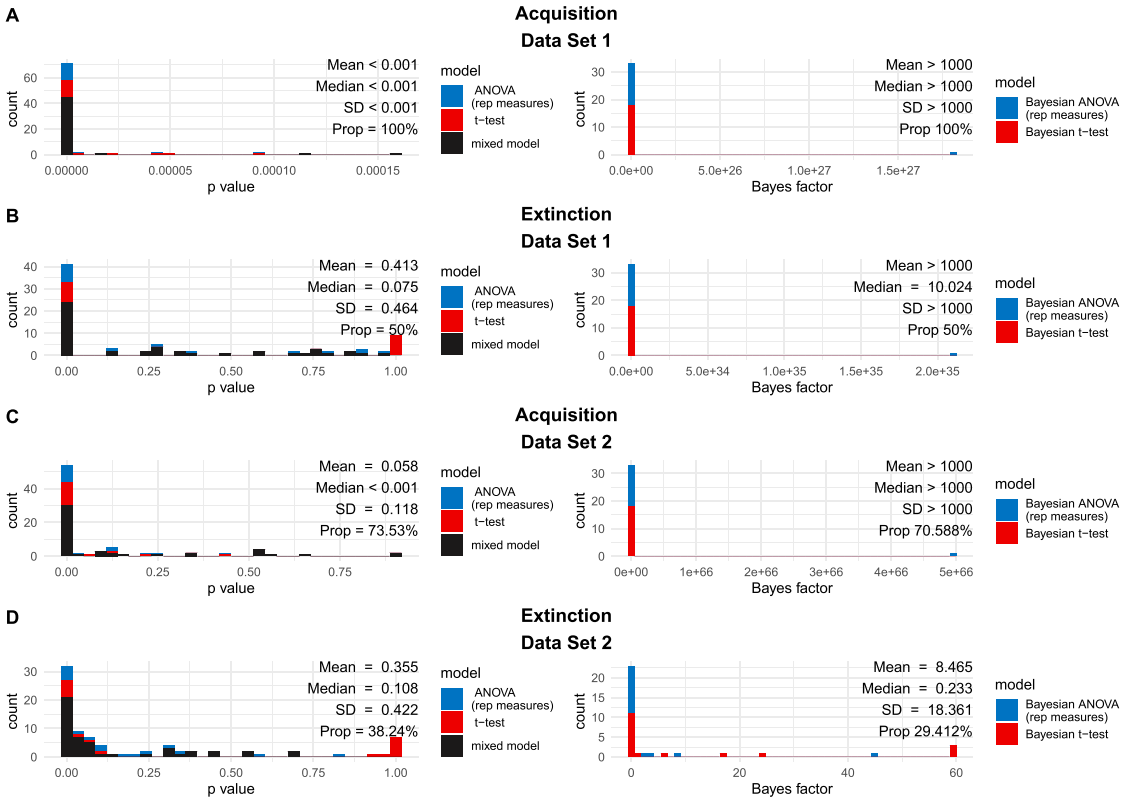

Fig. 2. Histogram of p-values (left panel) and Bayes factors (right panel) of the multiverse analyses for the acquisition training (panel A) and extinction training (panel B) phases of the data set 1, as well as for acquisition training (panel C) and extinction training (panel D) data set 2.

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 nods 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, Melinščak, 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 a 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

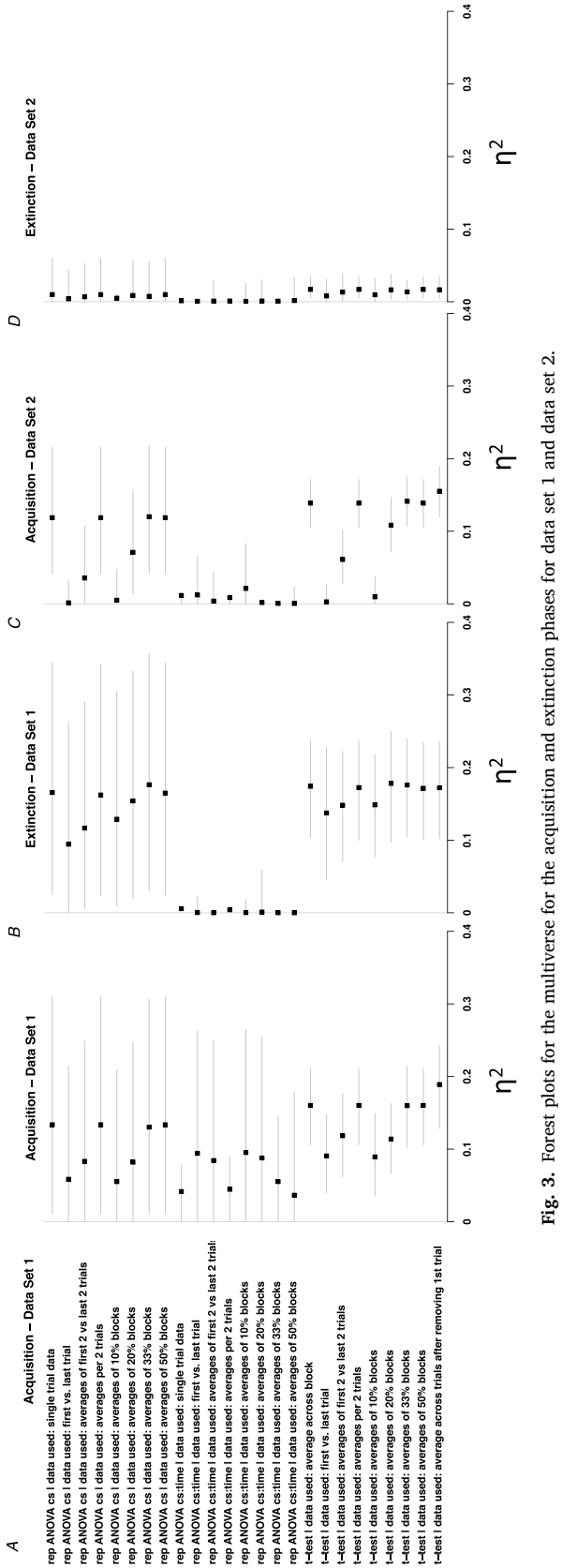

Fig. 3. Forest plots for the multiverse for the acquisition and extinction phases for data set 1 and data set 2.

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 consistency[5] 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 nods which then may dilute the effect of the justified or appropriate nods. 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 Guidice 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 dataset) 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., 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

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

Steegen, 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

Wethzels, 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