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

short title: SCR quantification approaches

Manuel Kuhn[1,2], Anna M.V. Gerlicher [3] & Tina B. Lonsdorf[1]

[1] Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

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

[3] Department of Clinical Psychology, University of Amsterdam, Amsterdam, The Netherlands

Corresponding author: Tina B. Lonsodorf, Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Martinistrasse 52, Hamburg, Germany +49 40 7410 55769, t.lonsdorf@uke.de

Abstract

Raw data have to be processed to be ready for statistical analyses and processing pipelines are often characterized by substantial heterogeneity. Here, we present a systematic literature search on different stimulus-evoked skin conductance response (SCR) quantification approaches used in the literature by using fear conditioning research as a case example. Next, we applied seven different approaches (trough-to-peak scoring, script-based baseline-correction, Ledalab as well as four different models implemented in the software PsPM) to two fear conditioning datasets differing in key procedural specifications (i.e., stimulus duration, reinforcement rate, number of trials). This can be viewed as a set of robustness analyses (i.e., same data subjected to different methods) aiming to investigate if and to what extent these methods yield comparable results. To our knowledge, no formal framework for the evaluation of robustness analyses exists to date, but we may borrow some criteria from a framework suggested for the evaluation of 'replicability' in general. Our results from seven different SCR quantification approaches applied to two datasets suggest that there may be no single approach that consistently yields larger effect sizes across both datasets. Yet, at least some of the approaches employed show consistent effect sizes within each dataset indicating comparability. Finally, we highlight substantial heterogeneity also within most quantification approaches and discuss implications and potential remedies.

1 Introduction

Scientific work rests fundamentally upon data, their measurement, processing, analysis, illustration and interpretation. Raw data have to be processed to be ready for statistical analyses and interpretation. While these processing pipelines can be well defined and standardized, they are often characterized by substantial heterogeneity - particularly in Biological Psychology and Cognitive Neuroscience (Botvinik-Nezer et al., 2020; Sandre et al., 2020). A commonly used measure in these scientific disciplines is skin conductance which is sensitive to emotional arousal, novelty and salience (Dawson et al., 2007) and thought to provide insight into sympathetic activation levels. Skin conductance is characterized by slowly changing tonic activity (skin conductance level, SCL) and faster changing phasic activity with a rather steep incline and slower return to baseline (skin conductance response, SCR). SCRs can occur as spontaneous non-specific fluctuations or stimulus-evoked (Boucsein et al., 2012) with the strength of the latter being the focus of this work. SCRs are typically recorded continuously and subsequently quantified off-line with a multitude of different response quantification approaches available, with any given study typically choosing only one of these options. Already in 1971, Lykken & Venables raised attention to the „[…] disconcerting diversity of electrodermal measurement technique which, at best, make it difficult to compare one set of results with another and sometimes even casts real doubt on the interpretation of the findings." (Lykken D & Venables P, 1971, p. 656). Now, nearly half a century later, basically everything has changed with respect to the equipment and techniques used to record SCRs, while on the other hand, the problem of disconcerting methodological diversity identified in 1971 still persists.

As a consequence, the interpretation of any single set of SCR results is difficult because it may hinge on the specific choices made – as already argued by Lykken half a century ago (Lykken D & Venables P, 1971, p. 656). As a potential solution to the problem of data processing and statistical heterogeneity, the multiverse approach has recently been suggested (T. Lonsdorf et al., 2021; Sjouwerman et al., 2021; Steegen et al., 2016): In data multiverse analysis, the same raw data is processed into a multiverse of datasets depending on different processing choices – all potentially equally reasonable in light of the absence of empirical and/or theoretical criteria to guide the researchers' decisions. This data multiverse inevitably implies a multiverse of statistical results given identical raw data and applied statistical models (T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019; T. B. Lonsdorf, Merz, et al., 2019; Silberzahn et al., 2018; Steegen et al., 2016) and can inform on the stability or robustness of the effect of interest against different processing pathways (see Del Giudice & Gangestad, 2021 for a critical discussion of multiverse-style analyses). Here, we focus on a small-scale multiverse-type of approach by comparing SCR quantification

approaches derived from a systematic literature search in two datasets and by using fear conditioning research as a case example.

1.1 Different response quantification approaches for skin conductance responses

The different currently employed approaches for SCR quantification can be roughly grouped into i) trough-to-peak (TTP) scoring ii) computational approaches such as Ledalab and PsPM and iii) what we here refer to as 'baseline correction' approaches. Of note, however, these approach categories are by no means homogeneous and we refer to our related work for an in-depth investigation of within-approach heterogeneity of specifications in the baseline correction approach (Sjouwerman et al., 2021). In the literature, these different approach categories are generally treated interchangeably despite the lack of empirical support for their equivalence in capturing the same underlying construct and biological process (jingle fallacy) - a problem that has been discussed in fear conditioning research (T. B. Lonsdorf, Merz, et al., 2019; T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019; Ojala & Bach, 2019; Sjouwerman et al., 2021) as well as for related fields in psychology and the neurosciences (Botvinik-Nezer et al., 2020; Garrett-Ruffin et al., 2021; Sandre et al., 2020). In the following, we briefly introduce these three different SCR quantification approach categories: trough-to-peak, computational and baseline correction approaches (as well as their sub-categories).

1.1.1 Trough-to-peak (TTP)

'Trough-to-peak' (TTP) scoring of SCRs quantifies the difference between the skin conductance at the peak of a response and its preceding trough in pre-specified time-windows according to a published set of criteria and publication recommendations (Boucsein et al., 2012): The onset latency, that is the footpoint of the SCR, is typically required to occur in a onset latency time-window (OLW) of 1-3s (Levinson & Edelberg, 1985), 1-3.5s (although stimulus-specific response windows were suggested, Sjouwerman & Lonsdorf, 2019), or 1-4s (Boucsein et al., 2012) after stimulus onset. The SCR peak value is then required to occur in a peak detection time window (PDW) of 0.5-5s after SCR onset (i.e., footpoint, Boucsein et al., 2012). More precisely, if the footpoint occurs 2s after the stimulus presentation the peak must occur in a time window of 2.5-7s after stimulus onset. In addition to the OLW and PDW, a minimum response - typically varying between 0.05µS to 0.01µS - is often applied (Boucsein et al., 2012; T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019). SCRssmaller than this minimum response are not considered as a valid response and included as non-response with a value of zero (T. B. Lonsdorf et al., 2017; i.e., "magnitude", Venables, P.H. & Christie, M.J., 1980). Consequently, TTP scoring can only yield SCR values with a zero or positive value. ~~magnitude~~

TTP scoring employing the above described criteria can be performed a) manually in most recording software, b) computer-assisted with the help of graphical user interfaces (commonly custom-made) which provide editable suggestions for each SCRs footpoint and peak, or c) supervised, but fully automatized ("Autonomate", Green et al., 2014)– even though the latter can also be used as a graphical user interface for visual inspection and/or computer-assisted scoring. Furthermore, d) also fully automatized custom-made scripts are employed. Automatized approaches iteratively apply the published TTP criteria (Boucsein et al., 2012) 2) while systematically dealing with the challenge of overlapping SCRs by searching for patterns in inflection points (Green et al., 2014). Fully automatized TTP scoring consequently reduces some of the drawbacks inherent to manual or computer-assisted (semi-manual) TTP scoring: being time consuming, sensitive to the scale invariance problem (i.e, depending of the scale used to view the data different inflection points may be detected through visual inspection), requiring long inter-stimulus intervals to avoid overlapping responses, and being susceptible to human bias. We highlight that most of the work on skin conductance response dates back to early research in the 70ies and new work has not re-investigated assumptions regarding and SCRs shape and temporal profile with newer technical equipment in detail.

1.1.2. Baseline correction (BLC) approach

In addition, an approach that we here refer to as 'baseline-correction approach' has been suggested that "does not require undertaking the complex process of mathematically modeling [skin conductance] data curves, identifying points of inflection that define a response onset and creating, or learning to use, software that accomplishes this process"(cf. Pineles et al., 2009 p. 993). Pineles suggested the use of an 'entire-interval response' which scores the highest SCR peak in the entire stimulus presentation time window (Pineles et al., 2009). The BLC approach suggested by Pineles employs an algorithm that identifies a response onset by stepping forward (or backward) until the slope changes from negative to positive (or from positive to negative). A response peak is found by locating the highest SC value after the identified onset and within the window specified for the peak (Pineles et al., 2009). Importantly, neither the onset nor the peak may be located at the first or last data point of their respective windows and if this happens, the algorithm will look for a new onset and peak in a shrunk window. If the window is iteratively shrunk to a zero-width, no response is calculated (i.e., zero). The entire interval response suggested by Pineles is accordingly calculated by subtracting the mean skin conductance level for the 2s immediately preceding stimulus onset from the highest SC level value during the entire stimulus presentation period (i.e., 8s; Pineles et al., 2009). Of note, this procedure can yield negative values when no stimulus-bound SCR is

observed or when it is comparably smaller than the (habituation) drift in SCRs. Some authors set these negative responses to "zero" during post-processing (Vogel et al., 2015). Now, BLC approaches are also often performed with custom made scripts that do not follow iterative algorithms, calculate the baseline in a pre-CS time window and subtract this baseline from the post-CS peak identified during a post-CS time window (Sjouwerman et al., 2021).

1.1.3 Computational or model-based approaches

Lastly, computational or model-based approaches are available in different software-packages for instance Ledalab (Benedek & Kaernbach, 2010a; Lim et al., 1997) and PsPM (Bach et al., 2009, 2013; Bach & Friston, 2013)(formerly labelled SCRalyze; Bach, Flandin, Friston, & Dolan, 2009) or cvxEDA (Greco et al., 2016). These approaches rely on (generative or forward) models that specify how a physiological or psychological state generates an observable skin conductance response and use model inversion to estimate these states from the data. The different model-based approaches differ in respect to the exact properties of the employed SCR response function, the treatment of slow-drifts in SCR data, the treatment of observation noise, and the applied model inversion. However, they all generally offer the advantage of automaticity and computational reproducibility. Furthermore, they are thought to improve discriminability of overlapping SCRs in paradigms with short inter-stimulus intervals as SCRs are slow responses, rapidly spaced stimuli with an inter stimulus interval (ISI) of 2-3s do not elicit visually distinguishable SCR peaks and generally appear as a single response (Benedek & Kaernbach, 2010a) – commonly referred to as overlapping responses.

Specifically, deconvolution-based approaches, such as *Ledalab*, decompose skin conductance data into slowly-varying tonic and fast-varying phasic activity (Benedek & Kaernbach, 2010a; Lim et al., 1997). The phasic component is suggested to reflect the time-course of sudomotor or sympathetic nerve activity. The latter is characterized by a zero-baseline and shorter time constant than the resulting SCR, making it possible to discern closely succeeding responses in rapid, quickly-spaced events with an ISI <3s. Ledalab offers a variety of different measures to quantify skin conductance responses within a defined response window, among them the estimated amplitude (which may differ from a TTP approach), sum of all SCRs detected, the average, the peak, and the area under the curve of the phasic driver response.

The software package PsPM (formerly SCRalyze) offers two different approaches: a general linear model (GLM) approach (Bach et al., 2009) and a non-linear dynamic causal modelling (DCM) approach (Bach et al., 2010). The GLM approach models event-onsets as delta-functions, convolves the onset-regressor with a canonical (or data-based) skin conductance response function and fits the data to the resulting time

series (Bach et al., 2009). Depending on whether the GLM onset-regressors comprise all trials of one condition ('condition-wise') or only one individual trial ('trial-wise'), the resulting parameter-estimates reflect condition-specific (e.g., CS+, CS-) or trial-specific SCR magnitudes (e.g., CS+ trial 1, CS+ trial 2, …, CS- trial 1). The non-linear DCM approach provides a causal model that describes how different inputs to sudomotor activity (e.g., spontaneous, evoked, anticipatory responses) map onto skin conductance data. Via model inversion the most likely contribution of each of these components to the observed data is estimated. For discussion and empirical evaluation of differences between Ledalab and the GLM or DCM approach implemented in PsPM, we refer to other sources (Bach, 2014; Bach et al., 2013; Staib et al., 2015).

1.1.4 Comparison between different SCR response quantification approaches

To date, few comparative studies addressing different SCR quantification approaches exist – and those that we are aware of (see Table 1 for a summary) all come from authors that have developed one of the approaches and performed comparisons for means of validation. What is striking from Table 1, is that even those comparative attempts are characterized by substantial heterogeneity with respect to the used SCR quantification approaches and experimental specifications and it is noteworthy that conclusions derived from these studies are as heterogeneous as the methods included. While Green and colleagues concluded that all methods produced comparable effect sizes and hence suggest that a number of suitable methods and software tools exist for SCR quantification analysis of SCRs" (Green et al., 2014), Bach and colleagues in contrast concluded that all SCRalyze basedmodel-based methods are more sensitive than the peak-scoring approach and providesignificantly higher predictive validity than any Ledalab measure in most of the tested contrast (Bach, 2014).

Potential explanations for these conflicting conclusions can be derived from Table 1: First, studies focused on CS discrimination between CS+ and CS- (Bach, 2014) or on SCRs averaged across stimulus types (Green et al., 2014). Second, computer-assisted TTP scoring, supervised automatized TTP scoring (Autonomate) and the GLM approach as implemented in SCRalyze were compared in one study (Green et al., 2014) while another one compared SCRalzye's GLM approach against unsupervised script-based TTP scoring (Bach, 2014; Bach et al., 2013) and script-based BLC approaches (Bach et al., 2010, 2013) but never to computer-assisted TTP scoring. Third, the paradigms employed in the studies differ in experimental strength with fear conditioning paradigms and film clips (Green et al., 2014) eliciting rather strong SCRs (Green et al., 2014) while passive picture viewing (Bach, 2014; Bach et al., 2013) typically elicits less strong SCRs. Fourth, CS duration differs markedly between studies with very short durations of 1s (Bach, 2014; Bach et al.,

2013), 4s duration (Green et al., 2014) and variable durations (Bach et al., 2010: 3.5 and 4-16s in two different datasets). Similarly, ITI length varies substantially between studies. Bach and colleagues have shown previously that short CS-US intervals (≤4s) may be suboptimal for the GLM approach implemented in SCRalyze (Bach et al., 2010; Bach & Melinscak, 2020) and consequently different regressors were built for CS+ trials reinforced with the US and not reinforced with only the latter included in statistical analyses (Bach et al., 2010). In contrast, both reinforced and non-reinforced CS+ trials were considered for statistics by Green et al. (2014). Lastly, the substantially larger number of trials in the fear conditioning paradigms employed by Bach and colleagues (i.e., 32- 90 per CS type; Bach et al., 2010) as opposed to the other fear conditioning paradigms (5 to 16 trials per CS type; see Table 1) may underlie different outcomes and conclusions due to habituation effects and/or differences in statistical power.

1.2 Overarching aim

Our work departs from the lack of conclusive and comprehensive comparative work addressing the question if and to what extent different SCR quantification approaches can be used interchangeably (jingle fallacy) as well as from the lack of empirical agreement on the existence of a single superior SCR quantification approach. Particularly in light of recent discussions on measurement challenges and their potential contributions to (non-) replicability (Flake & Fried, 2020), it is particularly timely to investigate to what extend a given effect can be formally 'replicated' by subjecting a single datasets to multiple theoretically equally justifiable SCR response quantification approaches (i.e., robustness analyses).

First, we need a synopsis of the different approaches of each approach employed in the literature as well as their abundance. Here, we provide an exemplary systematic literature search focusing on different SCR quantification approaches by using fear conditioning research as a case example.

Second, we provide an independent evaluation of seven commonly used and equally justifiable SCR response quantification approaches applied to two datasets differing in key procedural specifications that may impact on overlapping responses (i.e., CS duration, reinforcement rate) or reliability measures (i.e., number of trials, number of participants) . Note that we do not aim for a comparison between both datasets as these are characterized by different specifications but provide a small-scale multiverse analysis within each dataset. To our knowledge, no formal framework for the evaluation of robustness analyses exists to date, but we may borrow some criteria from a framework suggested for the evaluation of 'replicability' in general (LeBel et al., 2018), as robustness can be viewed as a sub-aspect of replicability.

Third and finally, we include TTP scoring from two independent raters per datasets (one experienced and one first-time rater) to address the question if computer-assisted TTP scoring is reproducible (i.e., obtaining 'the same' result when applying the same method to the same data).

If we find evidence for robustness of the results across the different SCR response quantification approaches this would argue in favor of being able to use the different SCR quantification approaches interchangeably. This would be really good news for the field. If we in turn observe a lack of robustness as defined by the above criteria, we have identified a challenge that we can then take into account when interpreting results and as a field working collaboratively towards solving this challenge.

Table 1: Overview over comparative work on different SCR response quantification approaches.

1 **OLW:** onset latency window (post-CS with CS onset serving as 0s), **PDW:** peak detection window (post SCR onset with SCR onset serving as 0s)

2 **BL**: baseline (prior to stimulus onset with stimulus onset serving as 0), Note that some publications refer to this as (standard) peak scoring (Bach et al., 2013; e.g., Privratsky et al., 2020) or peak measure (Bach et al., 2010). To avoid confusion between BC and TTP approaches which use different time windows, we do not adopt the term "peak scoring" here which has at times been used to subsume TTP and BC approaches (e.g., Privratsky et al., 2020).

3 **RI rate**: reinforcement rate (i.e., percentage of CS+ presentations paired with the US); only in fear conditioning paradigm.

4 **Acq**: Acquisition training, **Ext:** extinction, **Pre**: pre-conditioning, **Gen**: generalization, **Ren:** renewal

5 A condition-wise GLM was computed with one regressor for each experimental condition. In addition, two variants of single-trial GLMs were employed: (1) a single-trial GLM with one regressor for each trial, and (2) one single-trial GLM for each trial (i.e., number of trials = number of GLMs) with one regressor for that trial and one regressor for the remaining trials. As the authors state that the "the convolved design matrices underwent the same processing steps applied to SCR data in respective processing pipelines, as recommended previously (Bach, 2014; Staib et al., 2015)", we assume that SCRs were based on "reconstructed time-series" (see footnote [14] for details).

6 only non-reinforced trials were used for analyses across SCR response quantification approaches

7 SCR responses were scored using the event-related EDA analysis routine in the software program Acknowledge version 4.1 (Biopac). Scores were averaged across two manual raters.

8 Continuous decomposition analysis (CDA; Benedek & Kaernbach, 2010a) as implemented in Ledalab version 3.28 was run and SCRs reconstructed from an estimated driver of phasic activity were generated for each trial.

9 General linear models (GLM) utilizing a canonical impulse response function as implemented in SCRalyze version b2.1.3 were solved using the pseudoinverse, yielding parameter estimates for each condition in all subjects.

10 Stimuli were averaged across all trials and conditions (i.e, CS types) per phase for analyses. Consequently, the focus was not on CS+/CS- discrimination.

11 In addition to the first interval response (FIR) for which scoring criteria are reported in the table, also SIR (OLS_{SIR}: 4-8s PDW_{SIR}: 5-9.5s) and EIR using the entire CS-UCS interval (0-8s) were calculated. Response quantification for the EIR was performed in the software program Mathematica 6 with an iterative algorithm that identifies a response onset by finding the 'point of maximum curvature of the SCL data within a pre-specified onset window and then stepping forward (or backward) until the slope changes from negative to positive (or from positive to negative). This point of slope change defines the response onset. A response peak is found by locating the highest SC value after the identified onset and within the window specified for the peak" (cf. Pineles et al., 2009 p, 989). Importantly, neither the onset nor the peak may be located at the first or last data point of their respective windows and if this happens, the algorithm will look for a new onset and peak in a shrunk window. An exception is when the data are flat at the onset for a minimum of .03s when the onset occurs at the first datapoint. If the window is iteratively shrunk to a zero-width, no response is calculated. The EIR suggested by Pineles is accordingly calculated by subtracting the mean skin conductance level for the 2s immediately preceding stimulus onset from the highest SC level value during the entire stimulus presentation period (i.e., 8s; Pineles et al., 2009).

12 Response quantification was performed in the software program Matlab.

13 Mean of 4 measures using CDA & DDA approaches (DDA1, DDA2, CDA1, CDA2): Ledalab does not recommend one single estimate of sympathetic arousal but offers a choice of 4 measures. These were all analysed, without correction for multiple comparison. More precisely, both discrete decomposition analysis (DDA; Benedek & Kaernbach, 2010b) and continuous decomposition analysis (CDA; Benedek & Kaernbach, 2010a) was performed. Following the approach in the validation papers, SN peaks were extracted within a response window of 1–4 s after stimulus onset with a minimum threshold of 0.01 µS. Ledalab's response window refers to the time window during which the response is initiated and peaks. The respective SA indices for each method are: a) **AmpSum**: sum of SCRs of above-threshold in DDA 1 and b) in CDA 1, c) **AreaSum**: sum of SCR area of above-threshold SCRs (DDA 2), d) **SCR**: the average phasic driver (CDA 2). These four measures were then averaged across trials including zero responses, within each condition, as estimates of mean SA in this experimental condition.

14 Note that sympathetic arousal is not a GLM parameter estimate here, but the peak response amplitude of a reconstructed CS-specific time-course. More precisely, the CS-specific time-courses were reconstructed by multiplying the canonical skin conductance response function (SCRF) and its temporal derivative by their respective CS-specific GLM parameter estimates, and adding the thereby created (parameter-weighted) responses. '**Mean SA**' was calculated as peak amplitude for each experimental condition based on the reconstructed time-series.

15 *Magnitudes* including zero responses as well as *amplitudes* with excluding responses < 0.01µS were calculated. The software program used was not specified but it is assumed that Matlab was used.

16 A mean GLM was computed with one regressor for each experimental condition using either a canonical SCRF, SCRF with time derivative, SCRF with time and dispersion derivative, finite-impulse response (FIR) basis set with 15 or 30 sec post-stimulus time bins of 1 sec duration, cosine 4th or 8th order, or a subject-specific response-function. Furthermore, different filter settings were compared (i.e., uni- and bidirectional Butterworth high-pass filter at .005, .1, .0159 Hz, and from .02 to .10 Hz in steps of .005 Hz. In addition, for experiment 1 single-trial GLMs were computed with one regressor for each trial using either the SCRF or the SCRF with time derivative and a high-pass filter cut-off of .05 Hz.

17 Informed DCM: the DCM is informed about stimulus onsets and SN burst are assumed to occur 2000ms after stimulus onset; Uninformed DCM: the DCM is not informed about stimulus onsets and SN burst are assumed to occur anywhere within the experiment; each estimated SN response that caused an SCR that fell into a 1-4s post stimulus time window was extracted and the largest response in each time window was retained.

18 Two GLMs were computed with one regressor for each experimental condition using either the canonical SCRF or the SCRF with time derivative. In addition, a single-trial GLM was computed with one regressor for each trial using either the SCRF or the SCRF with time derivative. The parameter estimates of single-trial GLM were either directly entered into the comparison or were used as a basis for reconstructing the peak of the estimated response (see footnote 12).

19 SN burst assumed to occur during the CS duration (i.e., in an interval between CS onset (i.e., 0 sec) until the offset of the CS) which we refer to as "DCM full duration"

20 When the CS duration was <5s, then a PDW of 5s was used

	Trough-to-peak (TTP)			Computational				Validation datasets				Duration in s	
Author	**Computer assisted**[1]	**Autonomate**	**script-based**[1]	**Ledalab**	**SCRalyze GLM**	**SCRalyze DCM**	**BLC**[2]	**Paradigm**	**N**	**RI rate**[3]	**N°trials per condition**[4]	**CS/ cue**	**ITI**
Kuhn et al. (present study)	✓ $OLW_{CS(1)}$: 0.9-3.5s $OLW_{US(1)}$: 0.9-2.5s $OLW_{(2)}$: 0.9-4s PDW: 0-5s	✗	✗	✓ CDA, response window: 0.9-4s	✓ single-trial GLM (parameter estimates)	✓ full interval, onset only, restricted interval	✓ BL: -2s $PDW_{(1)}$: 6s $PDW_{(2)}$: 4.5	1. Fear conditioning (Lonsdorf, Klingelhöfer-Jens, et al., 2019)	118	100 %	Acq: 15	6-8	10-16
								2. Fear conditioning (Gerlicher, Tüscher, & Kalisch, 2018)	38	50%	Acq: 10	4.5	17, 18 or 19
(Privratsky et al., 2020)	✗	✗	✗	✗	✓[5] condition-wise GLM (reconstructed) two single-trial GLMs (reconstructed)	✗	✓ BL: -2s $PDW_{(1)}$: 1-3.5s $PDW_{(2)}$: 1-4s $PDW_{(3)}$: 1-4.5s $PDW_{(4)}$: 1-5s	1. Fear conditioning	65	50%[6]	Pre: 6 Acq: 18 Ext: 18	3	2-6s

(Green et al., 2014)	✓[7] OLW: 1-4s PDW: 0.5-5s	✓ OLW: 1-4s PDW: 0.5-5s	✗	✓[8] CDA	✓[9] condition-wise GLM	✗	✗	1. Fear conditioning (Huff, Hernandez, Blanding, & LaBar, 2009)[10]	20	31.3 %	Pre: 4 Acq: 16 Ext: 16 Ren: 16	4	11 ± 4
								2 Probablistic classification learning (Thomas & LaBar, 2008)	20	--	Day 1: 50 Day 2: 50	4	5.5 mean
								3 Fear conditioning (Dunsmoor, Mitroff, & LaBar, 2009)[10]	20	60%	Pre: 6 Acq: 10 Gen : 9	4	5-10
								4 Film clips (spontaneous SCRs) (Kragel & Labar, 2013)	20	--	2 per 6 conditions	120	--
(Pineles et al., 2009)	✗	✗	✓[11]	✗	✗	✗	✓ BL: -2s	Fear conditioning	287	100 %	Pre: 5 Acq: 5 Ext: 10	8	15-25

			OLW_{FIR}: 1-4s PDW_{FIR}: 2-6s				PDW: 0-8s						
(Bach, 2014)	✗	✗	✓ 12 OLW: 1-3s PDW: 0.5-5s	✓ 13 response window: 1-4s Mean of 4 measures (CDA, DDA)	✓ 14 condition-wise GLM (reconstructed)	✗	✗	1. Passive picture viewing (IAPS, neutral, aversive)	60	--	45	1s	7.65, 9, 10.35
								2. Passive picture viewing (IAPS, neutral, aversive, positive)	38	--	16	1s	4.4
								3. Passive face viewing (KDEF, angry, fearful, neutral)	42	--	38	1s	7.65, 9, 10.35
								1. 4. Neutral picture viewing (IAPS, neutral)	61	--	45	1s	7.65, 9, 10.35
(Bach et al., 2013)	✗	✗	✓ 15 OLW(1): 1-3s OLW(2): 1-4s PDW: 0.5-5s	✗	✓ 16 condition-wise GLM (parameter estimates) single-trial GLM	✓ 17 DCM informed, DCM uninformed	✓ BL: -1s OLW: 1-4s	1. Passive picture viewing (IAPS, negative, neutral)	60	--	45	1s	7.65, 9, or 10.35
								2. Passive picture	40	--	16	1s	4.4

					(parameter estimates)	about event onsets		viewing (IAPS, neutral, aversive, positive)					
(Bach et al., 2010)	✗	✗	✗	✗	✓ [18] condition-wise GLM (parameter estimates) single-trialGLM (parameter estimates /reconstructed)	✓ [19] full interval, first/second half of interval	✓ BL: -1s PDW: CS duration [20]	1. Fear conditioning	32	50% [6]	32	4, 10 and 16	14, 19 or 23
								2. Fear conditioning	20	50% [6]	90	3.5	7, 9 or 11

2 Methods

2.1 Systematic literature search

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 March 22nd 2019. This systematic literature search was performed to derive data intended to serve as case examples for a number of research projects such as our recently published work (T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019) and the present work. As described in Lonsdorf et al., (2019), the following search terms were used: 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. The original study was included in case author corrections were published within the search period unless the study itself was already included. From the identified 854 records listed in PubMed, stage 2 screening (abstract) included 152 records. For stage 3 screening (full text), 86 were retained. Screening served the aim that the final set of studies consisted of 50 records that reported results for (1) SCRs as an outcome measure from (2) the fear acquisition training phase (3) in human participants (a flow chart with details has been published in (T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019). A subset of the identified SCR quantification approaches was subsequently applied to two independent datasets (see below for details).

2.2 Participants and Experimental paradigms

2.2.1 Dataset 1: Hamburg

Participants: Dataset 1 consisted of the acquisition phase (i.e., day 1) from the baseline (T0) measurement of a longitudinal fear conditioning study in 120 participants. Data from two participants were excluded due to protocol deviations leaving 118 participants for analyses (78 females, mean ± SD age of 24.38 ± 3.7 years). 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). Dataset 1 has been included as a case example in a previous publication (T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019) focusing on methodological questions (i.e., exclusion of 'non-learner' and 'non-responder' in fear conditioning research).

Paradigm and Stimuli: The paradigm (for details see T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019) consisted of a two-day uninstructed fear conditioning paradigm with habituation and acquisition training taking place on day 1 and extinction training and recall 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. During all experimental phases, BOLD fMRI, fear ratings (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 exclusively on the methodological question of different approaches to SCR quantification and only data from the fear acquisition training phase at T0 were included. All datasets were trimmed to this period of interest starting 2s prior to the first event of interest (i.e., first CS presentation during acquisition training) and ending between 10s and 20s (20s trim cutoff value) after the last event of interest (i.e., last CS or US presentation during acquisition training). 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). Trial order was randomized in such a way that not more than two trials of the same type (i.e., CS +, CS−) succeeded each other. Allocation of the two visual stimuli to CS+ and CS– was counterbalanced between participants and the CS+ was followed by the US in all cases during fear acquisition training (100% reinforcement rate). 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 gray background and controlled by Presentation software (Version 14.8, Neurobehavioral Systems, Inc, Albany California, USA).

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.

2.2.2 Dataset 2: Mainz

Participants: Forty male participants (mean ± SD age of 28.1±2.7 years) were included in the dataset which was published previously (Gerlicher et al., 2018). All participants provided written informed consent and the protocol was approved by the local ethics committee (Ethikkommission der Landesärztekammer, Rheinland-Pfalz). Data of 2 participants on day 1 (fear acquisition) were excluded from the analyses of SCR data presented in this work due to recording artefacts, leaving the data of n = 38 participants for statistical analysis of each phase.

Paradigm and stimuli: Dataset 2 consists of a three-day 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 details see Gerlicher et al., 2018) with only the fear acquisition training phase used for the present work. During all experimental phases, BOLD fMRI, expectancy ratings (before and after each experimental phase) and skin conductance data were acquired. BOLD fMRI as well as expectancy ratings are, however, not included in the present work, as it focuses exclusively on the methodological question of different approaches to SCR quantification. Two black geometric symbols (square, rhombus) served as CS+ and CS− and were presented in the center of a computer screen. The CSs were superimposed on background pictures of either a kitchen or a living room. Assignment of symbols to CS+ or CS− and rooms to background pictures were randomized between participants. CSs were presented for 4.5 s. US delivery started at 4400 ms after CS onset and terminated with CS presentation. Inter-trial intervals lasted 17, 18, or 19 s (mean of 18.5 s). Trial order was randomized in such a way that not more than two trials of the same type (i.e., CS +, CS−) succeeded each other. During fear acquisition training on day 1, participants were presented with ten CS+ and ten CS− trials in context A. Five out of ten CS+ presentations (i.e., 50% reinforcement) were reinforced with an electric stimulus. Stimulus presentation was controlled by Presentation software (Version 14.8, Neurobehavioral Systems, Inc, Albany California, USA).

Electrical stimuli consisting of three square-wave pulses of 2 ms (50 ms interstimulus interval) were employed as US. The electrical stimuli were generated by a Digitimer DS7A constant current stimulator (Welwyn Garden City, Hertfordshire, UK) and delivered on the right dorsal hand through a surface electrode with platinum pin (Specialty Developments, Bexley, UK). Before the start of the experiment the intensity of the US was calibrated to a level described as painful, but still tolerable by the participant.

2.3 SCR recording and response quantification

2.3.1 SCR recording

Dataset 1 (Hamburg): 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 MP150-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). Data were recorded continuously at 1,000 Hz with a gain of 5 μΩ.

Dataset 2 (Mainz): 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 and the Biopac MP150 with EDA100C. All datasets were trimmed to 5s prior to the first event of interest (i.e., first CS presentation during acquisition training) and

22s after the last event of interest (i.e., last CS or US presentation during acquisition training). The signal was low-pass filtered with a second-order Butterworth filter with a cut-off frequency of 1 Hz using Matlab 2019a (Mathworks®, Natick, Massachusetts, USA).

2.3.2 SCR response quantification approaches employed

We applied three different response quantification approaches including their sub-categories to both datasets: was employed by two different raters for each datasets (two different rater for each dataset), one representative BLC approach (i.e., most commonly used specifications according to the literature search; (Sjouwerman et al., 2021)) as well as computational approaches as implemented in Ledalab (one representative setting) and PsPM (GLM-based as well as three different DCM-based settings). This was done for the full fear acquisition training phase for both datasets (results are presented in the main manuscript) as well as i) for the first and ii) second half of this phase separately and by using iii) the last two trials of fear acquisition training only (results are presented in the supplementary material). For Ledalab and PsPM, data used for i, ii and iii were derived from the same model as the full phase. Decision to include these additional phases was guided by the fact that the specific number of trials included in the statistical models to analyze the success of fear acquisition training is heterogeneous in the literature as revealed by the systematic literature search (T. Lonsdorf et al., 2021) and as illustrated for fear extinction (T. Lonsdorf et al., 2021; Ney et al., 2020).

Here, we do neither employ an unsupervised fully automated script based TTP approach nor include Autonomate because the supervised TTP approach offered through Autonomate's graphical user interface is reported (Green et al., 2014) to be procedurally nearly identical to the computer-assisted TTP approach employed here with identical OLW and PDWs. The choice of approaches was guided by the results of our systematic literature search described in 3.1.

2.3.2.1 Trough-to-peak (TTP)

SCRs were scored computer-assisted by using a custom-made computer program 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 an onset latency window (OLW) of 0.9 to 4s (Boucsein et al., 2012)) post stimulus onset and the peak was identified in a peak detection window (PDW) of maximally 5 s post SCR onset. In case of multiple peaks in the PDW, the first peak was considered. Raters 1 (TTP1) were experienced raters and raters 2 (TTP2) were first time raters for both datasets but different individuals for both datasets resulting in a total of 4 rater. For TTP1 in the Hamburg sample, a stimulus specific time window was used with the OLW defined for SCRs to the CS as 0.9 to 3.5s and the US as 0.9-2.5 post US

onset, as suggested recently based on an empirical evaluation of SCR onset latencies across stimulus types (Sjouwerman & Lonsdorf, 2019). This was done to have a direct empirical comparison between these recently suggested time-windows and the time windows suggested in the publication recommendations by Boucsein (Boucsein et al., 2012), which were applied for TTP2 (Hamburg) and both Mainz rater.

Both rater for the Hamburg sample were trained by the senior author and so was the experienced rater in the Mainz dataset (AMG) who then trained the first time rater in the Mainz dataset.

Data were down-sampled to 10 Hz. Each scored SCR was checked visually, and the scoring suggested by the custom-made computer program was corrected if necessary (e.g., the foot or trough when misclassified by the algorithm was manually corrected). For instance when the Data with recording artifacts (i.e., more than half of the trials) were treated as missing data points and excluded from the analyses. For the Hamburg datasets, 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 datasets was determined empirically by visually inspecting response specifically above and below this cutoff (T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019), which suggested that in this datasets, responses > 0.01 µS can be reliably identified. For the Mainz datasets a minimum amplitude criterion of 0.02 µS was used.

2.3.2.2 Baseline correction (BLC)

A custom-made script in Matlab, version R2019b implemented the BLC response quantification approach by subtracting the mean of the 2s time window prior to stimulus onset from the subsequent highest value identified in a peak detection window (PDW). The PDW spanned the minimal CS duration (6s; as CS duration was jittered between 6 and 8s) for the Hamburg sample and the full CS duration (4.5s) for the Mainz sample for both, CS and US, stimuli. In light of a substantial degree of heterogeneity in specification of the duration of the baseline time window and the PDW as revealed by the systematic literature search, these specifications were decided on because they were the most abundant ones in the literature search (n=3, see results and our related work for details on heterogeneity within the BLC approach, Sjouwerman, et al., 2021) and matched rather closely the criteria initially proposed by Pineles (BWL: -2s, PDW: full CS duration; Pineles et al., 2009) (as described in the introduction and in Table 1). Note, however, that Pineles employed an iterative algorithm in the program Mathematica for peak detection that prevents the identification of a peak despite the absence of a response (e.g., detection of the peak at the first data point in the PDW when no reaction is present but only a habituation drift). Here, however, we did not use such an iterative algorithm for the representative BC approach as no publication identified through the

systematic literature search used an iterative algorithm. A comprehensive discussion and evaluation of the different implementations of the BC approach will be discussed elsewhere (Sjouwerman, et al., 2021).

2.3.2.3 Ledalab

A continuous decomposition analysis (CDA) was conducted using Ledalab V3.4.9 (Benedek & Kaernbach, 2010a) running in Matlab 2019b (Mathworks®, Natick, Massachusetts, USA). CDA extracts a phasic information underlying the EDA signal. SCRs are deconvolved by the general response shape and are then decomposed into continuous phasic and tonic components. For data preprocessing, a second-order low-pass Butterworth filter was applied and data was down sampled to 10Hz. The 'optimize' function, as implemented in Ledalab, was used using default settings. The response window was defined as 0.9s to 4.0s after stimulus onset. Minimum thresholds of SCRs were 0.01 µS and 0.02µS for the Hamburg and the Mainz datasets, respectively. For statistics, the 'CDA.SCR' value was extracted, representing the phasic SCR activity most accurately without falling back on classic SCR amplitude, which may, however, differ from TTP ampliude (www.ledalab.de). According to the developers the CDA approach is the recommended approach in Ledalab and was, among the publications using Ledalab, also most frequently used according to our literature search.

2.3.2.4 PsPM

PsPM single-trial GLM: All Psycho-Physiological Modelling analyses (PsPM 4.3.0 (Bach et al., 2018)) were conducted in Matlab 2019b (Mathworks®, Natick, Massachusetts, USA). To capture the nature of increasing SCRs over time in the fear conditioning paradigm due to learning, single trial modelling was conducted. To estimate single-trial SCR, we employed a general linear model (Bach et al., 2009, 2013) comprising one regressor for each CS onset, one regressor for each US delivery and used a canonical skin conductance response function with time-derivative (Bach et al., 2010) and fixed response latency.

PsPM DCM fixed and flexible onset. Non-linear (dynamic causal, DCM) modelling in PsPM employs a non-linear inversion algorithm to infer single-trial estimates of sudomotor impulse response magnitude (Bach et al., 2010). Following the PsPM manual, in a first model, we applied a 'full interval' model in which the SCR onset, and its onset latency as implemented in PsPM, can be modelled within a time window that spans the entire CS duration (i.e., until US onset). In a second model we defined a time-window of 0-4 s ('restricted interval') to resemble the TTP (see 2.3.2.1) and Ledalab (see 2.3.2.3) approaches. In a third model, a fixed latency response at CS onset (i.e. DCM fixed onset) was defined. These different models were specified to elaborate on the most appropriate model and most appropriate time window in light of the PsPM manual indicating that DCM models that allow for a flexible response onset come with the risk

of absorbing SCR elicited by US and US omission and erroneously assigning it to the CS+. Thus, these models are not recommended for analyzing reinforced SCR trials which is particularly problematic for experimental designs with 100% or high reinforcement rate. More precisely, PsPM's manual states for the non-linear model, "for fear conditioning paradigms, the best way of modelling anticipatory SCR is currently under investigation. It is possibly suboptimal to model one anticipatory "flexible" response, in particular at longer CS/US SOAs when this flexible response may absorb SCR elicited by US or US omission" (cf. page 22, manual for PsPM 4.3.0, http://pspm.sourceforge.net/). In all three DCM models (i.e., fixed, full interval and restricted interval), the response latency was fixed at US onset and US omission for each trial.

2.4 Statistical analyses

All analyses were conducted in R version 4.0.2.

2.4.1 Within SCR quantification approach analyses

For all subject-specific mean stimulus SCRs as quantified by all here employed approaches, Bayesian paired two-sample t-tests as implemented in the 'BayesFactor' (https://CRAN.R-project.org/package=BayesFactor, version 0.9.12-4.2.) package (Morey & Rouder, 2018) were conducted in R to assess CS+/CS-discrimination. The package's ttestBF function was used with 1,000,000 iterations to extract the posterior of the effect size for CS discrimination for each iteration per subject. The median effect size and its 95% credible intervals (CrIs) was calculated and the Bayes factor extracted using the extractBF function. To provide complementary analyses that provide results based on most commonly employed inferential statistics to assess mean differences between CS+ and CS- (CS+/CS- discrimination), parallel analyses employed paired t-tests for all approaches using R's t.test function yielding p-values and 95% confidence intervals.

2.4.2 Evaluation of robustness of the effect against and consistency of the effect between different SCR response quantification approaches

Here, we adopted criteria for the evaluation of a set of robustness analyses from criteria suggested for the evaluation of outcomes from replication attempts (LeBel et al., 2018). The robustness analyses presented here test whether different SCR quantification approaches applied to an identical dataset yield results that justify interpreting and using the different approaches interchangeably. More precisely, we aim to empirically evaluate whether different approaches can be considered exact/very close replications or have to be considered far (or conceptual) replications in the datasets used here. Even though LeBel et al. used a NHST framework to evaluate replicability, while we use a Bayesian approach to evaluate

robustness, we consider the criteria to be generally applicable to our purposes. More precisely, we adopt the following criteria that we will apply to our data:

a) Is a signal detected within each approach? A signal is considered detected when the 95% CrI around the effect size point estimate does *not* include zero.
b) How precise is the effect size estimate within each approach? How wide are the CrI's within the different SCR quantification approaches?
c) Are the effect size estimates consistent across approaches? Consistency between two effects is considered given when the effect size point estimate of one approach is included in the other effect size's CrIs.

2.4.3 Measures of agreement across SCR quantification approaches

Most commonly the intra-class correlation coefficient (ICC) has been used in comparative research. The ICC is a 'measure of agreement, corrected for the agreement expected by chance' (cf. Bland & Altman, 1990) and is based on data that are centered and scaled using a pooled mean and standard deviation (in 'traditional', Pearson's correlation, each variable is centered and scaled by its individual mean and standard deviation). The ICC is commonly used to assess the consistency of measurements made by multiple observers (Shrout & Fleiss, 1979), in this case multiple response quantification approaches. However the use of the ICC has been criticized (Bland & Altman, 1990) and problematically, in case of systematic differences across approaches, which likely do exist here, the ICC is a composite of *intra-observer* and *inter-observer* variability (with observer here being approach) and may yield implausible results. In light of these criticisms which will not be reiterated in full detail (Shrout & Fleiss, 1979), the ICC is not considered the optimal tool for the assessment of inter-rater or inter-method agreement. Thus, we use an alternative measure that has the advantage of 'high flexibility regarding the measurement scale, the number of raters, [and] can handle missing data' (cf. Zapf et al., 2016) : the alpha coefficient suggested by Krippendorff (Krippendorff, 1970) as comprehensively described by Zapf and colleagues (Zapf et al., 2016). We use Krippendorff's α to investigate the agreement between two raters using the TTP approach a) across all trials, b) trial-by-trial and c) per CS type. Furthermore, we assess the agreement across all approaches investigated here including both TTP raters (n=8 approaches) a) across all trials, b) trial-by trial and c) per CS type. We also provide trial-by-trial pair-wise agreement between the different approaches (n=8) across all CS types and per CS-type respectively. Finally, we assessed trial-by-trial agreement between all possible pairs of quantification approaches. Krippendorff's α is a reliability coefficient with values ranging from -1 to 1, where -1 is perfect disagreement and 1 is perfect agreement.

According to Krippendorff, and α of ≥ .8 is required for agreement (Krippendorff, 2004). Benchmark values have been suggested (Landis & Koch, 1977) for interpretation of the strength of agreement as substantial (.61-.8), moderate (.41-.6) and fair (.21-.40). All analyses were conducted in R 4.0.2 using the script provided by Zapf et al. (2016) selecting ordinal measurement scaling, a two-sided type one error of 5%, and 1000 bootstrap samples.

3 Results

3.1 Systematic literature search

The systematic literature search revealed that TTP scoring (n=24) and baseline correction (BC) approaches (n=18 including two that used SCL rather than SCR but applied a baseline-correction approach) were most abundant in the publications exemplarily screened (published between 06/2018 and 02/2019) while model-based approaches (n=5) were less frequently employed (see **Figure 1A**). Of the model-based approaches, n=4 used Ledalab (n=3 CDA with varying time windows, n=1 DDA) and n=1 study used the GLM approach as implemented in SCRalyze. Within the TTP approach category, manual or computer-assisted TTP scoring are subsumed under the term "computer-assisted" and was most commonly applied (n=19) and the software Autonomate was applied in three studies (n=3) while a custom-made script was used in two (n=2) studies. Of note, it was oftentimes unclear which software program (e.g., Matlab, Acknowledge, custom-made) was used for TTP scoring and procedures were often described very rudimentary to an extent that it is possible that some studies actually used custom-made scripts rather than computer-assisted TTP scoring. Furthermore, it was often not clear if the time-window described referred to the time-window in which the onset (OLW) or the peak (PDW) had to occur. In light of the slow responding SCR, this is a crucial difference. Three studies were excluded: Two studies reported skin conductance level rather than SCR which was quantified through other means than BLC and one did record SCR but did not report methods for response quantification or SCR results as they did fail to observe differential responding (i.e. CS+>CS-) in SCRs. Thus, from the 50 publications included, 47 reported methods for SCR quantification.

Of note, these categories of approaches (TTP, BLC, model-base) were not homogeneous in themselves as across studies different criteria were applied to define a valid response, which is – at least in part – attributable to different procedural specifications (e.g., CS and ITI durations). For conciseness, we here selected one representative set of criteria for each approach (i.e., TTP, Ledalab, BC, see methods for justification for the choice of specifications for each approach) and included four different implementations offered by PsPM (see **Figure 1B**). The latter decision was based on a look into the future

for which we envision enhanced reproducibility of SCR response quantification which can be achieved optimally through model-based approaches.

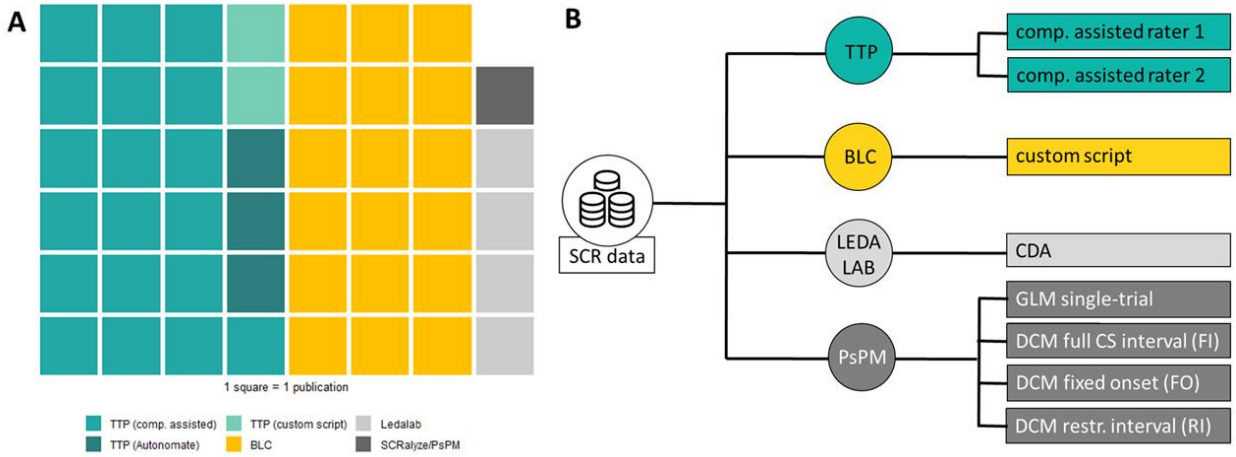


Figure 1 *(A) Frequency of different SCR quantification approaches exemplified from the systematic literature search which included 47 publications, published between 06/2018 and 02/2019. And (B) Illustration of the different SCR response quantification approaches employed to the two independent datasets in the current work: Trough-to-peak (TTP), Baseline correction (BLC), Ledalab as well as PsPM (formerly SCRalyze) with four different specifications.*

3.2 Descriptive presentation of trial-by-trial SCR trajectories and average values across SCR quantification approaches

Here, we present trial-by-trial SCR trajectories for the CS+, CS- and US during fear acquisition training as derived from the different SCR quantification approaches employed here for both datasets (see **Figure 2A and B**) as well as averaged SCR values across all trials per stimulus type (i.e., CS+, CS-, US, see **Figure 2C and D**). On a descriptive level, in both datasets (Hamburg, Mainz) the trial-by-trial trajectories appear to follow a similar pattern when responses are quantified through the TTP, Baseline correction, Ledalab approach or the single-trial GLM approach implemented in PsPM . The trial-by-trial trajectories based on the three different DCM approaches implemented in PsPM (i.e., full interval [FI], fixed onset [FO], restricted interval [RI]) deviate on a descriptive level from the trajectories derived from the above mentioned approaches. More precisely, data derived from the DCM FI approach (for both the Hamburg and Mainz datasets) and the RI approach (primarily Mainz dataset) apparently yielding larger CS+ responses but substantially smaller US responses. This was particularly pronounced in the Mainz datasets in which the CS duration was shorter than in the Hamburg study (Mainz: 4s, Hamburg: 6-8s jittered) and the reinforcement ratio was partial (50%) while it it was fully (100%) in the Hamburg study. This might be

indicative of an overestimation of CS+ responses at the cost of underestimation of US responses. This is in line with the PsPM manual noting that PsPM DCM models that allow for a flexible response onset come with the risk of absorbing SCRs elicited by US and US omission and erroneously assigning it to the CS+. Indeed in the Mainz data the DCMs with flexible response onset (FI, RI) and in the Hamburg dataset the DCM modelling the full interval (FI) seem to under-estimate US responses and instead over-estimate reinforced CS+ responses (note, order of CS+ responses differed between participants). Thus, these models do not seem suitable for analyzing reinforced SCR trials which is particularly problematic in paradigms with 100% or high reinforcement rate as all CS+ trials are reinforced. Yet, also when only analyzing unreinforced CS+ trials, this results in a reduced number of CS+ trials which necessarily impacts on the variance of the data which may turn out to be different between the CS- and the CS+ due to the different number of trials included in the analyses.

Furthermore, for the Mainz sample, the trajectories yielded by PsPM's FI model (i.e., modelling the full CS duration of 4.5s) and the restricted interval model (i.e., modelling 0-4s post CS onst) unsurprisingly result in near-identical results as the CS duration (4.5s) was only 0.5s longer than the definition of the restricted interval (i.e., 4s post CS onst). In the Hamburg dataset in which the CS duration was longer (6 - 8s jittered), however, both approaches differ substantially (i.e., full interval modelled 0-6, 0-7 or 0-8s, restricted interval: 0-4s). It is striking that in the Hamburg sample, in which the CS-US interval is much longer than in the Mainz sample, the trajectory derived from the DCM RI model (i.e., CS modelled as 0-4s post CS onset) resembled the trajectories of the TTP, BLC, Ledalab and PsPM GLM model approaches despite apparently smaller differences between the CS+ and the CS- (see also 3.3. for statistics). Yet, the US trajectory is rather comparable.

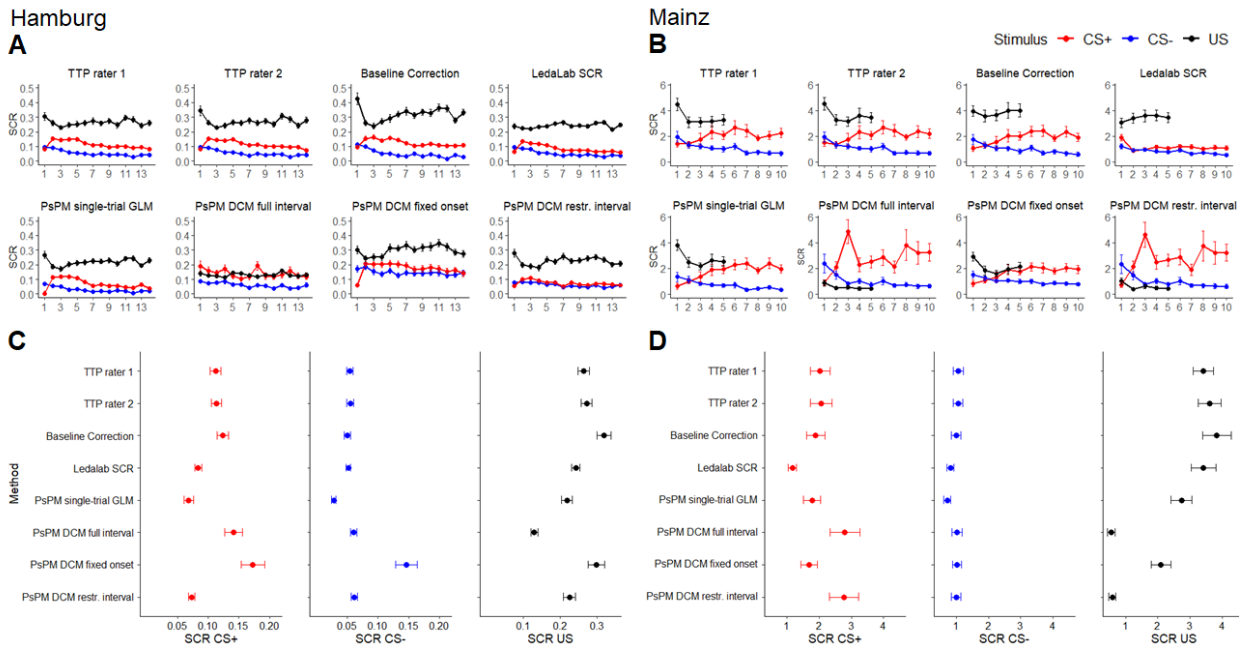


Figure 2. *Trial-by-trial trajectories for the CS+ (red), CS- (blue) and US (black) during fear acquisition training for the Hamburg* ***(A)*** *and Mainz* ***(B)*** *sample illustrated for all different SCR response quantification approaches employed: TTP rater 1 and TTP rater 2, Baseline correction (BC), Ledalab, PsPM single-trial GLM, PsPM DCM with flexible response onset in full CS-interval (FI), PsPM DCM with fixed response at CS onset (FO) and PsPM DCM with flexible response onset in a restricted interval (RI). Furthermore, the averaged raw SCRs (plus standard error) for the CS+ (red), CS- (blue) and US (black) for each SCR quantification approach employed in the Hamburg* ***(C)*** *and Mainz* ***(D)*** *datasets are shown. Supplementary Figure 1 illustrates trial-by-trial average values derived from the different quantification approaches in a single figure and Supplementary Figure 2 shows the averaged raw SCRs split up for the first and second half of fear acquisition training. Note that in the Hamburg sample a 100% reinforcement rate was employed while a 50% reinforcement rate was employed in the Mainz sample resulting in a reduced number of US responses available. As indicated in the PsPM manual PsPM DCM models that allow for a flexible response onset (here: FI and RI) come with the risk of absorbing SCRs elicited by US and US omission and erroneously assigning it to the CS+ when the CS-US interval is short.*

3.3 CS discrimination and effect sizes for the different SCR response quantification approaches

Both inferential (**Figure 3A** and **B**) and Bayesian (**Figure 3C** and **D**) paired two sample t-tests indicate significant CS discrimination during fear acquisition training for data derived from all different SCR quantification approaches employed (all p's<0.003, BF's >7.16 see **Figure 3A** and **B**), even though CS discrimination values and their CI's differed numerically between approaches. Similarly, resulting effect size estimates derived from Bayesian paired two-sample t-tests (**Figure 3C** and **D**) differed between response quantification approaches with marked variation in the Hamburg sample and lower variation between effect sizes but also wider credible intervals in the smaller Mainz sample.

It is striking that there is no clear pattern between both datasets that can be taken to identify a specific SCR response quantification approach that results in generally higher or lower effect sizes across both paradigms which differ in CS duration (4.5s vs. 6-8s), number of trials (10 vs. 14) and reinforcement rate (50% vs. 100%) as well as sample size (38 vs. 118 participants).

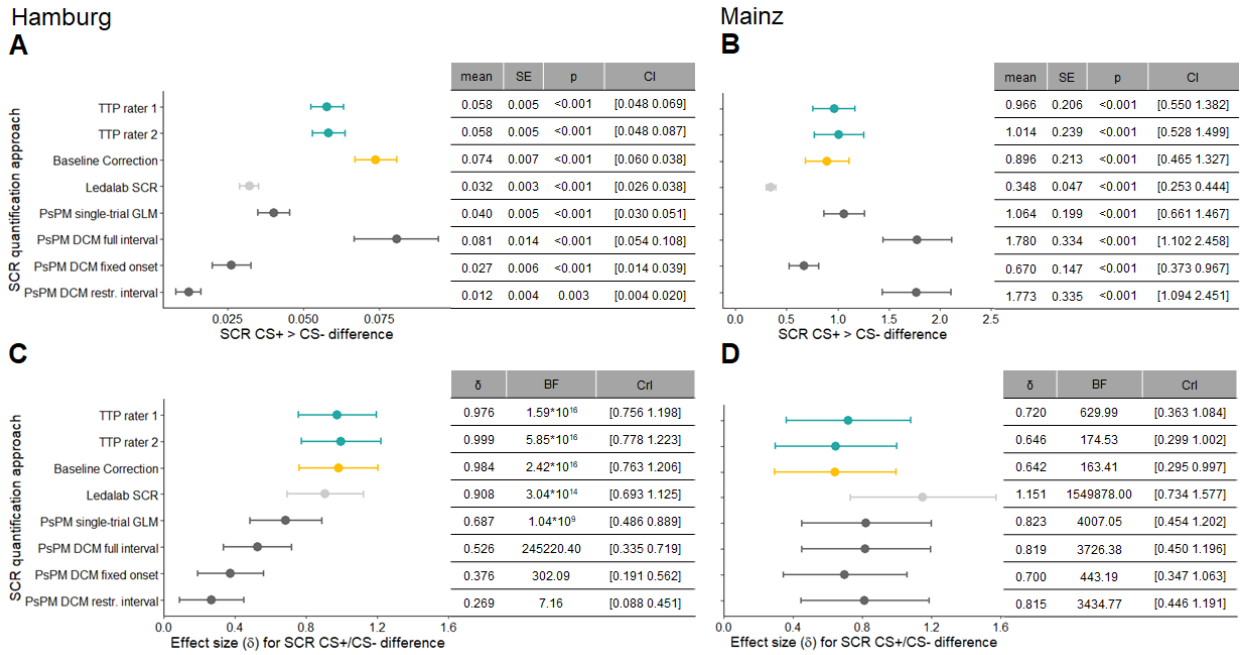


Figure 3*. CS discrimination (based on raw values per CS type during fear acquisition training) based on data derived through different SCR response quantification approaches in the Hamburg (A, C) and Mainz (B, D) datasets including results (i.e., mean, p-values, CI) from paired-sample t-tests as well as corresponding effect sizes, Bayes Factors, and credible intervals as derived from the Bayesian paired two-sample t-tests for the Hamburg (C) and Mainz (D) datasets. Supplementary Figure 3 shows this split up for the first and second half of fear acquisition training.*

Normalization (e.g. z-scoring) can naturally increase effect sizes. In our data, z-scoring does not change the general pattern of heterogeneous effect size point estimates between quantification methods (see Supplementary Figure 4).

3.4. Formal comparison of robustness of results across SCR response quantification approaches

Here we evaluate the results of the sets of robustness analyses based on three criteria borrowed from a framework suggested for the evaluation of 'replicability': (1) the existence of a signal, (2) its precision and (3) the pairwise consistency of results.

First, as described above (see 3.3.) a signal is defined here as larger SCRs to the CS+ as compared to the CS- averaged across all trials of the fear acquisition training phase. A signal is obtained for SCRs quantified from any of the eight approaches employed here in both the Hamburg and the Mainz sample.

Second, effect sizes are more precise in the larger Hamburg datasets (CrI width (min-max): 0.363 - 0.445) as compared to the smaller Mainz dataset (CrI width (max - min): 0.702 - 0.843), $t(7)=-16.12$, $p<0.001$, but are rather similar within different approaches applied to the data of one dataset.

Third, the pairwise consistency of effect sizes as indicated by the point estimate of one effect size falling within the 95% CrI of the other estimate is summarized in **Table 2**. For both the Hamburg (black) and Mainz (blue) dataset, effect sizes derived from TTP1 and TTP 2 as well as TTP1 and BLC and TTP2 and BLC were consistent with each other. For the Hamburg dataset, effect sizes derived from these three approaches (TTP1, TTP2, BLC) were consistent with those derived from Ledalab while they were inconsistent with those derived from Ledalab in the Mainz dataset with Ledalab resulting in larger effect sizes than any of the other approaches.

For the Mainz dataset all pairwise comparisons between effect sizes derived from any of the four PsPM models and the four other approaches (TTP1, TTP2, BLC, Ledalab) yielded consistent effects sizes with the exception of Ledalab yielding inconsistently larger effect sizes than the DCM fixed onset (FO), TTP1, TTP2 and BLC approaches. Yet, it has to be highlighted that the 95% CrI in the smaller Mainz dataset are wide and larger sample sizes may result in a different conclusion.

In the Hamburg dataset in turn effect sizes derived from PsPM's single trial models were inconsistent (i.e., smaller) with effect sizes derived with the aforementioned four approaches (TTP1, TTP2, BLC, Ledalab). In fact, for the Hamburg sample, effect sizes derived from any of the PsPM-based approaches, were smaller than these four approaches (TTP1, TTP2, BLC, Ledalab) and have to be evaluated as inconsistent with

these as their respective point estimates fall outside of the 95% CrI of any of these approaches. Within the different PsPM approaches, effect sizes derived from the single trial GLM model and the DCM full interval (FI) model are consistent with each other while the effect size derived from the single-trial GLM model is inconsistent with the fixed onset (FO) and restricted interval (RI) models with larger effect sizes derived from the GLM model as compared to the FO and the RI models.

The fixed onset (FO) model's effect sizes were consistent with both the full (FI) and restricted interval (RI) models' effects sizes but the effect sizes derived from the full interval (FI) model were inconsistently larger than those derived from the reduced interval (RI) model.

				PsPM			
	TTP2	BLC	Ledalab	GLM	DCM FI	DCM FO	DCM RI
TTP1	✓✓	✓✓	✓✗	✗✓	✗✓	✗✓	✗✓
TTP2		✓✓	✓✗	✗✓	✗✓	✗✓	✗✓
BLC			✓✗	✗✓	✗✓	✗✓	✗✓
Ledalab				✗✓	✗✓	✗✗	✗✓
GLM					✓✓	✗✓	✗✓
DCM FI						✓✓	✗✓
DCM FO							✓✓
DCM RI							

Table 2. *Pairwise consistency between different SCR quantification approaches with ✓ indicating consistency and ✗ indicating non-consistency for the Hamburg sample (in black: ✗ ✓) and the Mainz sample (in blue: ✗ ✓) for Trough-to-peak (TTP), baseline correction (BLC), Ledalab as well as four different models in PsPM including the trial-wise general-linear model (GLM) as well as three dynamic causal modeling (DCM) models with full interval (FI), flexible onset (FO) and restricted interval (RI).*

3.5 Agreement between different SCR response quantification approaches

Across all SCR quantification approaches, trials-wise agreement in the Hamburg sample (see **Figure 4B**) was mostly moderate to substantial but for some trials also fair. In the Mainz sample it was poor to substantial (**Figure 5B**). In the Hamburg sample, substantial agreement was observed for the CS+ trials (average [range]: 0.618 [0.533 to 0.708]) as well as for the US trials (average [range]: 0.631 [0.577 to 0.715]). Agreement for the CS- trials, however, was only moderate (average [range]: 0.500 [0.311 to 0.673]) in the Hamburg sample. In the Mainz sample in turn, substantial agreement was observed for the

CS+ (average [range]: 0.639 [0.449 to 0.769]) and the CS- (average [range]: 0.726 [0.653 to 0.805]) while agreement was only fair for the US (average [range]: 0.226 [0.165 to 0.330]).

When excluding the three PsPM DCM models which may not be optimally suited for the analyses of fear conditioning data derived from the experimental designs employed here (see above and see PsPM manual 4.3.0, page 22), agreement in the Hamburg sample remained substantial for the CS+ (average [range]: 0.778 [0.567 to 0.861]) and US (average [range]: 0.800 [0.734 to 0.845]) and remained moderate for the CS- ((average [range]: 0.578 ([0.376 to 0.720]). For the Mainz sample, agreement for the CS+ trials (average [range]: 0.808 [0.493 to 0.875]) and CS- (average [range]: 0.749 [0.619 to 0.842] also remained substantial when excluding the three PsPM DCM models while agreement for the US trials improved from fair to substantial (average [range]: 0.668 ([0.543 to 0.882]).

The trial-wise agreement between pairs of SCR quantification approaches in the Hamburg sample (**Figure 4A**) and the Mainz sample (**Figure 5A**) differed substantially with some approaches showing consistent and near perfect agreement across stimulus types (e.g., TTP1 vs. TTP2) and datasets. Yet, the pattern of pairwise agreement was often not consistent across both datasets. In the Hamburg dataset, agreement seems to be lowest for CS- trials while in the Mainz sample agreement seems to be lowest for the US trials.

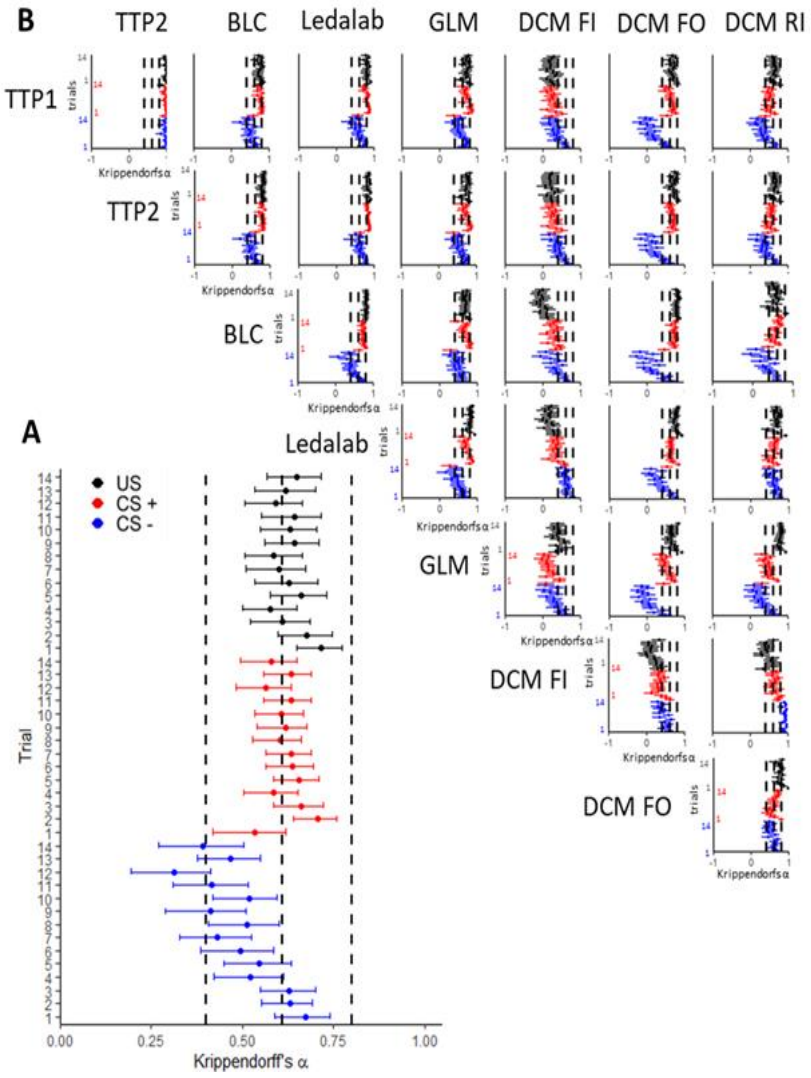


Figure 4. *Krippendorff's alpha (and CI's) as a measure of agreement between SCR quantification approaches, as calculated in the Hamburg sample (A) across all eight approaches employed for each trial during fear acquisition training. and as calculated (B) for pair-wise comparisons between the eight different approaches employed here (including the three DCM models). Different stimulus types are color coded with the CS+ in red, CS- in blue and the US in black. Vertical lines are positioned at .8 and .4 highlighting benchmarks for near perfect agreement (>.80) and fair to poor (<.41) according to the benchmarks suggested by Landis and Koch (1977). According to the benchmarks by Landis et al. (1977),values can be interpreted using the following benchmarks for Krippendorffs a <0 ``poor''agreement,0 to .2 ``slight'', .21 to .40 ``fair'', .41 to .60 ``moderate'', .61 - .80 ``substantial'', and .81 to 1 ``near perfect''.*

Note that trial sequences on the y-axis in the smaller tiles in panel B are identical to the trial sequence on the y axis in B.

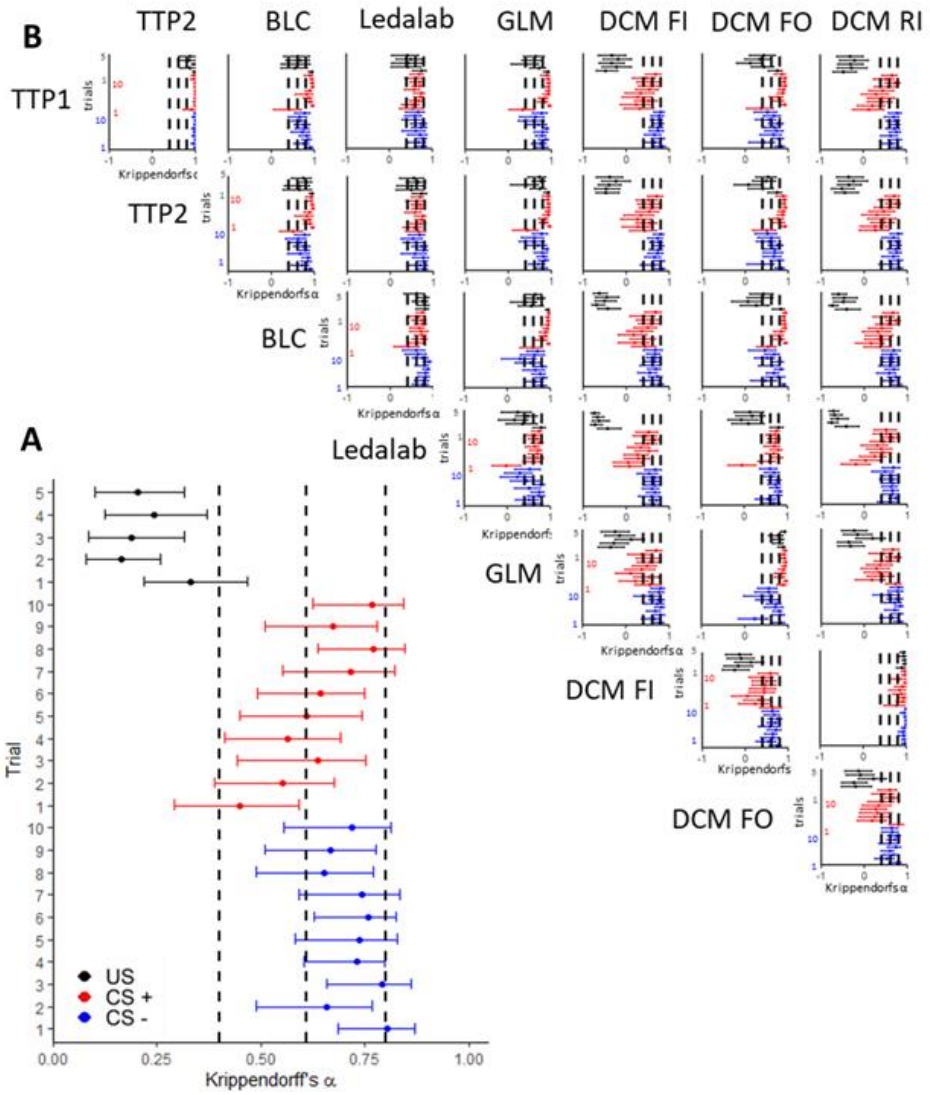


Figure 5. *Krippendorff's alpha (and CI's) as a measure of agreement between SCR quantification approaches as calculated in the Mainz sample (A) across all eight approaches employed for each trial during fear acquisition training and as calculated (B) for pair-wise comparisons between the eight different approaches employed here (including the four DCM models). Different stimulus types are color coded with the CS+ in red, CS- in blue and the US in black. Vertical lines are positioned at .8 and .4 highlighting benchmarks for near perfect agreement (>.80) and fair to poor (<.41) according to the benchmarks suggested by Landis and Koch (1977). According to the benchmarks by Landis et al. (1977), values can be interpreted using the following benchmarks for Krippendorffs a <0 ``poor'' agreement, 0 to .2 ``slight'', .21*

to .40 ``fair'', .41 to .60 ``moderate'', .61 - .80 ``substantial'', and .81 to 1 ``near perfect''. Note that trial sequences on the y-axis in the smaller tiles in panel C and D are identical to the trial sequence on the y axis in A and B.

3.6 Secondary question: Inter-rater comparisons for computer assisted TTP scoring

For both datasets (Hamburg, Mainz), two independent raters quantified SCRs though computer-assisted TTP scoring whereof rater 1 at both sites was experienced and rater 2 at both sites was a first-time rater (note that rater 1 and 2 were different individuals for both sites, that is, there were in total 4 rater). Note, however, that for Hamburg rater 1 and rater 2 used slightly different scoring criteria (i.e., 0.9-3.5 and 0.9-4.5s OLW's). Formal inter-rater reliability coefficients using Krippendorffs alpha indicate near perfect agreement across all trials and CS types (Hamburg sample: average Krippendorff's alpha [lower/upper bounds of CIs]: 0.962 [0.955, 0.969]; Mainz sample: 0.973 [0.954 0.991]). Reliability coefficients calculated separately for the stimulus types also revealed near perfect agreement for the CS+ (Hamburg sample: 0.961 [0.948, 0.974]; Mainz sample: 0.990 [0.977,0.998]), the CS- (Hamburg sample: 0.948 [0.934, 0.962]; Mainz sample: 0.992 [0.984,0.997]), and the US (Hamburg sample: 0.961 [0.946, 0.975]; Mainz sample: 0.919 [0.823, 0.986]).

Finally, the range of trial-wise agreement (see Supplementary Table 1) revealed near perfect agreement across trials for the Hamburg sample [(0.845 to 0.996) and the Mainz sample alike (0.860 to 1).

Figure 6 illustrates the excellent inter-rater reliability on a CS-type level (i.e., averaged SCR magnitude per stimulus type for rater 1 and rater 2) per individual. Note that the figure illustrates this descriptively on an individual level (i.e., connects the average SCR magnitude value as scored by rater 1 and rater 2 for data from the same participant, while the analyses described above (i.e., Krippendorff's alpha) do not include the individual subject level.

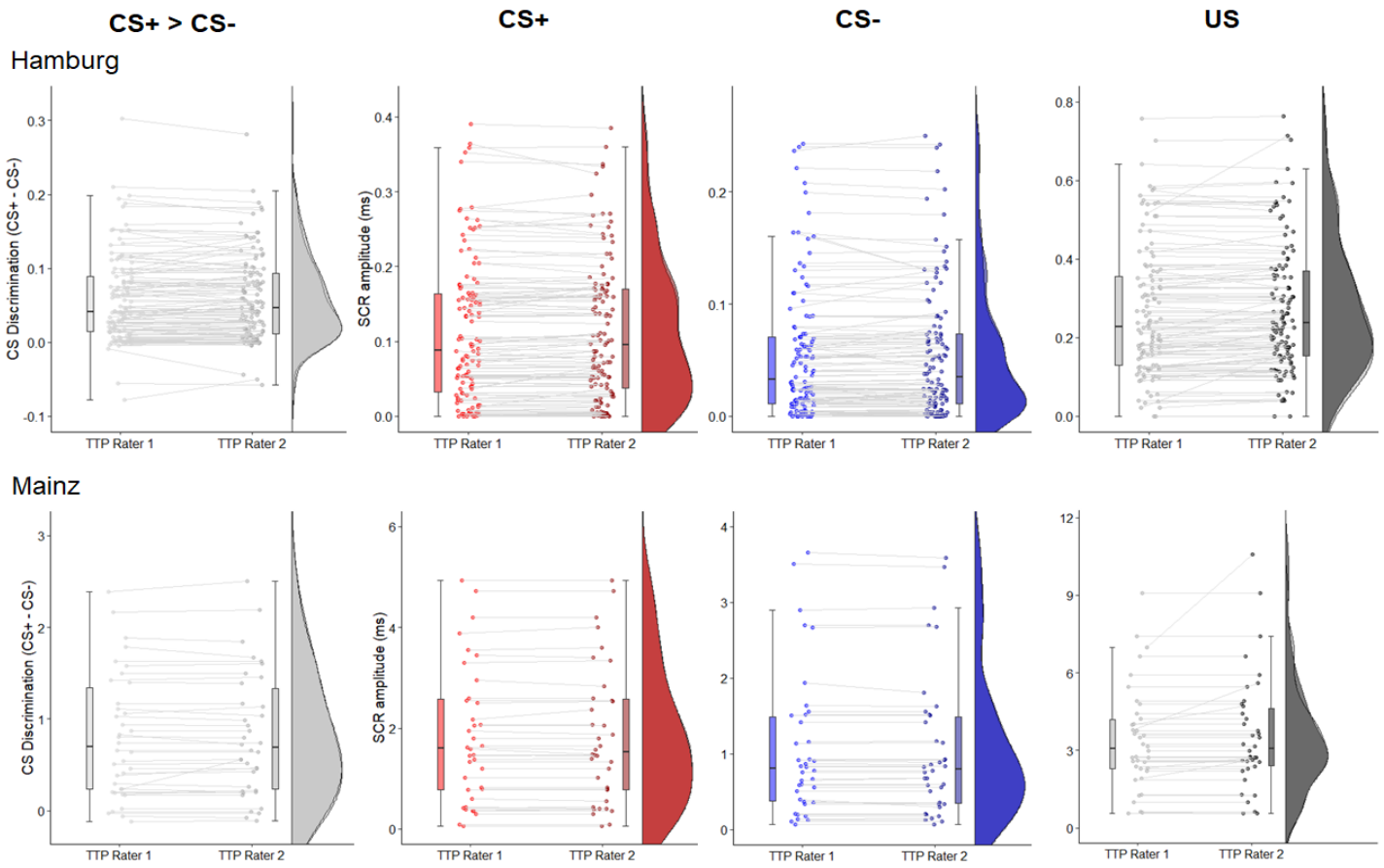


Figure 6. *Inter-rater comparisons between TTP rater 1 and TTP rater 2 for the Hamburg sample (upper row) and the Mainz sample (lower row) for single trial discrimination (light grey) as well as single trial SCRs for the CS+ (red), CS- (blue) and the US (dark grey) during fear acquisition training. Subplots show single trial or pair-wise discrimination values as well as box plots and densities for both raters with identical trials connected through lines. Note that densities are nearly completely overlapping. Note that rater 1 and 2 were different individuals in the Hamburg and Mainz sample. Also note that both rater used exactly the same criteria in the Mainz sample while in the Hamburg sample both rater used slightly different criteria to allow for a direct comparison of two previously suggested sets of criteria (see Methods for details).*

4 Discussion

Here, we provide a comparison between seven different SCR response quantification approaches across two datasets characterized by different procedural details. The overarching aim of this work was to a) provide a synopsis of which approaches are used in the literature by using fear conditioning research as a case example and b) evaluate if and to what extent the different approaches used in the literature lead to comparable results as well as c) investigate the inter-rater agreement between two individuals performing TTP scoring in two datasets.

4.1 Take home message from the systematic literature search

The systematic literature search revealed three main take-home messages: **First**, (computer-assisted) TTP scoring and BLC through custom-made scripts seem to be the prevailing approaches for SCR response quantification in fear conditioning research to date. Our literature search, however, covers only articles published in a six-month period until early 2019 and we anticipate that the model-based approaches may become increasingly attractive with increasing appreciation of the value and importance of computational reproducibility.

Second, the SCR quantification approaches identified (i.e., TTP, BLC, Ledlab, PsPM) do not represent unitary methods but come in heterogeneous specifications (see for instance **Table 1**). This likely originates - at least partly - from differences in experimental paradigms, particularly timing and duration of stimulus presentation. This, however, is unlikely to be obvious for novices or researchers outside the field and we thus recommend to explicitly and clearly justify specific choices for response quantification criteria including appropriate references. More precisely, TTP and BLC approaches differ in the definition of onset latency, baseline and peak detection time window and a comprehensive overview has been provided by Pineles et al. (2009). Similarly, a number of different settings and approaches are offered by software programs that implement model-based approaches such as Ledalab (http://www.ledalab.de/documentation.htm) and PsPM (e.g., GLM-based, DCM-based with different possible settings each, http://pspm.sourceforge.net/documentation/). The specific model, the chosen settings and, if applicable, the selected output measure (e.g., parameter estimate, reconstructed response, area under the curve, etc.) need to be reported in enough detail to allow for computational reproducibility, which is often not the case as revealed by our literature search. We refer to our related work (Sjouwerman et al., 2021) for an investigation of within-approach heterogeneity with a focus on the BLC method.

Third, we noticed that navigating among the different SCR response quantification approaches and terminology employed in the literature can be rather challenging even for researchers familiar with the field. For instance, TTP scoring has sometimes been referred to as (standard) 'peak scoring', a term which has also been used to subsume TTP and BLC approaches (Privratsky et al., 2020). This distinction is, however, important as the *onset latency* window (OLW) for TTP scoring cannot be employed as a *peak detection* window (PDW) in BLC approaches (as done in Privratsky et al., 2020) simply as the *onset* of a stimulus included SCR (i.e., OLW) occurs with a different timing from CS onset as the *peak* (i.e., PDW) and hence the peak may be missed. This is rather likely when employing windows as short as 0-3s (Privratsky et al., 2020) taken from the OLW as PDW. To avoid this jingle (i.e., assuming erroneously that two different things are the same because they bear the same name)-jangle (i.e., two identical things are erroneously considered to be different because they carry different names) fallacy, we suggest using standard terminology and to describe methods and procedures as precisely and transparently as possible. This includes ensuring that references used really refer to the procedure employed in all details, which was not always true for the publications included in the systematic literature search. It was most striking that many publications employing the BLC approaches oftentimes cited the study by Pineles (2009) as a reference, which, however, used an iterative algorithm and often different time windows than the citing literature. The articles identified through the literature search, however, were exclusively based on custom-made scripts that did not seem to include an iterative algorithm but were also not shared with the articles.

4.2 Comparison between different approaches

Here, we applied seven different SCR response quantification approaches to two independent datasets in a small-scale data multiverse approach: computer-assisted TTP scoring, a representative BLC approach, CDA as implemented in the software Ledalab as well as four different models offered by the software PsPM (GLM single trial, DCM full interval, DCM fixed onset and DCM restricted interval). Furthermore, two independent raters performed TTP scoring in both datasets - whereof one first-time rater and one experienced rater to allow for the assessment of inter-rater reliability.

4.2.1 (Computational) reproducibility and concordance between TTP raters

From a computational reproducibility perspective (i.e., obtaining the same results when applying the same methods to the same data), fully unsupervised and fully automatized procedures offer practical and methodological advantages and are available for the TTP approach (i.e., "Autonomate"; Green et al., 2014), inherent in the model-based computational approaches (e.g., PsPM, Ledalab) and implemented in

the script-based BLC approaches. Yet, reproducibility is limited as particularly the custom-made scripts were not publicly available (but might be available on request which we did not attempt). Computer-assisted or manual TTP scoring approaches in turn require extensive training prior to performing the scoring, are never completely free from scorer bias and human errors and require substantial time investments when a large number of trials and/or a large number of participants are included. From a reproducibility perspective, however, within-lab inter-rater concordance rates reported here are near perfect for both datasets even with a slight change in employed criteria (i.e., TTP1 and TTP2 in the Hamburg sample) and one rater being experienced while one was a first-time rater. This matches high concordance rates as reported in previous reports (average ICC: .982; Green et al., 2014) and together suggests that reliability and reproducibility may not be a major concern for computer assisted TTP scoring - provided raters are well trained. Our results are reassuring and echo previous findings that suggest that reliability of TTP scoring is excellent.

4.2.2 Robustness of the CS discrimination effect against different response quantification approaches

The application of different SCR quantification approaches to the same datasets can be viewed as a set of robustness analyses (i.e., applying different processing or analysis pipelines to the same data) with the overarching aim to investigate if and to what extent the different methods lead to comparable results. As we are not aware of a formal framework for the evaluation of the outcome of robustness analyses, we here borrowed some criteria from a framework suggested for the evaluation of 'replicability' in general (LeBel et al., 2018). More precisely, we evaluated whether there was a) a signal. This is in the context of this work defined as significant CS discrimination. We furthermore evaluated b) whether the effect size of this signal was consistent across the different approaches, and whether c) the (relative) precision of the effect differed across the different SCR response quantification approaches.

In sum, a **signal** (i.e., significant CS discrimination) was universally observed in both datasets irrespective of quantification approach. As we focused on the average responding during the full fear acquisition training phase in which strong CS discrimination is typically observed, it cannot be excluded that a focus on a more subtle effect in different experimental phases such as a return of fear test or recall phase may lead to different results across SCR quantification approaches. This would be important to address in future work.

Furthermore, the **precision** of the resulting estimates did not differ significantly between different SCR response quantification approaches applied within the datasets.

Yet, the effect sizes yielded by the different approaches were not universally **consistent**: In the Hamburg sample (N=118, 100% reinforcement rate, CS duration: 6-8s), both TTP raters (TTP1 and TTP2), the BLC approach as well as the CDA approach implemented in Ledalab yielded consistent effect sizes while effect sizes generated though any of the PsPM models were smaller and inconsistent with all of the aforementioned approaches. In addition, the four PsPM models did not yield consistent effect sizes either when compared to each other in the Hamburg dataset. In the smaller Mainz sample (N=38, 50% reinforcement rate, CS duration: 4.5s), however, most approaches yielded consistent effect sizes even though it has to be noted that the CrI's were wider as in the larger Hamburg sample. Still, the CDA approach as implemented in Ledalab yielded an effect size that was inconsistent with and larger than those yielded by TTP1, TTP2, BLC as well as one of the PsPM models (i.e., DCM FO).

4.2.3. Comparable results yielded by the TTP and representative BLC approach

From this pattern of (in)consistency, we conclude that in the two datasets investigated here, only few SCR response quantification approaches yielded comparable effect sizes in both datasets, despite numeric difference between the CS+ and the CS- (CS discrimination): TTP and the representative BLC approach employed as well as some of the PsPM models (i.e., GLM and DCM FI; DCM FI and DCM FO as well as DCM FO and DCM RI).

With respect to the TTP and BLC approach, the time-window during which the peak SCR was to be identified were relatively similar in TTP (i.e., up to 5s post CS onset) and BLC (i.e., full CS duration which corresponds to 0-6s in the Hamburg and 0-4.5s post CS onset in the Mainz sample). The trough of the response, however, is defined very differently (i.e., BLC: average SCL 2s prior to CS onset; TTP: onset in a OLW of 0.9-4.5s post CS onset). This group-level comparability between both approaches is both striking and surprising given the prominent differences between both approaches. For instance, the BLC approach can yield negative values as the highest value in the PDW may be lower than the average baseline when there is a strong habituation drift in the data while such negative values are implausible in TTP scoring. Furthermore, as the BLC approach was employed in a script-based manner without visual inspection and without the implementation of adaptive algorithms (as in Pineles et al., 2009), a value for a response is always identified while the TTP approach may score both missing (e.g., electrode artifacts) and zero responses. The latter is for instance the case, when there is only a habituation trend but no response, which would correspond to a negative value in the BLC approach. We refer to our related work using a full multiverse approach covering 150 combinations of time-windows used in the BLC approach for an in-

depth discussion about the differences between TTP and BLC approach and the resulting problems (Sjouwerman et al., 2021).

Despite a number of major problems with the BLC approach discussed in our related work (Sjouwerman et al., 2021), our results are re-assuring that TTP and the representative BLC approach to SCR response quantification seem to yield comparable results - at least for the design specifications included here and average responding at the group level. As these are the currently two most abundantly used approaches to SCR response quantification in the field of fear conditioning research, this is good news for the field.

4.2.4. Comparison of model-based approaches implemented in PsPM across datasets

Furthermore, it is noteworthy that the four PsPM models yielded more consistent results not only in comparison to each other but also with any of the other approaches in the Mainz than the Hamburg datasets. We can only speculate on potential reasons beyond the generally wider CrI in the smaller Mainz sample. For instance, the stimulus durations in the studies included in previous PsPM comparative work (Bach, 2014; Bach et al., 2010, 2013) was with 1s to 3.5s rather short. The CS duration of 4.5s in the Mainz dataset is closer to this than the 6-8s duration in the Hamburg dataset. It remains to be investigated systematically whether the model-based approaches in PsPM are optimized for shorter duration CSs, and short it is or work equally well with longer duration stimuli that are more common in fear conditioning research. In addition, reinforced CS+ trials were excluded in the studies validating PsPM in fear conditioning data and also in the only study included in our systematic literature search that used PsPM's GLM model (Taylor et al., 2018). We did not exclude reinforced trials in the Mainz sample and this was impossible to do for the Hamburg sample as all CS+ trials were followed by the US - in fact this may be a major reason why the PsPM models were inconsistent with any other models in the Hamburg data. Of note, two of the here employed DCM approaches seemed to erroneously assign SCRs elicited by the US to the CS in both samples. Thus, the DCM approaches may not be optimal for response quantification in paradigms with full or high reinforcement rates or when not excluding reinforced trials (see PsPM manual 4.3.0, page 22). Of note, excluding reinforced trials as modelling a flexible CS response onset may absorb SCR elicited by US or US omission leads to unequal number of trials for the CS+ and the CS-. These unequal numbers of trials resulting from excluding reinforced trials may result in different variances, reliability estimates and statistical power which may also be problematic. Another difference between previous comparative work focusing on SCRalyze/PsPM is that these previous studies included a (substantially) higher number of trials per condition (i.e., 16-90 trials) as our work (i.e., 10-15) which may result in differences in statistical power and different impact of the fast habituation typically seen in skin conductance responding.

In sum, the software package PsPM offers a number of different model specifications that - likely depending on experimental specifications - can substantially impact on the results. Thus, data processing and model specification need to be reported in detail to ensure computational reproducibility and that the models need to be empirically evaluated against typical paradigm specification details such as reinforcement rate and stimulus duration (see e.g., Bach et al., 2010).

4.3 Implications for post-processing and data analyses

Here, we have illustrated that different commonly used SCR response quantification approaches used in fear conditioning research do not necessarily yield converging and comparable effect sizes for group-level CS-discrimination despite all yielding significant CS+/Cs- discrimination in the same direction. The different effect sizes and different numeric values for CS+ and CS- responses as well as CS+/CS- discriminination may also have implications for the application of commonly used post-processing or data-cleaning procedures such as minimal response criteria as well as the identification of performance-based exclusion of SCR non-responder and SCR non-learner (for a critical evaluation and discussion see T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019). For instance, responses quantified through the TTP approach cannot be smaller than zero while the BLC approach can yield negative values. Further, it is clear from the average CS+, CS- and CS disrimination values (see **Figure 2**) yielded by the different response quantification approaches, that identical cut-offs for non-learning are likely to lead to different results across approaches. Yet, we did not investigate this empirically and hence can only speculate here.

4.4 Is there a single best approach for SCR response quantification?

The common aim of all SCR response quantification approaches is to infer an unobservable psychological process (sympathetic arousal) from observable skin conductance data. It has been proposed that we may identify the 'best' approach for SCR response quantification by means of 'retrodictive validity', formerly referred to as 'predictive validity' (Bach et al., 2020; Bach & Melinscak, 2020). The method with the highest retrodictive validity is the method that has the highest chance of recovering this unobservable process. It has further been suggested that this can be achieved by comparing two conditions that are known to induce strong differences in sympathetic arousal (Bach, 2014) such as viewing of aversive (strong arousal) and neutral (weak arousal) pictures or a condition predictive of an aversive event (i.e., CS+) and a control condition (i.e., CS-). According to retrodictive validity, the best method would be the method that best separates both conditions. In the context of this work, the method that produces the strongest CS discrimination.

When interpreting the results of our work in a 'retrodictive validity framework', we do not observe a single superior approach. More precisely, our results from two different datasets differing primarily in number of participants (118 vs. 38), reinforcement rate (100% vs. 50%) and CS duration (6-8s vs. 4.5s) suggest that there may not be a single universally superior method as there is no single method that yields a consistently higher effect size as compared to other methods in both datasets.

Rather than suggesting a single universally superior approach, we echo the notion, that assumptions about the shape and timing of an SCR across different quantification approaches are mostly similar, but that "they are implemented using different algorithms which may impact their performance and comparability across different paradigms or experimental contexts" (cf. Green et al., 2014, p. 192). Consequently, a single 'best' or 'superior' method may not exist as the most suitable method most likely depends on design and sample specifics. This is a complicated scenario that does not allow for an easy solution. As a consequence, we call for caution in light of the recent suggestion (e.g., Bach & Melinscak, 2020; Privratsky et al., 2020) that PsPM-based SCR quantification *generally* leads to a massive reduction in required participants as opposed to other approaches due to substantially higher statistical power and retrodictive validity (as also discussed in Bach & Melinscak, 2020). More precisely, our data suggest that depending on the design and sample specifications, sometimes the opposite may be true: for instance, we observed smaller effect sizes for CS discrimination (i.e., retrodictive validity) for all PsPM-based approaches as opposed to the TTP, BLC and Ledalab-based SCR response quantification in the Hamburg sample. Given that the evidence to date is limited, we echo the call (Bach & Melinscak, 2020) for more comparative (multiverse-type of) studies and thorough validation of new methods in different experimental and design settings until a single method can be recommended, in particular recommended as superior.

4.5 Limitations

Here, we compare seven different SCR response quantification approaches as identified through a literature review. Yet, the full multiverse for SCR processing steps that may potentially impact on the results include a number of additional steps not considered here in depth such as transformations (see also supplementary material) , cut-off criteria (T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019), data exclusion (T. B. Lonsdorf, Klingelhöfer-Jens, et al., 2019), and filtering (see e.g., Privratsky et al., 2020)). Furthermore, also different specifications in the experimental design may have an impact – for instance shorter ITI durations or even shorter CS durations that may lead to an increase in overlapping SCRs which model-based approaches may be particularly suitable for when appropriately designed. Future work may

systematically focus on these additional steps or cover the full data multiverse systematically (see Sjouwerman et al., 2021 for a full multiverse focusing on within approach heterogeneiety in the BLC method).

SCRs were relatively larger in the Mainz as compared to the Hamburg sample. This difference may be explained by the usage of a more aversive US in the Mainz sample: US intensity was calibrated to a level perceived as 'maximally painful, but still tolerable' as compared to 'maximally uncomfortable, but not painful' in the Hamburg sample. Empirical and theoretical work suggests that stronger US intensity is associated with larger conditioned responses (Morris & Bouton, 2006; Rescorla & Wagner, 1972). The difference could also be explained by the different reinforcement rates employed in both datasets as SCRs have been suggested to reflect the associability of a stimulus (Li et al., 2011; Seymour et al., 2005; Tzovara et al., 2018; Zhang et al., 2016). Finally, differences in external conditions, such as room temperature, differences in hardware could also account for these differences.

Finally, our comparison of different SCR response quantification approaches across two datasets focused on average group-level responding and future work focusing on individual-level responding would be a logical extension of our and previous work.

4.6. Summary and Outlook

Our results illustrate heterogeneity in the exact specification and implementation of SCR response quantification approaches and illustrate partly inconsistent outcomes for CS discrimination when applying seven different SCR response quantification approaches to the same data. Our results challenge the existence of a universally 'best' or 'superior' SCR response quantification approach across design and sample specifications and call for more and systematic comparative (multiverse-type of) studies.

5 References

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

Bach, D. R., Castegnetti, G., Korn, C. W., Gerster, S., Melinscak, F., & Moser, T. (2018). Psychophysiological modeling: Current state and future directions. *Psychophysiology*, *55*(11), e13214. https://doi.org/10.1111/psyp.13209

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

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

Bach, D. R., & Friston, K. J. (2013). Model-based analysis of skin conductance responses: Towards causal models in psychophysiology. *Psychophysiology*, *50*(1), 15–22. https://doi.org/10.1111/j.1469-8986.2012.01483.x

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

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

Benedek, M., & Kaernbach, C. (2010a). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*(1–5), 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028

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

Bland, J. M., & Altman, D. G. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in Biology and Medicine*, *20*(5), 337–340. https://doi.org/10.1016/0010-4825(90)90013-F

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

Dawson, M. E., Schell, A. M., Filion, D. L., & Berntson, G. G. (2007). The Electrodermal System. In *Handbook of Psychophysiology* (Third edition). Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511546396.007

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), 2515245920954925. https://doi.org/10.1177/2515245920954925

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Garrett-Ruffin, S., Hindash, A. C., Kaczkurkin, A. N., Mears, R. P., Morales, S., Paul, K., Pavlov, Y. G., & Keil, A. (2021). Open science in psychophysiology: An overview of challenges and emerging solutions. *International Journal of Psychophysiology*, *162*, 69–78. https://doi.org/10.1016/j.ijpsycho.2021.02.005

Gerlicher, A. M. V., 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

Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., & Citi, L. (2016). cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing. *IEEE Transactions on Biomedical Engineering*, *63*(4), 797–804. https://doi.org/10.1109/TBME.2015.2474131

Green, S. R., Kragel, P. A., Fecteau, M. E., & LaBar, K. S. (2014). Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *91*(3), 186–193. https://doi.org/10.1016/j.ijpsycho.2013.10.015

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70. https://doi.org/10.1177/001316447003000105

Krippendorff, K. (2004). Reliability in Content Analysis. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. JSTOR. https://doi.org/10.2307/2529310

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*. https://doi.org/10.1177/2515245918787489

Levinson, D. F., & Edelberg, R. (1985). Scoring criteria for response latency and habituation in electrodermal research: A critique. *Psychophysiology*, *22*(4), 417–426. https://doi.org/10.1111/j.1469-8986.1985.tb01626.x

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

Lim, C. L., Rennie, C., Barry, R. J., Bahramali, H., Lazzaro, I., Manor, B., & Gordon, E. (1997). Decomposing skin conductance into tonic and phasic components. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *25*(2), 97–109. https://doi.org/10.1016/s0167-8760(96)00713-1

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

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., Shiban, 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., & 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

Lonsdorf, T., Gerlicher, A., Klingelhöfer-Jens, M., & Krypotos, A.-M. (2021). *Multiverse analyses in fear conditioning research*. PsyArXiv. https://doi.org/10.31234/osf.io/2z6pd

Lykken D & Venables P. (1971). Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology*, *8*, 656–672. https://doi.org/10.1111/j.1469-8986.1971.tb00501.x

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & 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

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), 371–385. https://doi.org/10.1037/0097-7403.32.4.371

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), e13650. https://doi.org/10.1111/psyp.13650

Ojala, K., & Bach, D. R. (2019). *Measuring learning in human classical threat conditioning: A review of translational, cognitive and methodological considerations* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/2dzkj

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

Privratsky, A. A., Bush, K. A., Bach, D. R., Hahn, E. M., & Cisler, J. M. (2020). Filtering and model-based analysis independently improve skin-conductance response measures in the fMRI environment: Validation in a sample of women with PTSD. *International Journal of Psychophysiology*, *158*, 86–95. https://doi.org/10.1016/j.ijpsycho.2020.09.015

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). 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., & Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, *8*(9), 1234–1240. https://doi.org/10.1038/nn1527

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

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

Sjouwerman, R., Illius, S., Kuhn, M., & Lonsdorf, T. (2021). *A data multiverse analysis investigating non-model based SCR quantification approaches*. PsyArXiv. https://doi.org/10.31234/osf.io/q24t8

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

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

Steegen, 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

Taylor, V. A., Roy, M., Chang, L., Gill, L.-N., Mueller, C., & Rainville, P. (2018). Reduced Fear-Conditioned Pain Modulation in Experienced Meditators: A Preliminary Study. *Psychosomatic Medicine*, *80*(9), 799–806. https://doi.org/10.1097/PSY.0000000000000634

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

Venables, P.H. & Christie, M.J. (1980). Electrodermal activity. In Martin, I & Venables, P.H. (Eds.), *Thechniques in psychophysiology*. Wiley.

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16*(1), 93. https://doi.org/10.1186/s12874-016-0200-9

Zhang, S., Mano, H., Ganesh, G., Robbins, T., & Seymour, B. (2016). Dissociable Learning Processes Underlie Human Pain Conditioning. *Current Biology: CB*, *26*(1), 52–58. https://doi.org/10.1016/j.cub.2015.10.066

6 Author notes

Grant funding: This work was funded by grants from the German Research Foundation (DFG) to TBL (DFG LO1980/7-1 and CRC58; subproject B07, INST 211/633-1).

Acknowledgements: The authors thank Karita Ojala for helpful comments on the manuscript. The authors thank Claudia Immisch for data acquisition (Hamburg sample), Karoline Rosenkranz for TTP scoring of the Hamburg sample and Maren Klingelhöfer-Jens for TTP scoring of the Hamburg sample as well as for data processing and preparation. We thank Raffael Kalisch and Oliver Tüscher for funding acquisition of the Mainz data (CRC1193; subproject C01 to RK and subproject C04 to OT) and their support with data collection. Furthermore, we thank Anita Schick, Merve Ilhan, Julian Behr, Petra Seyfahrt, Kenneth Yuen und Amgad Droby for support with data collection in Mainz, and Danielle Stibbe for TTP scoring of the Mainz sample.

We also thank Matthias Gamer for providing EDAview for computer-assisted TTP scoring.

Conflict of interest statement: The authors do not report any conflict of interest.

Name and e-mail address for reprints: Tina B. Lonsdorf, t.lonsdorf@uke.de

The authors made the following contributions (CRediT-Statement):

MK: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft

AMVG: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft

TBL: conceptualization, funding acquisition, methodology, project administration, resources, supervision, visualization, writing – original draft

Supplementary Material for

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 models

short title: SCR quantification approaches

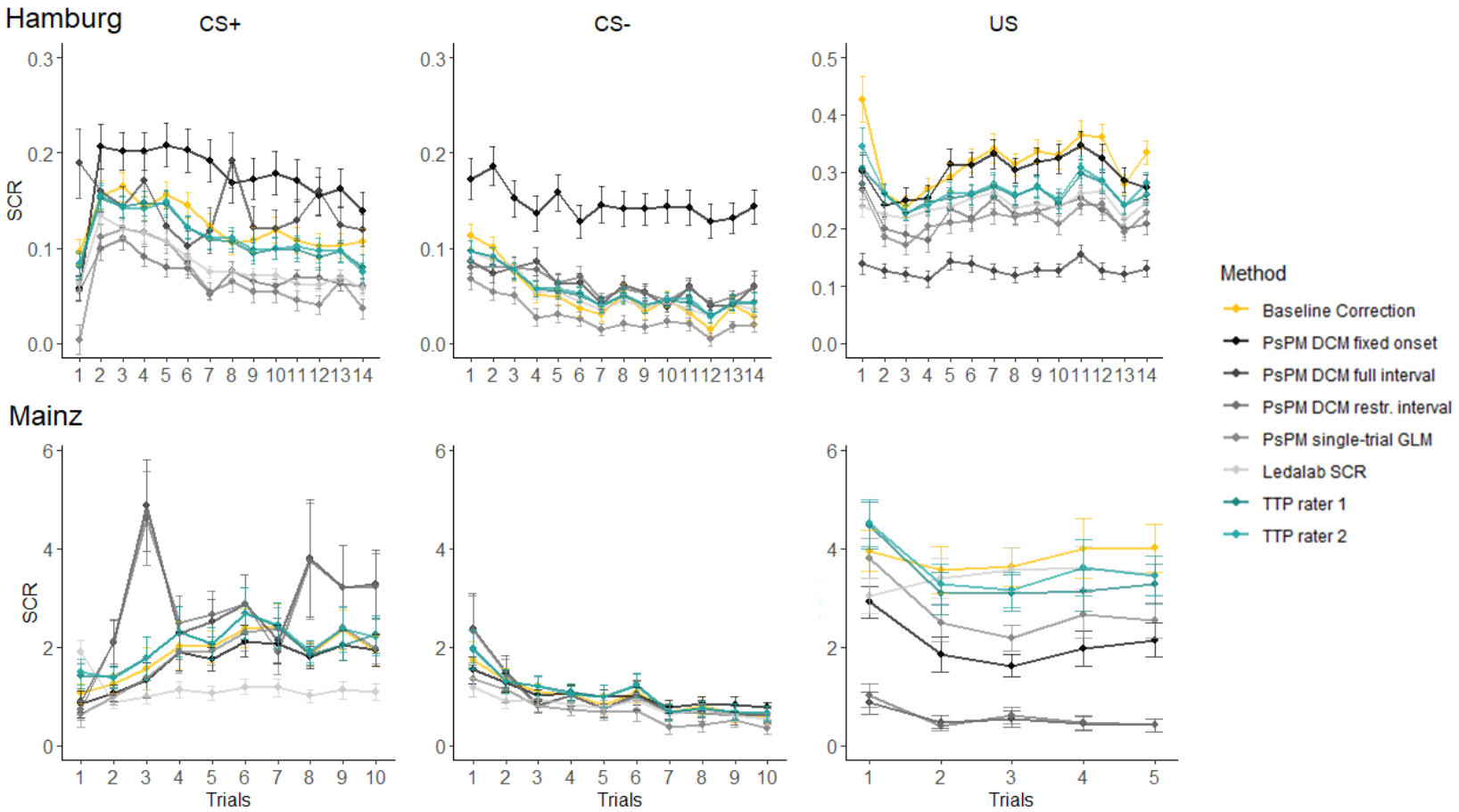
Manuel Kuhn[1,2], Anna M.V. Gerlicher [3] & Tina B. Lonsdorf[1]

[1] Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

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

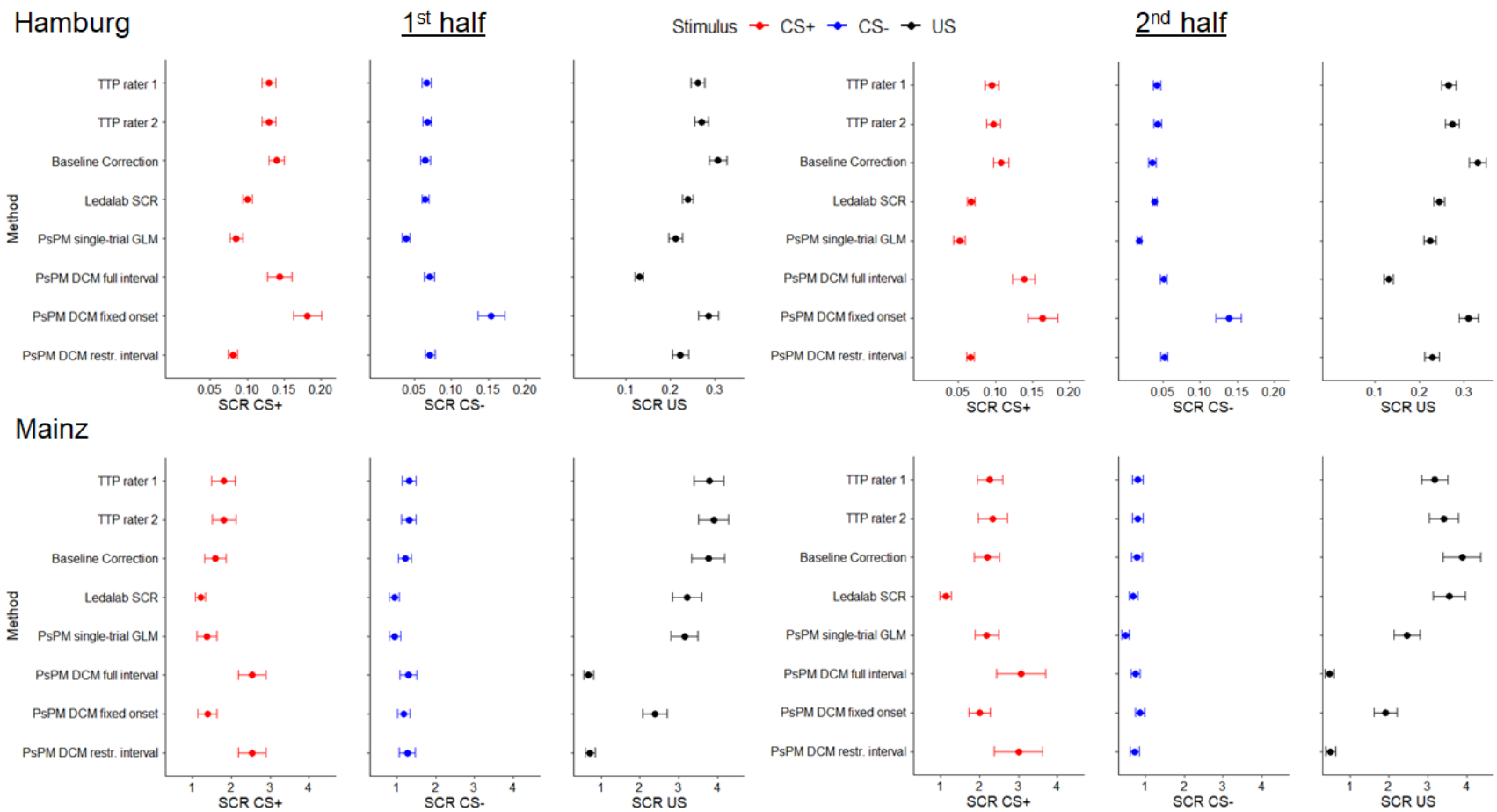
[3] Department of Clinical Psychology, University of Amsterdam, Amsterdam, The Netherlands

Corresponding author: Tina B. Lonsodorf, Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Martinistrasse 52, Hamburg, Germany +49 40 7410 55769, t.lonsdorf@uke.de

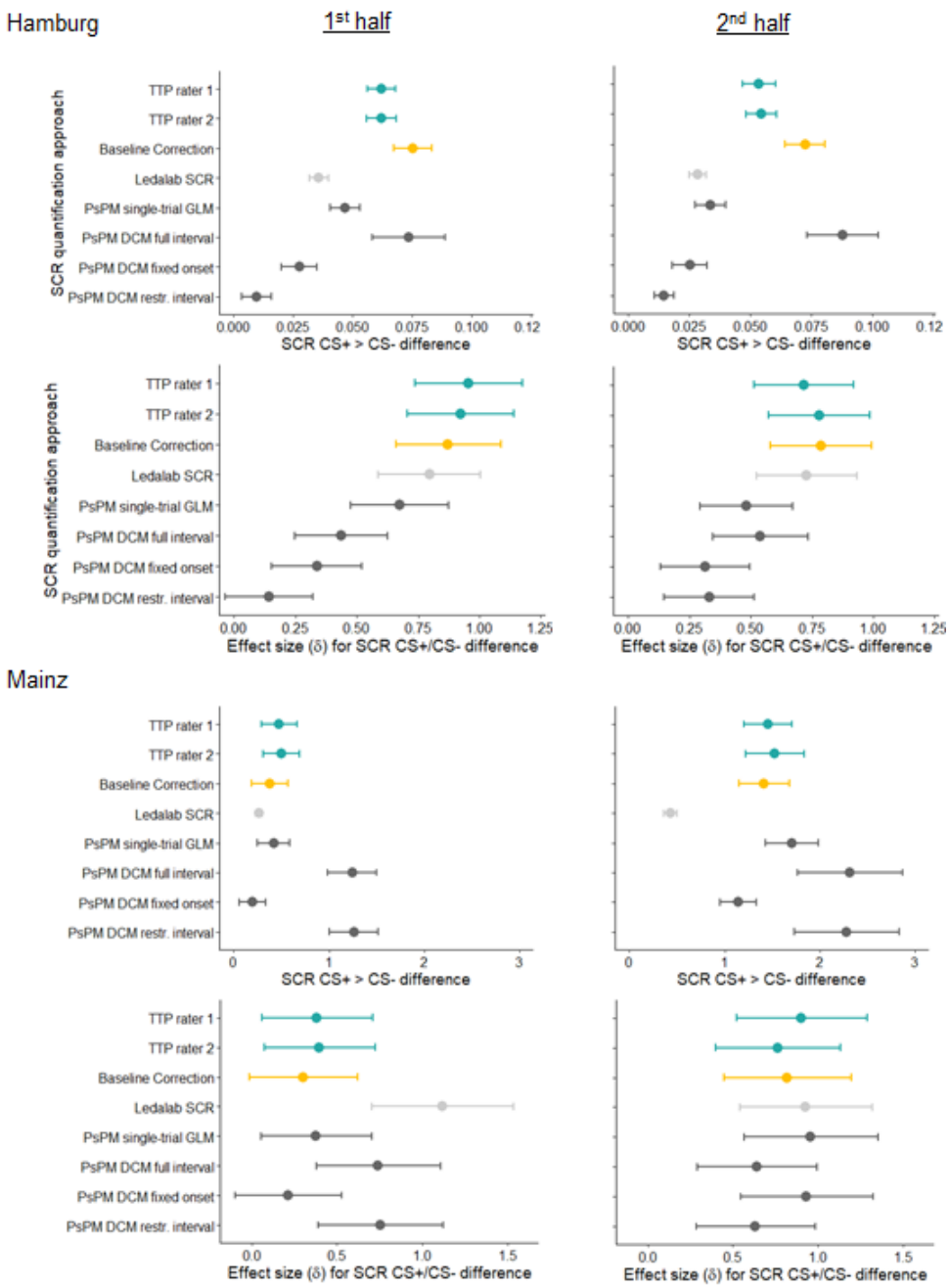


Supplementary Figure 1. *Trial-by-trial averages values (averaged across participants) for the CS+ (left), CS- (middle) and the US (right) in the Hamburg sample (upper row) and the Mainz sample (bottom row).*

Note that the scales used by the different approaches are not necessarily directly comparable.



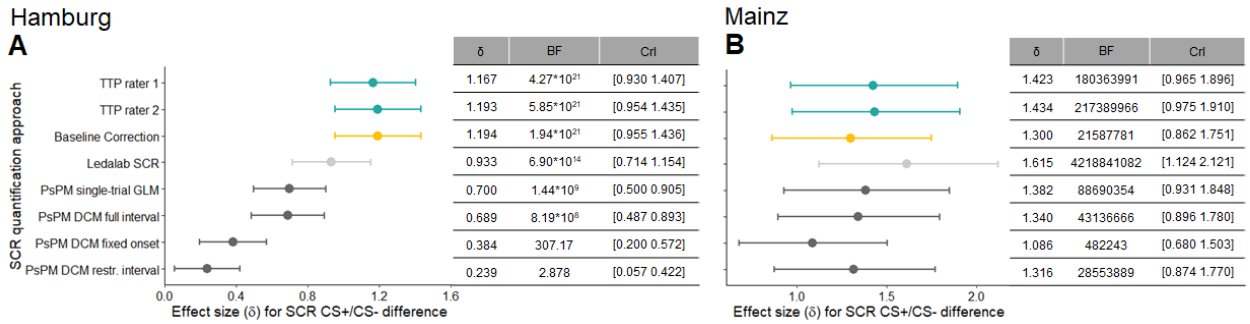
Supplementary Figure 2. *Averaged raw SCRs (plus standard error) for the CS+ (red), CS- (blue) and US (black) for each SCR quantification approach employed in the Hamburg and Mainz datasets split up for the first half of acquisition training (left) and the second half of acquisition training (right).*


Hamburg
1st half
2nd half
SCR quantification approach
TTP rater 1
TTP rater 2
Baseline Correction
Ledalab SCR
PsPM single-trial GLM
PsPM DCM full interval
PsPM DCM fixed onset
PsPM DCM restr. interval
0.000 0.025 0.050 0.075 0.100 0.12
SCR CS+ > CS- difference
0.00 0.25 0.50 0.75 1.00 1.25
Effect size (δ) for SCR CS+/CS- difference
Mainz
0 1 2 3
0.0 0.5 1.0 1.5

Supplementary Figure 3: *Average CS discrimination (+/- CI) based on raw values per CS type during fear acquisition training based on data derived through different SCR response quantification approaches in the Hamburg and Mainz datasets as corresponding effect sizes and credible intervals as derived from the Bayesian paired-sample T-tests for the first half of acquisition training (left) and the second half of acquisition training (right).*

Supplementary Table 1. *Trial-wise agreement (Krippendorff-alpha as well as lower and upper CI bounds) for Trough-to-peak (TTP) rater 1 and 2 in the Hamburg sample (left) and the Mainz sample (right). Note that there were fewer trials in general in the Mainz sample and that only 50% of the CS+ was followed by the US.*

		Hamburg			Mainz		
stimulus	trial N°	K's α	CI lower bound	CI upper bound	K's α	CI lower bound	CI upper bound
US	14	0.9491816	0.8941711	0.9889613			
	13	0.9340825	0.8253001	0.9959611			
	12	0.9807042	0.9355861	0.9987993			
	11	0.9857542	0.959771	0.9975319			
	10	0.9378392	0.8400543	0.9969566			
	9	0.9944447	0.9852576	0.9985307			
	8	0.9783618	0.9444765	0.9981308			
	7	0.9927716	0.9819344	0.9983336			
	6	0.9941949	0.9835254	0.9977273			
	5	0.9426647	0.8605444	0.9956077	0.8787643	0.6012345	1
	4	0.946771	0.8787656	0.9938299	0.8599314	0.5789249	1
	3	0.9680347	0.9027922	0.9985239	0.983124	0.9317327	1
	2	0.9958881	0.9897625	0.9983421	0.8755138	0.6650003	1
	1	0.8664107	0.7492566	0.9512877	0.9880684	0.9475097	1
CS+	14	0.9581475	0.9130992	0.9904578			
	13	0.9667328	0.9262767	0.990461			
	12	0.9113926	0.8209098	0.9824749			
	11	0.9319578	0.8545438	0.9914481			
	10	0.9908385	0.9738787	0.9993003	0.9962208	0.9756577	0.9992429
	9	0.9617094	0.9108156	0.9952467	0.9778019	0.9111852	0.9997573
	8	0.9378922	0.8537252	0.9942212	0.9727613	0.8860747	0.9997294
	7	0.9759806	0.9384362	0.9955748	0.9970024	0.9808084	0.9999459
	6	0.9628767	0.9032798	0.992602	0.9960039	0.9780103	0.9999729
	5	0.9784465	0.9475419	0.9965377	0.9988121	0.9920419	0.9999729
	4	0.9752519	0.9428901	0.994724	0.999946	0.9990497	1
	3	0.9891962	0.9699742	0.9975678	0.9984066	0.9861044	1
	2	0.9830422	0.9565487	0.9975757	0.9907849	0.952272	1
	1	0.9093658	0.8210891	0.9695272	0.9156519	0.7221096	0.9992701
CS-	14	0.9894342	0.9634818	0.9999341			
	13	0.9201136	0.8461459	0.9749781			
	12	0.9682093	0.9188963	0.9998659			
	11	0.878636	0.7832007	0.9536765			
	10	0.9748552	0.9365602	0.999217	0.9764351	0.9141222	1
	9	0.9400161	0.8767708	0.9833651	0.9985776	0.9898428	0.9999728
	8	0.9627718	0.9069285	0.9944651	0.9851123	0.9417591	0.9999724
	7	0.8448468	0.7286466	0.9382485	0.9999438	0.9982612	1
	6	0.9618257	0.9145197	0.9942109	0.9953226	0.972953	1
	5	0.9837512	0.9638328	0.9976714	0.9776089	0.9139633	0.9999721
	4	0.9493747	0.8954762	0.9852882	0.9911666	0.9613912	0.9999729
	3	0.9760664	0.9522766	0.9921131	0.9797595	0.9188521	0.9999187
	2	0.9778288	0.9572033	0.990262	0.9990548	0.9901168	1
	1	0.9109558	0.8179052	0.9719682	0.998137	0.9884094	0.9997834



Supplementary Figure 4. *Effect sizes, Bayes Factors, and credible intervals as derived from the Bayesian paired two-sample t-tests for the Hamburg (A) and Mainz (B) datasets based on z-transformed data (based on a reviewer's request).*