Received: 3 December 2021 | Revised: 21 February 2022 | Accepted: 8 March 2022

DOI: 10.1111/psyp.14058

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

Navigating the manyverse of skin conductance response quantification approaches – A direct comparison of trough-to-peak, baseline correction, and model-based approaches in Ledalab and PsPM

Manuel Kuhn[1,2] | **Anna M. V. Gerlicher**[3,4] | **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, Massachusetts, USA

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

[4]Department of Experimental Psychology, Utrecht University, Utrecht, The Netherlands

Correspondence
Tina B. Lonsdorf, Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Martinistrasse 52, Hamburg, Germany.
Email: t.lonsdorf@uke.de

Funding information
This work was funded by grants from the German Research Foundation (DFG) to TBL (DFG LO1980/7-1, LO 1980/4-1, and CRC58; subproject B07, INST 211/633-1)

Abstract

Raw data are typically required to be processed to be ready for statistical analyses, and processing pipelines are often characterized by substantial heterogeneity. Here, we applied seven different approaches (trough-to-peak scoring by two different raters, script-based baseline correction, Ledalab as well as four different models implemented in the software PsPM) to two fear conditioning data sets. Selection of the approaches included was guided by a systematic literature search by using fear conditioning research as a case example. Our approach can be viewed as a set of robustness analyses (i.e., same data subjected to different processing pipelines) aiming to investigate if and to what extent these different quantification approaches yield comparable results given the same data. 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 data sets with different paradigms suggest that there may be no single approach that consistently yields larger effect sizes and could be universally considered "best." Yet, at least some of the approaches employed show consistent effect sizes within each data set indicating comparability. Finally, we highlight substantial heterogeneity also within most quantification approaches and discuss implications and potential remedies.

KEYWORDS
anxiety disorders, fear, multiverse, psychophysiology, replicability crisis

1 | INTRODUCTION

Scientific work rests fundamentally upon data, their measurement, processing, analysis, illustration, and interpretation. Raw data are typically required to be processed to be ready for statistical analyses and interpretation. Although these processing pipelines can be well defined and standardized, they are often characterized by substantial heterogeneity, particularly in

[Correction added on April 23, 2022 after first online publication: The portrait format of table 1 has been changed to Landscape.]

Biological Psychology and Cognitive Neuroscience (Botvinik-Nezer et al., 2020; Lonsdorf, Klingelhöfer-Jens, et al., 2019; Sandre et al., 2020). A commonly used measure in these scientific disciplines is skin conductance that 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 nonspecific 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. This can be done with a multitude of different response quantification approaches, with any given study typically choosing only one of these options. Already in 1971, Lykken and 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 & Venables, 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 & Venables, 1971, p. 656). As a potential solution to the problem of data processing and statistical heterogeneity, the "multiverse approach" has recently been suggested (Steegen et al., 2016): In data multiverse analyses, the same raw data are processed into a multiverse of processed data sets (referred to as "universes") 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 (i.e., the sum of all universes) inevitably imply a multiverse of statistical results, given a single set of identical raw data and applied statistical models (Lonsdorf et al., 2021; Lonsdorf, Klingelhöfer-Jens, et al., 2019; Lonsdorf, Merz, & Fullana, 2019; Silberzahn et al., 2018; Sjouwerman et al., 2021; Steegen et al., 2016), and can inform on the stability or robustness of the effect of interest against different processing pathways. To this end, multiverse-type of studies have been proposed to explicitly facilitate debates on what (processing or analytical) specifications should be used (Del Giudice & Gangestad, 2021; Simonsohn et al., 2020). Of note, the "full" multiverse consists of an infinite number of options and hence, it has been recognized that many other decisions could be considered than what is typically referred to as "full multiverse" in these types of studies (Del Giudice & Gangestad, 2021). Often, it can be advantageous to focus on a more limited set of decision nodes and investigate these in more depth. Here, we focus on a small-scale multiverse-type of approach (referred to as "manyverse") by comparing SCR quantification approaches derived from a systematic literature search in two data sets and by using fear conditioning research as a case example. As (systematic) robustness analyses such as multiverse-type of studies are per definition applied to the same set of data, we acknowledge that we do not aim for a direct comparison between both data sets as these differ in key experimental specifications. Hence, we provide an SCR response quantification manyverse approach *within* each data set.

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 model-based approaches such as Ledalab (Benedek & Kaernbach, 2010a; Lim et al., 1997) and Psycho-Physiological Modelling (PsPM; Bach et al., 2009, 2013; Bach & Friston, 2013), and (iii) what we here refer to as "baseline correction" approaches. Of note, however, these approach categories are by no means homogeneous and different specifications and settings can be applied. We refer, for instance, to our related work that focuses on an in-depth investigation of *within*-approach heterogeneity of specifications used 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, for instance, in fear conditioning research (Lonsdorf, Klingelhöfer-Jens, et al., 2019; Lonsdorf, Merz, & Fullana, 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, model-based approaches, and baseline correction approaches (as well as their subcategories).

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 prespecified 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 an onset latency time window (OLW) of 1–3 s (Levinson & Edelberg, 1985), 1–3.5 s (although stimulus-specific response windows were suggested, Sjouwerman & Lonsdorf, 2019), or 1–4 s (Boucsein et al., 2012) after stimulus onset. The SCR peak value is then required to occur in a peak detection time window (PDW) of 0.5–5 s after SCR onset (i.e., footpoint; Boucsein et al., 2012). More precisely, if the footpoint occurs 2 s after the stimulus presentation, the peak must occur in a time window of 2.5–7 s after stimulus onset. Some authors have also used the full stimulus duration (or even longer) as the PDW without explicitly distinguishing between OLW and PDW. In addition, a minimum response—typically varying between 0.05 and 0.01 μs—is often applied (Boucsein et al., 2012; Lonsdorf, Klingelhöfer-Jens, et al., 2019). SCRs smaller than this minimum response are not considered as a valid response and included as nonresponse with a value of zero (Lonsdorf et al., 2017; i.e., "magnitude," Venables & Christie, 1980). Consequently, TTP scoring can only yield SCR values with a zero or a positive value.

TTP scoring employing the above-described criteria can be performed as follows: (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) 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 on the scale used to view the data different inflection points may be detected through visual inspection), requiring long interstimulus 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 70 s and new work has not reinvestigated assumptions regarding an 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 the "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" that 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 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 2 s immediately preceding stimulus onset from the highest SC level value during the entire stimulus presentation period (i.e., 8 s; 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 postprocessing (e.g., Vogel et al., 2015). Today, BLC approaches are most 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 (for a discussion, see Sjouwerman et al., 2021).

1.1.3 | Computational model-based approaches

Last, 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 labeled SCRalyze; Bach et al., 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 function, the treatment of slowdrifts 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 interstimulus intervals as SCRs are slow responses and rapidly spaced stimuli with an interstimulus interval (ISI) of 2–3 s 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 <3 s. 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), the 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 nonlinear dynamic causal modeling (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 nonlinear 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 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 detailed summary of the status quo) 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 it is noteworthy that conclusions derived from these studies are similarly heterogeneous. 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 model-based methods as implemented in SCRalyze are more sensitive than the "peak-scoring" approach and provide significantly higher predictive validity than any Ledalab measure in most of the tested contrasts (Bach, 2014).

Speculations on potential explanations for these conflicting results and conclusions may be derived from Table 1. As this is, however, beyond the main aim of the present work, we refer the interested reader to the Supplementary Material for an in-depth discussion.

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 (when applied to an identical data set) can be used interchangeably (jingle fallacy). 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 extent a given effect can be formally "replicated" by subjecting single data sets to multiple theoretically equally justifiable SCR response quantification approaches (i.e., robustness analyses).

First, we need a synopsis of the different approaches employed in the literature as well as their abundance to guide the decision on approaches to compare. 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 data sets. Note that we do not aim for a direct comparison between both

TABLE 1 Overview over comparative work on different SCR quantification approaches

	Trough-to-peak (TTP)			Computational				Validation datasets				Duration	
Author	**Computer assisted[a]**	**Autonomate**	**script-based[a]**	**Ledalab**	**SCRalyze GLM**	**SCRalyze DCM**	**BLC[b]**	**Paradigm**	**N**	**RI rate[c]**	**Nr of trials per condition[d]**	**CS/cue (in s)**	**ITI (in s)**
Kuhn et al. (2022), present study	✓ $OLW_{CS(1)}$: 0.9–3.5 s $OLW_{US(1)}$: 0.9–2.5 s $OLW_{(2)}$: 0.9–4 s PDW: 0–5 s	✗	✗	✓ CDA, response window: 0.9–4 s	✓ Single-trial GLM (parameter estimates)	✓ Full interval, onset only, restricted interval	✓BL: −2 s $PDW_{(1)}$: 6 s $PDW_{(2)}$: 4.5	1. Fear conditioning (Lonsdorf, Klingelhöfer-Jens, et al., 2019)	118	100%	Acq: 15	6–8	10–16
								2. Fear conditioning (Gerlicher et al., 2018)	38	50%	Acq: 10	4.5	17, 18 or 19
Privratsky et al. (2020)	✗				✓[e] Condition-wise GLM (reconstructed) two single-trial GLMs (reconstructed)	✗	✓ BL: −2 s $PDW_{(1)}$: 1–3.5 s $PDW_{(2)}$: 1–4 s $PDW_{(3)}$: 1–4.5 s $PDW_{(4)}$: 1–5 s	Fear conditioning	65	50%[f]	Pre: 6 Acq: 18 Ext: 18	3	2–6
Green et al. (2014)	✓[g] OLW: 1–4 s PDW: 0.5–5 s	✓ OLW: 1–4 s PDW: 0.5–5 s	✗	✓[h] CDA	✓[i] Condition-wise GLM	✗	✗	1. Fear conditioning (Huff, Hernandez, Blanding, & LaBar, 2009)[j]	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)[j]	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	–

(Continues)

TABLE 1 (Continued)

	Trough-to-peak (TTP)			Computational				Validation datasets				Duration	
Author	Computer assisted[a]	Autonomate	script-based[a]	Ledalab	SCRalyze GLM	SCRalyze DCM	BLC[b]	Paradigm	N	RI rate[c]	Nr of trials per condition[d]	CS/cue (in s)	ITI (in s)
Pineles et al. (2009)	✗	✗	✓[k] OLW_{FIR}: 1–4 s PDW_{FIR}: 2–6 s	✗	✗	✗	✓BL: −2 s PDW: 0–8 s	Fear conditioning	287	100%	Pre: 5 Acq: 5 Ext: 10	8	15–25
Bach (2014)	✗	✗	✓[l] OLW: 1–3 s PDW: 0.5–5 s	✓[m] Response window: 1–4 s Mean of 4 measures (CDA, DDA)	✓[n] Condition-wise GLM (reconstructed)	✗	✗	1. Passive picture viewing (IAPS, neutral, aversive)	60	–	45	1	7.65, 9, 10.35
								2. Passive picture viewing (IAPS, neutral, aversive, positive)	38	–	16	1	4.4
								3. Passive face viewing (KDEF, angry, fearful, neutral)	42	–	38	1	7.65, 9, 10.35
								4. Neutral picture viewing (IAPS, neutral)	61	–	45	1	7.65, 9, 10.35
Bach et al. (2013)	✗	✗	✓[o] OLW(1): 1–3 s OLW(2): 1–4 s PDW: 0.5–5 s	✗	✓[p] Condition-wise GLM (parameter estimates) single-trial GLM (parameter estimates)	✓[q] DCM informed, DCM uninformed about event onsets	✓BL: −1 s OLW: 1–4 s	1. Passive picture viewing (IAPS, negative, neutral)	60	–	45	1	7.65, 9, or 10.35
								2. Passive picture viewing (IAPS, neutral, aversive, positive)	40	–	16	1	4.4

TABLE 1 (Continued)

	Trough-to-peak (TTP)			Computational				Validation datasets				Duration	
Author	Computer assisted[a]	Autonomate	script-based[a]	Ledalab	SCRalyze GLM	SCRalyze DCM	BLC[b]	Paradigm	N	RI rate[c]	Nr of trials per condition[d]	CS/cue (in s)	ITI (in s)
Bach et al. (2010)	✗	✗	✗	✗	✓[r] condition-wise GLM (parameter estimates) single-trial GLM (parameter estimates / reconstructed)	✓[s] full interval, first/second half of interval	✓ BL: -1s PDW: CS duration[t]	1. Fear conditioning	32	50%[f]	32	4, 10 and 16	14, 19 or 23

Notes: This table serves the purpose to provide in in-depth overview on the status quo on comparative studies on different SCR quantification approaches and to provide the interested reader with detailed methodological information that may guide future investigations (see supplementary material for a discussion).

[a]OLW, onset latency window (post-CS with CS onset serving as 0 s), PDW, peak detection window (post SCR onset with SCR onset serving as 0 s).

[b]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 BLC 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 BLC approaches (e.g., Privratsky et al., 2020).

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

[d]Acq, acquisition training; Ext, extinction; Pre, pre-conditioning; Gen, generalization; Ren, renewal.

[e]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 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).

[f]Only non-reinforced trials were used for analyses across SCR response quantification approaches.

[g]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.

[h]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.

[i]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.

[j]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.

[k]In addition to the first interval response (FIR) for which scoring criteria are reported in the table, also the so called second interval response (SIR; OLS_{SIR}: 4–8 s PDW_{SIR}: 5–9.5 s) and the so called "entire interval response" (EIR) using the entire CS-UCS interval (0–8 s) 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 0.03 s 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 2 s immediately preceding stimulus onset from the highest SC level value during the entire stimulus presentation period (i.e., 8 s; Pineles et al., 2009).

[l]Response quantification was performed in the software program Matlab.

(Continues)

TABLE 1 (Continued)

[m]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 analyzed, 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.

[n]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.

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

[p]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 s post-stimulus time bins of 1 s 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 0.005, 0.1, 0.0159 Hz, and from 0.02 to 0.10 Hz in steps of 0.005 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 0.05 Hz.

[q]Informed DCM: the DCM is informed about stimulus onsets and SN burst are assumed to occur 2000 ms 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 1–4 s post stimulus time window was extracted and the largest response in each time window was retained.

[r]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).

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

[t]When the CS duration was <5 s, then a PDW of 5 s was used.

data sets as these differ in more than a single specification (e.g., CS and ITI duration, reinforcement rate, sample size) but provide a manyverse analysis within each data set. Note that the multiverse approach focuses on applying different pipelines to the same underlying data. 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 subaspect of replicability. While multiverse analyses often focus on the distribution of *p* values across the multiverse (e.g., Steegen et al., 2016), we extend this somewhat limited focus by also considering effect sizes and precision of the estimates.

Third and finally, we include TTP scoring from two independent raters per data set (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 the robustness of the results across the different SCR quantification approaches, this would argue in favor of the interchangeable use of different SCR quantification approaches. 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 making analysis decisions and comparing SCR results.

2 | METHOD

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 6 months prior to March 22, 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 (Lonsdorf, Klingelhöfer-Jens, et al., 2019) and the present work. As described in Lonsdorf, Klingelhöfer-Jens, 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 Lonsdorf, Klingelhöfer-Jens, et al., 2019). A subset of the identified SCR quantification approaches was subsequently applied to two independent data sets (see below for details). The literature search here served the purpose to guide our decision on which approaches to apply here and to obtain an overview of what is commonly used in the literature. Hence, the literature search can be considered a tool rather than an aim in its own right.

2.2 | Participants and experimental paradigms

2.2.1 | Data set 1: Hamburg

Participants

Data set 1 consisted of the acquisition phase (i.e., Day 1) from the baseline (T_0) 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). Data set 1 has been included as a case example in a previous publication (Lonsdorf, Klingelhöfer-Jens, et al., 2019) focusing on methodological questions (i.e., exclusion of "nonlearner" and "nonresponder" in fear conditioning research).

Paradigm and stimuli

The paradigm (for details, see Lonsdorf, Klingelhöfer-Jens, et al., 2019) consisted of a 2-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 (T_0) and a follow-up measurement (T_1) 6 months 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 T_0 were included. All data sets were trimmed to this period of interest starting 2 s prior to the first event of interest (i.e., first CS presentation during acquisition training) and ending between 10 and 20 s (20 s trim cutoff

value) after the last event of interest (i.e., last CS or US presentation during acquisition training). Two light gray fractals served as conditioned stimuli that 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 intertrial 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 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 the middle finger. US intensity was individually calibrated in a standardized step-wise procedure aiming at an unpleasant, but still tolerable level.

2.2.2 | Data set 2: Mainz

Participants

Forty male participants (mean ± SD age of 28.1 ± 2.7 years) were included in the data set that 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 artifacts, leaving data of $n = 38$ participants for statistical analysis of each phase.

Paradigm and stimuli

Data set 2 consists of a 3-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 and 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. Intertrial intervals lasted 17, 18, or 19 s (mean of 18.5 s). The 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 10 CS+ and 10 CS− trials in context A. Five out of 10 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 the 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 a 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

Data set 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 1000 Hz with a gain of 5 μΩ/V and a 1.0 Hz hardware filter.

Data set 2 (Mainz)

Electrodermal activity was recorded from the thenar and hypothenar of the nondominant 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 data sets were trimmed to 5 s prior to the first event of interest (i.e., first CS presentation during acquisition training) and 22 s 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 cutoff frequency of 1 Hz using Matlab 2019a (Mathworks®, Natick, Massachusetts, USA).

2.3.2 | SCR quantification approaches employed

We applied three different response quantification approaches including their subcategories to both data sets: TTP was employed by two different raters for each data set, 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 data sets 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. The 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 (Lonsdorf et al., 2021) and as illustrated for fear extinction (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 Section 3.1.

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 0.9–4 s (Boucsein et al., 2012) poststimulus 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. This approach corresponds to what has been recommended by the Society for Psychophysiological Research (see Boucsein et al., 2012) and corresponds to what has been referred to as the so-called "first-interval response" in fear conditioning research. Provided the CS–US interval is sufficiently long (i.e., 6–10 s; Stewart et al., 1961) three SCR components that map onto different underlying processes can be distinguished in fear conditioning studies which have been referred to as the first-interval (FIR), second-interval (SIR), and third-interval responses (TIR). More precisely, the FIR (SCR onset: 1–4 s post-CS onset) is considered an orienting response while the SIR (SCR onset: 4 s post-CS onset to 1 s after CS onset) is thought to reflect anticipatory responding to the soon to be presented US and typically occurs only after contingency learning (Ohman, 1972). Finally, the TIR is the response to the US itself. This work on the three different components dates back to the 70s (Ohman, 1972; Prokasy & Ebel, 1967; Rescorla & Wagner, 1972) but the distinction between these three intervals has not been universally adopted (for a summary and critique, see Pineles et al., 2009). In fact, "Of the two anticipatory response components, the first is usually larger than the second and, because it is highly sensitive to conditioning manipulations, it is frequently the only one reported" (Lipp, 2006), possibly also because the FIR has been shown to have higher reliability than the SIR (Fredrikson et al., 1993). It is also important to note that the assessment of the SIR is often not possible when the CS–US interval is too short or when startle probes are included in the experimental design (i.e., the SCR to the probe confounds the SIR).

Raters 1 (TTP1) were experienced raters and Raters 2 (TTP2) were first-time raters for both data sets but different individuals for these data sets resulting in a total of 4 raters. 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 3.5 s and the US as 0.9–2.5 s 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 published recommendations by Boucsein et al. (2012), which were applied for TTP2 (Hamburg) and both Mainz rater.

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

Data were downsampled 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, see Supplementary Material for examples). Data with recording artifacts (i.e., in more than half of the trials) were treated as missing data points and excluded from the analyses. For the Hamburg data sets, SCRs below 0.01 μS or the absence of any SCR (i.e., flat line or habituation drift) within the defined time window were classified as non-responses and set to 0. The threshold of 0.01 μS for this data set was determined empirically by visually inspecting response specifically above and below this cutoff (Lonsdorf, Klingelhöfer-Jens, et al., 2019), which suggested that in this data sets, responses >0.01 μS can be reliably identified. For the Mainz data sets, a minimum amplitude criterion of 0.02 μS was used.

Baseline correction (BLC)

A custom-made script in Matlab version R2019b (Mathworks®, Natick, Massachusetts, USA) implemented the BLC response quantification approach by subtracting the mean of the 2 s 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 (6 s; as CS duration was jittered between 6 and 8 s) for the Hamburg sample and the full CS duration (4.5 s) for the Mainz sample for both, CS and US, stimuli. In light of a substantial degree of heterogeneity in the 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: −2 s, 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 BLC 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 BLC approach will be discussed elsewhere (Sjouwerman et al., 2021).

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 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 were downsampled to 10 Hz. The "optimize" function, as implemented in Ledalab, was used using default settings. The response window was defined as 0.9–4.0 s after stimulus onset. The minimum thresholds of SCRs were 0.01 and 0.02 μS for the Hamburg and the Mainz data sets, 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 amplitude (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.

PsPM

PsPM single-trial GLM. All Psychophysiological Modeling analyses (PsPM 4.3.0 [Bach et al., 2018]) were conducted in Matlab 2019b. To capture the nature of increasing SCRs over time in the fear conditioning paradigm due to learning, single-trial modeling 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 and 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. Nonlinear modeling (dynamic causal, DCM) in PsPM employs a nonlinear inversion algorithm to infer single-trial estimates of sudomotor impulse response magnitude (Bach et al., 2010). Following the PsPM manual, in the first model, we applied a "full interval" model in which the SCR onset, and its onset latency as implemented in PsPM, can be modeled 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 the US and US omission and erroneously assigning it to the CS+. Thus, these models are not recommended for analyzing reinforced SCR trials that are particularly problematic for experimental designs with 100% or high reinforcement rates. More precisely, PsPM's manual states for the nonlinear 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, 2015) were conducted in R to assess CS+/CS− discrimination. The package's *t* test BF function was used with 1000,000 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) were calculated and the Bayes factor was extracted using the extractBF function. To provide complementary analyses that provide results based on most commonly employed frequentist 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 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 data set to 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 should be considered far (or conceptual) replications in the data sets used here. Even though LeBel et al. used a frequentist 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 intraclass 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 *intraobserver* and *interobserver* 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 interrater or intermethod 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 (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 a trial-by-trial pairwise 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, α of ≥0.8 is required for agreement (Krippendorff, 2004). Benchmark values have been suggested (Landis & Koch, 1977) for interpretation of the strength of agreement as substantial (0.61–0.8), moderate (0.41–0.6), and fair (0.21–0.40). 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 trough-to-peak (TTP) scoring ($n = 24$) and baseline correction (BLC) 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), whereas 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 PsPM/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 as 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-based) were not homogeneous in themselves as

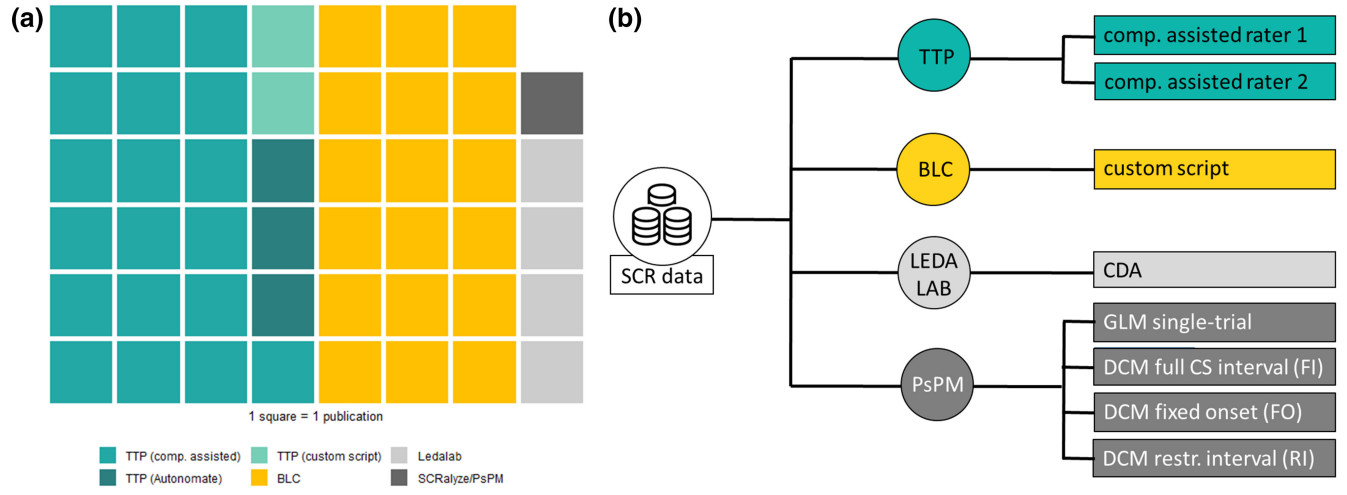

FIGURE 1 (a) Frequency of different SCR quantification approaches exemplified from the systematic literature search which included 47 publications, published between 06/2018 and 02/2019. (b) Illustration of the different SCR quantification approaches employed to the two independent data sets in the current work: trough-to-peak (TTP), baseline correction (BLC), Ledalab, as well as PsPM (formerly SCRalyze) with four different specifications

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, BLC, 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.

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 for both data sets (see Figure 2a,b) as well as averaged SCR values across all trials per stimulus type (i.e., CS+, CS−, and US, see Figure 2c,d). On a descriptive level, in both data sets (Hamburg, Mainz), the trial-by-trial trajectories appear to follow a similar pattern when responses are quantified through the TTP, BLC, 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], and 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 the Mainz data sets) and the RI approach (primarily Mainz data set) apparently yielded larger CS+ responses but substantially smaller US responses. This was particularly pronounced in the Mainz data sets in which the CS duration was shorter than in the Hamburg study (Mainz: 4 s, Hamburg: 6–8 s jittered) and the reinforcement ratio was partial (50%) while it was full (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

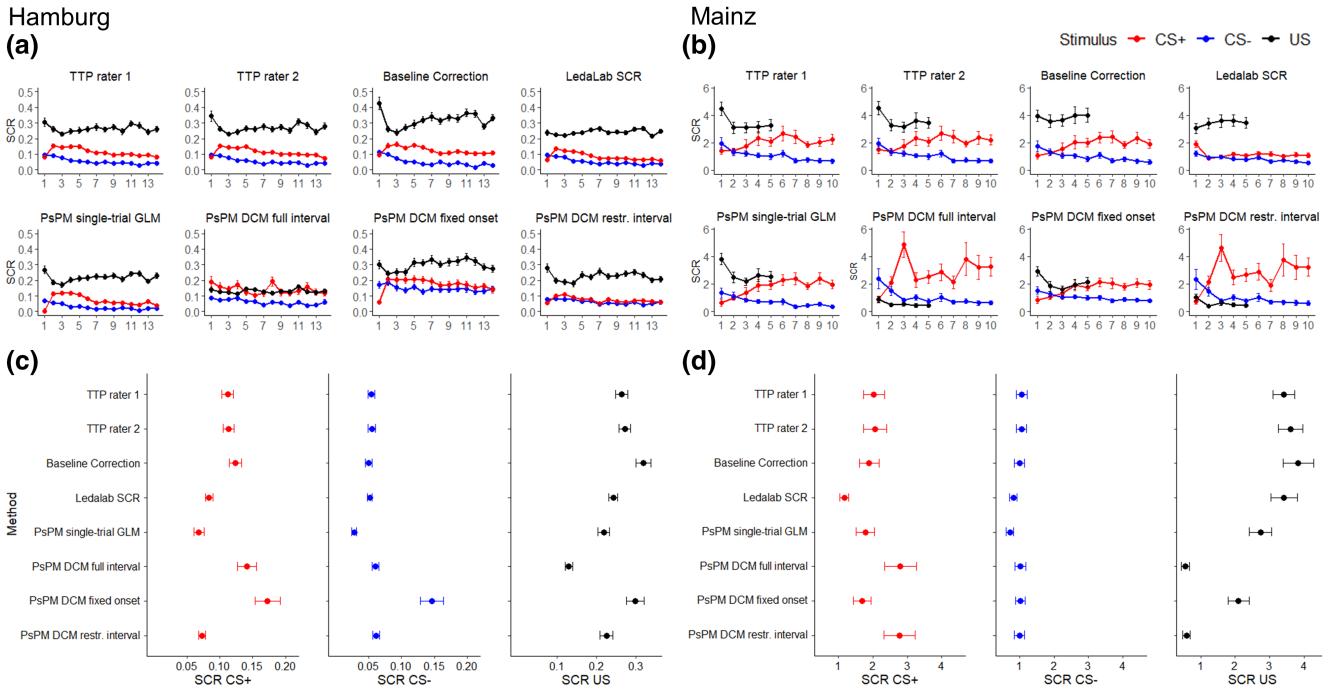

FIGURE 2 Trial-by-trial trajectories for the CS+ (red), CS− (blue), and US (black) during fear acquisition training for the Hamburg **(a)** and Mainz **(b)** samples illustrated for all different SCR quantification approaches employed: TTP Rater 1 and TTP Rater 2, baseline correction (BLC), Ledalab, PsPM single-trial GLM, PsPM DCM with flexible response onset in full CS interval (FI), PsPM DCM with the 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)** data sets are shown. Supplementary Figure S1 illustrates trial-by-trial average values derived from the different quantification approaches in a single figure, and supplementary Figure S2 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, whereas 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 the US and US omission and erroneously assigning it to the CS+ when the CS–US interval is short

noting that PsPM DCM models that allow for a flexible response onset come with the risk of absorbing SCRs elicited by the 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 data set, the DCM modeling the full interval (FI) seem to underestimate US responses and instead overestimate reinforced CS+ responses (note, order of CS+ responses differed between participants). Thus, these models do not seem suitable for analyzing reinforced SCR trials that are 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., modeling the full CS duration of 4.5 s) and the restricted interval model (i.e., modeling 0–4 s post-CS onset) unsurprisingly result in near-identical results as the CS duration (4.5 s) was only 0.5 s longer than the definition of the restricted interval (i.e., 4 s post-CS onset). In the Hamburg data set in which the CS duration was longer (6–8 s jittered), however, both approaches differ substantially (i.e., full interval modeled 0–6, 0–7, or 0–8 s, restricted interval: 0–4 s). 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 modeled as 0–4 s 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.

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

Both frequentist (Figure 3a,b) and Bayesian (Figure 3c,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 < .003, BFs > 7.16, see Figure 3a,b), even though CS discrimination values

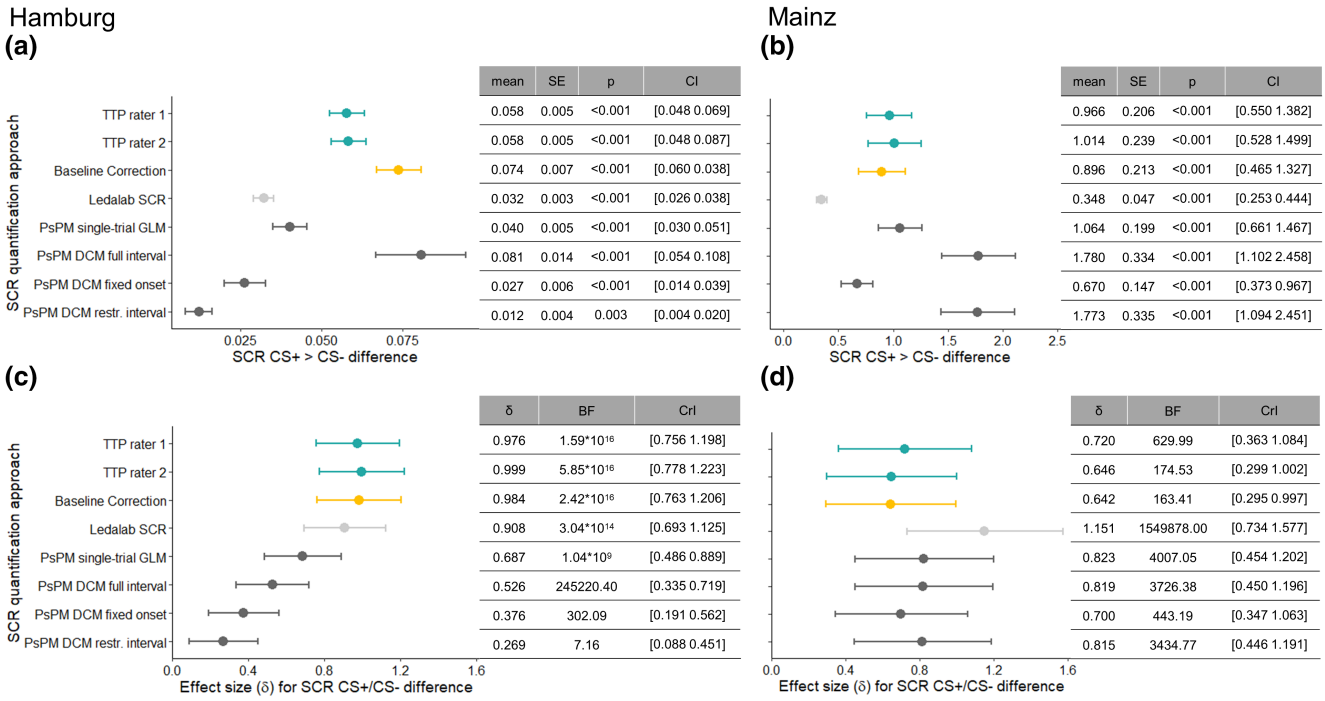

FIGURE 3 CS discrimination (based on raw values per CS type during fear acquisition training) based on data derived through different SCR quantification approaches in the Hamburg (a, c) and Mainz (b, d) data sets. A and B show mean CS discrimination (±standard error) and results as a table (i.e., mean, *p* values, confidence intervals) from paired-sample *t* tests, whereas c and d show corresponding effect sizes (±credible intervals) as well as results as a table (i.e., Bayes factors and credible intervals) as derived from the Bayesian paired two-sample *t* tests for the Hamburg (c) and Mainz (d) data sets. Supplementary Figure S3 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 S4)

differed numerically between approaches. Similarly, resulting effect size estimates derived from Bayesian paired two-sample *t* tests (Figure 3c,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 data sets 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 (4.5 s vs. 6–8 s), a number of trials (10 vs. 14), and reinforcement rate (50% vs. 100%) as well as the sample size (38 vs. 118 participants).

3.4 | Formal comparison of robustness of results across SCR 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+ compared with 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 samples.

Second, effect sizes are more precise in the larger Hamburg data sets (CrI width [min–max]: 0.363–0.445) compared with the smaller Mainz data set (CrI width [max–min]: 0.702–0.843), $t(7) = -16.12$, $p < .001$, but are rather similar within different approaches applied to the data of one data set.

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) data set, effect sizes derived from TTP1 and TTP2 as well as TTP1 and BLC and TTP2 and BLC were consistent with each other. For the Hamburg data set, effect sizes derived from these three approaches (TTP1, TTP2, and BLC) were consistent with those derived from Ledalab while they were inconsistent with those derived from Ledalab in the Mainz data set with Ledalab resulting in larger effect sizes than any of the other approaches.

For the Mainz data set, all pairwise comparisons between effect sizes derived from any of the four PsPM models and the four other approaches (TTP1, TTP2, BLC, and Ledalab) yielded consistent effect 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 data set are wide and larger sample sizes may result in a different conclusion.

In the Hamburg data set, 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, and 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, and 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

TABLE 2 Pairwise consistency between different SCR quantification approaches with ✓ indicating consistency and ✗ indicating nonconsistency 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)

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

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 compared with 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' effect sizes but the effect sizes derived from the full interval (FI) model were inconsistently larger than those derived from the reduced interval (RI) model.

3.5 | Agreement between different SCR quantification approaches

Across all SCR quantification approaches, trial-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 data sets. Yet, the pattern of pairwise agreement was often not consistent across both data sets. In the Hamburg data set, the agreement seems to be lowest for the CS- trials, whereas in the Mainz sample, the agreement seems to be lowest for the US trials.

3.6 | Secondary question: Interrater comparisons for computer-assisted TTP scoring

For both data sets (Hamburg and Mainz), two independent raters quantified SCRs through computer-assisted TTP scoring whereof Rater 1 at both sites was experienced and Rater 2 at both sites was the first-time rater (note that Raters 1 and 2 were different individuals for both sites, i.e., there were a total of 4 raters). Note, however, that Hamburg Rater 1 and Rater 2 used slightly different scoring criteria (i.e., 0.9–3.5 and 0.9–4.5 s OLWs). Formal interrater reliability coefficients using Krippendorff's 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 S1) revealed near-perfect agreement across trials for the Hamburg sample [0.845, 0.996] and the Mainz sample alike [0.860, 1].

Figure 6 illustrates the excellent interrater 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.

4 | DISCUSSION

Here, we provide a comparison between seven different SCR quantification approaches in two data sets. The overarching aim of this work was to (a) evaluate if and to what extent seven different approaches lead to comparable results as well as (b) investigate the interrater agreement between two individuals performing TTP scoring in two data sets.

4.1 | Take-home message from the systematic literature search

Our work departs from a systematic literature search that was intended to guide our selection of the to be included

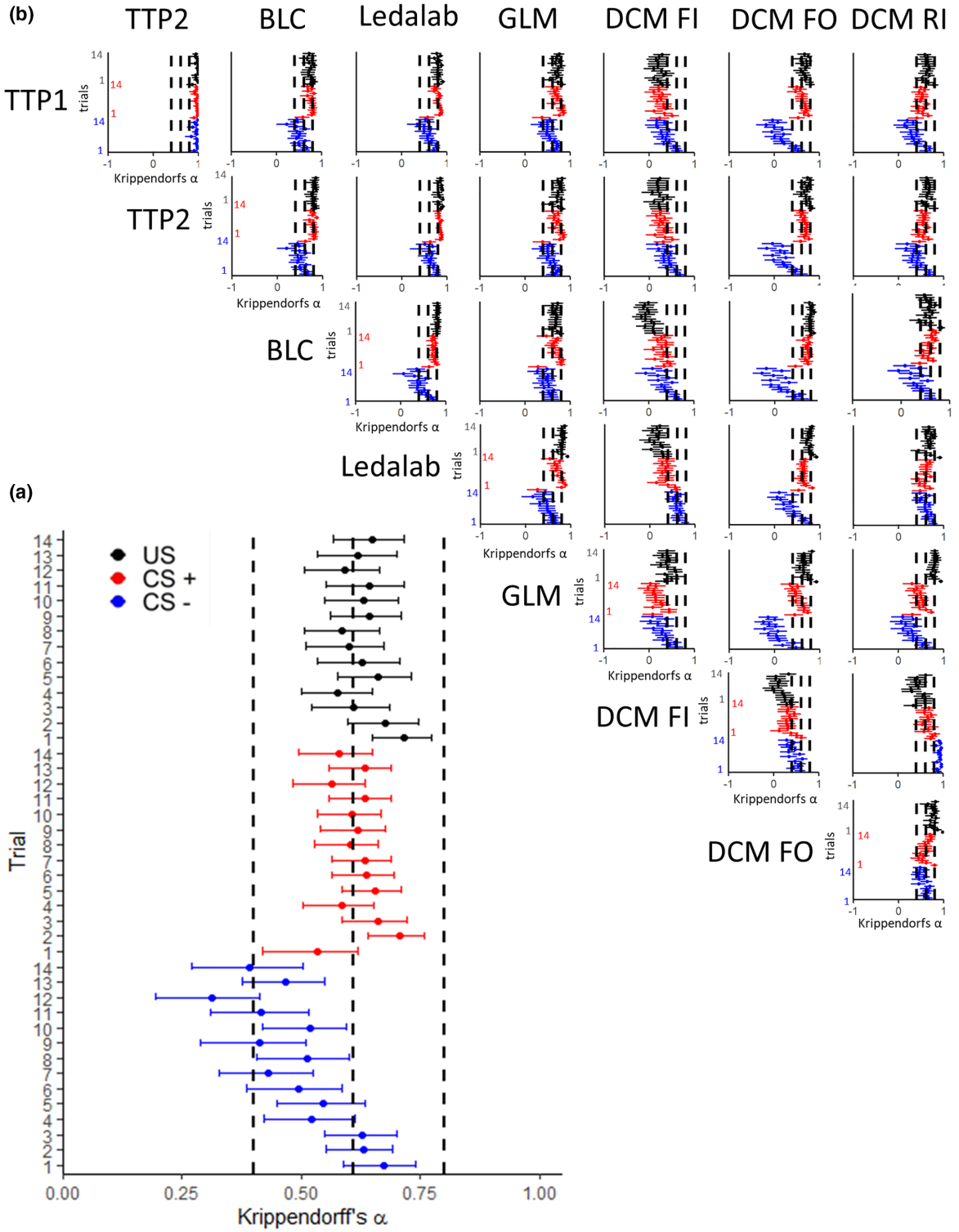

FIGURE 4 Krippendorff's alpha (and CIs) as a measure of agreement between SCR quantification approaches, as calculated in the Hamburg sample (a) across all eight approaches employed for each trial during fear acquisition training. And as calculated (b) for pairwise comparisons between the eight different 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 0.8 and 0.4 highlighting benchmarks for near-perfect agreement (>0.80) and fair to poor (<0.41) according to the benchmarks suggested by Landis and Koch (1977). According to the benchmarks by Landis and Koch (1977), values can be interpreted using the following benchmarks for Krippendorff's $\alpha < 0$ "poor" agreement, 0 to 0.2 "slight," 0.21 to 0.40 "fair," 0.41 to 0.60 "moderate," 0.61–0.80 "substantial," and 0.81 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

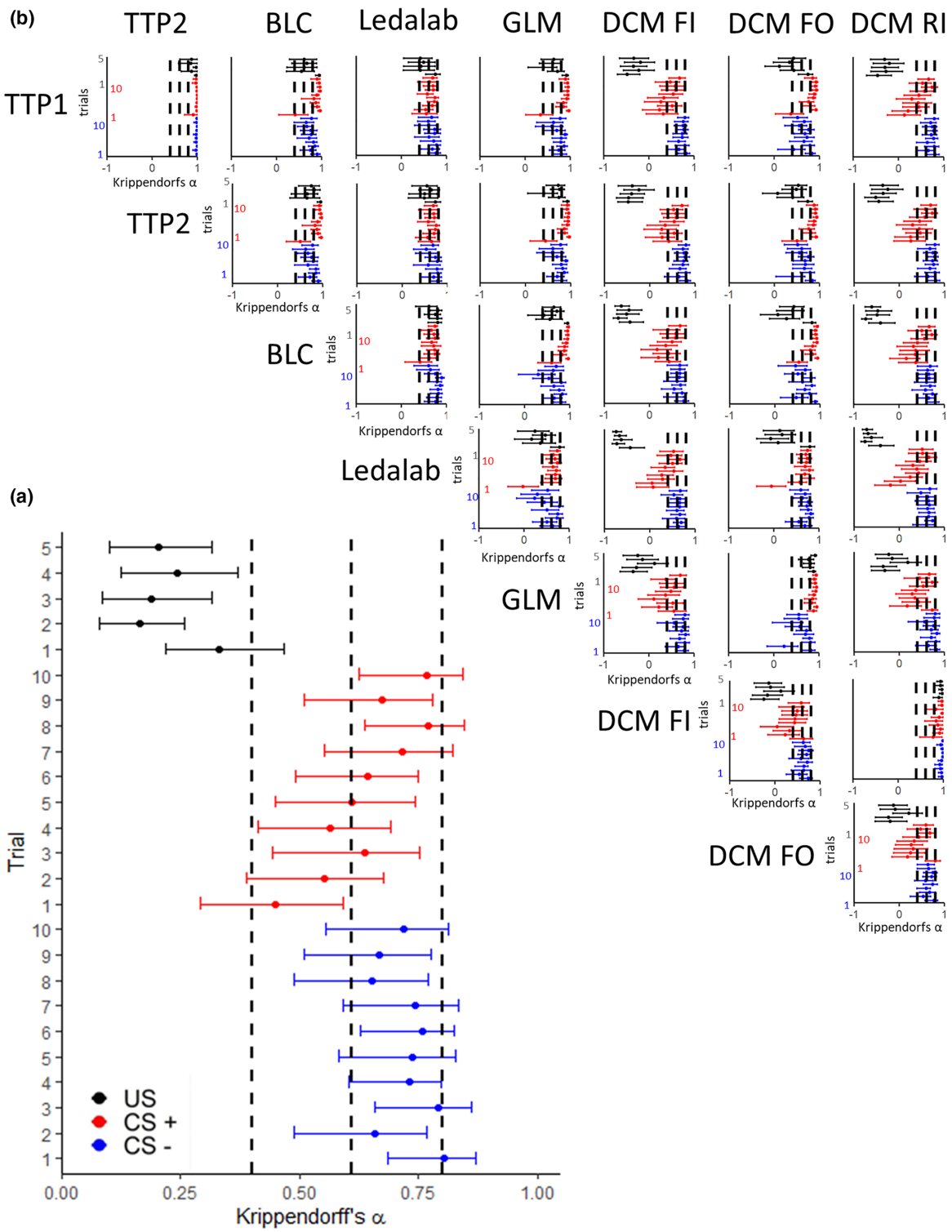

FIGURE 5 Krippendorff's alpha (and CIs) as a measure of agreement between SCR quantification approaches as calculated in the Mainz sample (a) across all eight approaches employed for each trial during fear acquisition training and as calculated (b) for pairwise comparisons between the eight different 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 0.8 and 0.4 highlighting benchmarks for near-perfect agreement (>0.80) and fair to poor (<0.41) 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 Krippendorff's $\alpha < 0$ "poor" agreement, 0 to 0.2 "slight," 0.21 to 0.40 "fair," 0.41 to 0.60 "moderate," 0.61–0.80 "substantial," and 0.81 to 1 "near perfect." Note that trial sequences on the *y* axis in the smaller tiles in panels C and D are identical to the trial sequence on the *y* axis in A and B

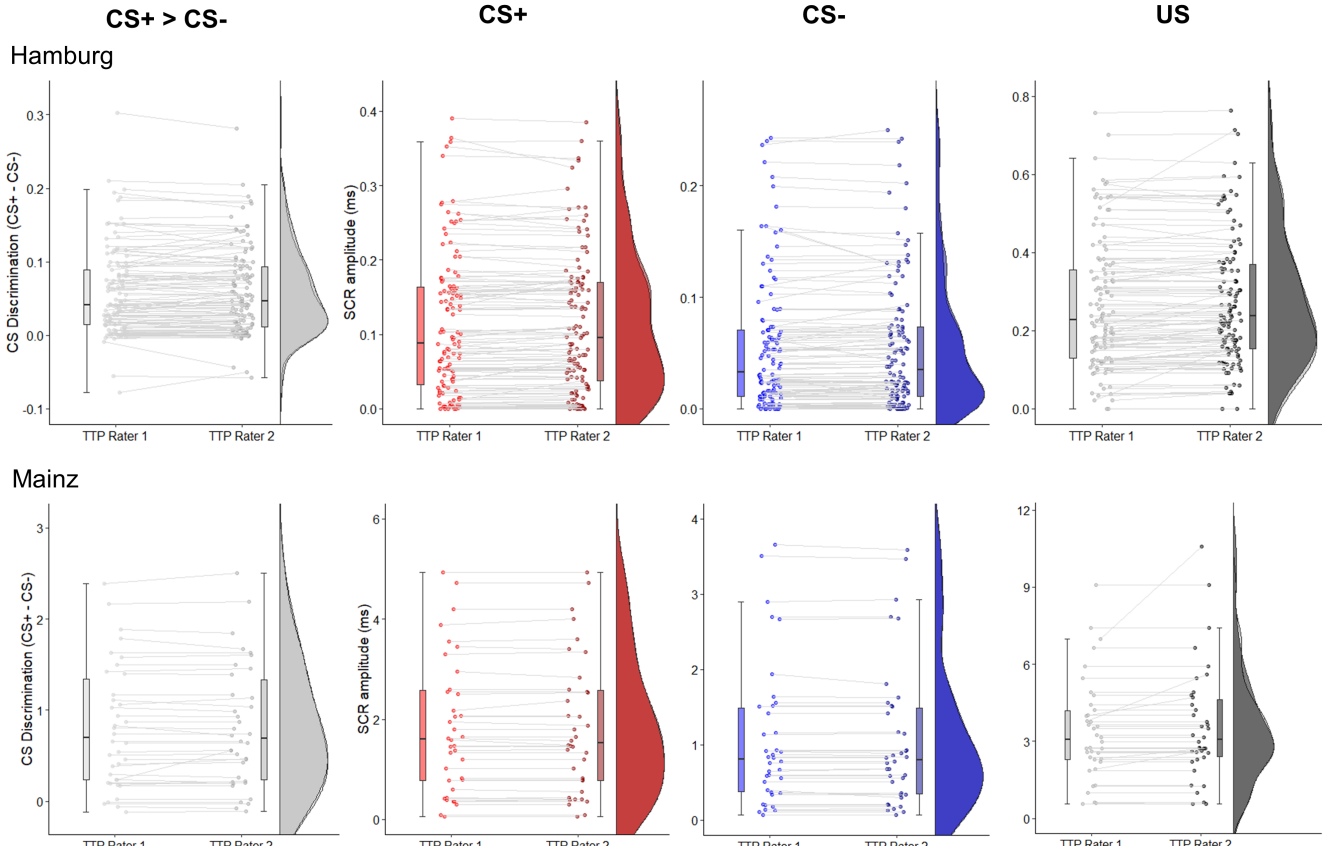

FIGURE 6 Interrater comparisons between TTP Rater 1 and TTP Rater 2 for the Hamburg sample (upper row) and the Mainz sample (lower row) for single-trial discrimination (light gray) as well as single-trial SCRs for the CS+ (red), CS− (blue), and the US (dark gray) during fear acquisition training. Subplots show single-trial or pairwise 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 Raters 1 and 2 were different individuals in the Hamburg and Mainz samples. Also note that both raters used the same criteria in the Mainz sample, whereas in the Hamburg sample, both raters used slightly different criteria to allow for a direct comparison of two previously suggested sets of criteria (see Methods for details)

SCR quantification approaches. Even though the literature search hence mainly served as a tool, some important take-home messages can be derived: *First* (computer-assisted) TTP scoring and BLC through custom-made scripts seem to be the prevailing approaches for SCR quantification in fear conditioning research to date. Our literature search, however, covers only articles published in a 6-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. Yet, a recently published study that focuses on different filter settings in SCR quantification also included a systematic literature search of fear conditioning studies covering 2019 and 2020 (Privratsky et al., 2020) and the frequencies that can be derived from the Supplementary Material seem comparable to what we found.

Second, the SCR quantification approaches identified (i.e., TTP, BLC, Ledlab, and PsPM) do not represent unitary methods but come in heterogeneous specifications (see, e.g., 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 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, the 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 as an in-depth discussion is beyond the scope of the present work.

Third, we noticed that navigating among the different SCR 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 that 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 induced 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 0–3 s (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 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 et al. (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. In conclusion, we see an urgent need for more standardization in the field with respect to the definition of time windows, peak detection (first, largest), and reporting standards.

4.2 | Comparison between different approaches

Here, we applied seven different SCR quantification approaches to two independent data sets in a manyverse 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 data sets—whereof one first-time rater and one experienced rater to allow for the assessment of interrater 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. 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 interrater concordance rates reported here are near perfect for both data sets 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 the 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 the reliability of TTP scoring is excellent. Note, however, that all four raters were directly (both raters for the Hamburg data set, experienced rater for the Mainz data set) or indirectly (new rater for the Mainz data set) trained by the senior author (T.B.L.) and it cannot be excluded that agreement between raters trained in different research groups may yield less-consistent results. A future direction could be to have different labs using the TTP approach scoring the same data set and investigating the convergence rates (i.e., many labs approach).

Relatedly, we note also substantial heterogeneity in the time windows and peak definitions (e.g., first peak, highest peak) used for TTP scoring in the literature. For instance, our literature search revealed that some authors use what corresponds to the First-interval response (FIR) in fear conditioning research (i.e., onset latency window, 0.9–4 s or 0.9–3.5 s) as used here, whereas others identify a peak in the entire CS duration (or entire CS duration +0.5 s) window

starting from CS onset, CS onset +0.5 s, or CS onset +1 s or in a time window that spans the full CS duration (or starting from CS onset +1 s) to 2 s after CS offset (the latter of which likely partly captures the SCRs to the US as this also seems to be applied to reinforced CS+ trials). Hence, future work should also focus on the role of between-study heterogeneity in TTP scoring between different laboratories which could also be done in a many labs approach.

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

The application of different SCR quantification approaches to the same data sets 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 within each data set. 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 quantification approaches.

In sum, a *signal* (i.e., significant CS discrimination) was universally observed in both data sets irrespective of the 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 subtler 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 quantification approaches applied within the data sets, which is novel and relevant information that has not been addressed before.

Yet, the effect sizes yielded by the different approaches were not universally consistent: In the Hamburg sample ($N = 118$, 100% reinforcement rate, CS duration: 6–8 s), 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 through 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 data set. In the smaller Mainz sample ($N = 38$, 50% reinforcement rate, CS duration: 4.5 s), however, most approaches yielded consistent effect sizes even though it has to be noted that the CrIs 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 data sets investigated here, only a few SCR quantification approaches yielded comparable effect sizes in both data sets, despite numeric differences 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 5 s post-CS onset) and BLC (i.e., full CS duration which corresponds to 0–6 s in the Hamburg and 0–4.5 s post-CS onset in the Mainz sample). The trough of the response, however, is defined very differently (i.e., BLC: average SCL 2 s prior to CS onset; TTP: onset in an OLW of 0.9–4.5 s post-CS onset). This group-level comparability between both approaches is 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 which 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 approaches and the resulting problems (Sjouwerman et al., 2021). Note that the work

by Sjouwerman et al. (2021) is complementary to the work presented here. While we here investigate whether seven different SCR response quantification approaches result in convergent results (i.e., comparison *between* different approaches), our related work focuses on *within-approach heterogeneity* in parameter specification (e.g., time windows) in one of the approaches used here (i.e., the BLC approach).

Despite a number of major problems with the BLC approach discussed in depth in our related work (Sjouwerman et al., 2021), our results are reassuring that TTP and the representative BLC approach to SCR 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 quantification in the field of fear conditioning research, this is good news for the field even though we highlight stimulus (i.e., CS−)-specific reduced agreement.

4.2.4 | Different model-based approaches as implemented in PsPM

Furthermore, it is noteworthy that the four PsPM models yielded more consistent results not only in comparison with each other but also with any of the other approaches in the Mainz than the Hamburg data sets. 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) were with 1–3.5 s rather short. The CS duration of 4.5 s in the Mainz data set is closer to this than the 6–8 s duration in the Hamburg data set. It remains to be investigated systematically whether the model-based approaches in PsPM are optimized for shorter duration CSs, and short ITIs 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 modeling a flexible CS response onset may absorb SCR elicited by US or US omission leads to an 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 a 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 the results. Thus, data processing and model specification need to be reported in detail to ensure computational reproducibility, and 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 postprocessing and data analyses

Here, we have illustrated that different commonly used SCR 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− discrimination may also have implications for the application of commonly used postprocessing or data-cleaning procedures such as minimal response criteria as well as the identification of performance-based exclusion of SCR nonresponder and SCR nonlearner (for a critical evaluation and discussion, see 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 (for an empirical investigation, see Sjouwerman et al., 2021). Further, it is clear from the average CS+, CS-, and CS discrimination values (see Figure 2) yielded by the different response quantification approaches that identical cutoffs for nonlearning 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 it realistic to assume the existence of a single and universally best approach for SCR quantification?

It has been proposed that we may identify the "best" approach for SCR quantification by means of "retrodictive validity," formerly referred to as "predictive validity" (Bach et al., 2020; Bach & Melinscak, 2020). More precisely, it has been proposed that the method with the highest retrodictive validity is the method that has the highest chance of recovering an unobservable (psychological) process from skin conductance data. 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 the retrodictive validity idea, 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 or the largest effect size. Even though an in-depth discussion on the retrodictive validity idea is beyond the scope of this work, we would like to note that an exclusive focus on effect size falls short of appreciating measurement precision as an important criterion.

When interpreting the results of our work in a "retrodictive validity framework," there is no evidence for a single, universally superior approach. More precisely, our results from two different data sets differing primarily in the number of participants (118 vs. 38), reinforcement rate (100% vs. 50%), and CS duration (6–8 s vs. 4.5 s) reveal no single method that yields a consistently higher effect size compared with other methods in both data sets.

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 may depend 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 (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 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 as universally superior. This is particularly important as the authors note that the toolbox PsPM has "been evaluated only in limited experimental circumstances and by a small group of researchers" (cf. Bach & Melinscak, 2020). We echo their call for more methodological research in order to establish "a clearer picture on what the best measurement approach is in different research scenarios" (cf. Bach & Melinscak, 2020) and with the present work provide the first step into this direction.

4.5 | Limitations

Here, we compare seven different SCR quantification approaches as identified through a literature review. Yet, the "full" multiverse of possible SCR processing steps includes a number of additional steps not considered here in-depth such as transformations (see also Supplementary Material), cutoff criteria (Lonsdorf, Klingelhöfer-Jens, et al., 2019), data exclusion (Lonsdorf, Klingelhöfer-Jens, et al., 2019), and filtering (see, e.g., Privratsky et al., 2020). Aiming to cover all potentially relevant decision nodes is infinite and a focus on "a" multiverse rather than "the" multiverse still provides valuable information. This can help to deflate the multiverse and leaves only the relevant specifications (i.e., those that have not been shown to be clearly inferior in the more focused investigations) for the construction of a larger and more comprehensive multiverse. Future work may systematically focus on these additional decision nodes or cover different parts of the full data multiverse systematically (see Sjouwerman et al., 2021 for a multiverse focusing on within-approach heterogeneity in the BLC method).

SCRs were relatively larger in the Mainz compared with 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" compared with "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 data sets 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 and differences in hardware could also account for these differences.

Furthermore, our literature search covered only a limited time frame (6 months in 2019) and hence the results may not be fully representative. Yet, a different literature search (Privratsky et al., 2020; full details provided in the supplementary material) covering more than 90 articles in the field of fear conditioning from 2019 and 2020 shows a similar picture with BLC and TTP being most abundantly used (subsumed as "peak scoring" by the authors, which is a problematic term, however) and with substantially fewer studies using Ledalab, few studies using PsPM, or other approaches (e.g., "area under the curve", cvxEDA). Even though the literature search provided here served primarily as a tool to guide the selection of the to-be included SCR quantification approaches, the results by Privratsky are reassuring the frequencies reported here are representative despite the short time window.

Finally, our comparison of different SCR quantification approaches across two data sets 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 | Prospects and challenges of a multiverse-type of approach

Multiverse-type of approaches (Del Giudice & Gangestad, 2021; Simonsohn et al., 2020; Steegen et al., 2016) have recently gained momentum in the field of psychophysiology—for instance, in research using EEG (Clayson et al., 2021; Kołodziej et al., 2021; Nikolin et al., 2022; Sandre et al., 2020; Wacker, 2017) or in fear conditioning research with a focus on SCRs (Lonsdorf et al., 2021; Lonsdorf, Klingelhöfer-Jens, et al., 2019; Lonsdorf, Merz, & Fullana, 2019; Sjouwerman et al., 2021). Multiverse-type of approaches can be considered an attempt to empirically optimize processing pipelines and an intermediate step toward more standardization in fields that are characterized by substantial heterogeneity in data (recording) and processing steps. More precisely, multiverse-type of analyses examine the impact of a (large) set of different equally justifiable methodological decisions on the robustness of an effect of interest. By empirically identifying and subsequently deprioritizing unsuitable paths, they can help to deflate the multiverse of *possible (equally justifiable)* data analysis paths. The most critical step in setting up a multiverse-type of analysis is the selection of the to-be included decision nodes and their specifications. Specifically, it is inherently challenging to define which methodological decisions can be considered "equally justifiable" (for discussions, see Del Giudice & Gangestad, 2021; Lonsdorf et al., 2021) in particular in light of often underspecified theories in psychology that leave much room for different definitions and hence operationalization of (latent) constructs (discussed for fear conditioning research in Lonsdorf et al., 2021; Lonsdorf, Merz, & Fullana, 2019). In addition, it is important to note that not all equally justifiable paths necessarily belong to the (exact) same multiverse. For instance, a statistical model with an included covariate tests a different underlying hypothesis than a model without that covariate and, hence, is—in a strict sense—not part of the same (model) multiverse (Del Giudice & Gangestad, 2021). Along the same lines, it may also be debatable whether model-based approaches and TTP/BLC approaches belong to the same multiverse as they may measure different constructs (e.g, estimated sudomotor nerve activity vs. observable physiological response, respectively). As these approaches are, however, used interchangeably in the literature, we combined them in the same multiverse here. We have chosen to depart from a systematic literature search as a means to objectively decide on the to-be included paths by defining "equally justifiable" as approaches that are used interchangeably in the literature. Other approaches that have been used to guide the decision on which specifications to include are based on expert agreement (Wacker, 2017) and/or multiple analyst approaches (Silberzahn et al., 2018). An advantage of our approach is that the different quantification approaches included mirror the actual multitude of decisions a researcher is presently faced in the field when aiming to quantify SCRs. Hence, our approach provides empirical evidence whether it can indeed be considered justifiable to use the different included approaches interchangeably in the field.

4.7 | Summary and outlook

Our results illustrate heterogeneity in the exact specification and implementation of SCR response quantification approaches derived from a systematic literature search and a thorough summary of the available comparative studies. Empirically, we illustrate partly inconsistent outcomes for effects sizes of CS discrimination when applying seven different SCR quantification approaches to the same data. Our results challenge the existence of a universally best or superior SCR quantification approach and call for more and systematic comparative (multiverse-type of) studies focusing on different decision nodes during data processing but also on different experimental design

specifications which, however, requires specifically tailed experimental designs. Finally, we call for more consideration to measurement and reliability questions and for more systematic and collaborative efforts to solve these challenges as a research field and work toward more exchange, more homogenization in research methods, as well as detailed reporting.

ACKNOWLEDGMENTS

The authors thank Karita Ojala for their 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 the 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, and 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. Open Access funding enabled and organized by Projekt DEAL.

AUTHOR CONTRIBUTIONS

Manuel Kuhn: Conceptualization; data curation; formal analysis; investigation; methodology; software; visualization; writing – original draft. **Anna Gerlicher:** Conceptualization; funding acquisition; methodology; project administration; resources; supervision; visualization; writing – original draft. **Tina B Lonsdorf:** Conceptualization; funding acquisition; methodology; project administration; supervision; visualization; writing – original draft.

CONFLICT OF INTEREST

The authors do not report any conflict of interest.

DATA AVAILABILITY STATEMENT

Data and code are available on the OSF https://osf.io/ft86v/.

ORCID

Anna M. V. Gerlicher https://orcid.org/0000-0001-7364-9845
Tina B. Lonsdorf https://orcid.org/0000-0003-1501-4846

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

Clayson, P. E., Baldwin, S. A., Rocha, H. A., & Larson, M. J. (2021). The data-processing multiverse of event-related potentials (ERPs): A roadmap for the optimization and standardization of ERP processing and reduction pipelines. *NeuroImage*, *245*, 118712. https://doi.org/10.1016/j.neuroimage.2021.118712

Dawson, M. E., Schell, A. M., Filion, D. L., & Berntson, G. G. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed.). Cambridge University Press. https://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

Dunsmoor J. E., Mitroff S. R., & LaBar K. S. (2009). Generalization of conditioned fear along a dimension of increasing fear intensity. *Learning & Memory*, *16*(7), 460–469. http://dx.doi.org/10.1101/lm.1431609

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

Fredrikson, M., Annas, P., Georgiades, A., Hursti, T., & Tersman, Z. (1993). Internal consistency and temporal stability of classically conditioned skin conductance responses. *Biological Psychology*, *35*(2), 153–163.

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*, *91*(3), 186–193. https://doi.org/10.1016/j.ijpsycho.2013.10.015

Huff N. C., Hernandez J. A., Blanding N. Q., & LaBar K. S. (2009). Delayed extinction attenuates conditioned fear renewal and spontaneous recovery in humans. *Behavioral Neuroscience*, *123*(4), 834–843. http://dx.doi.org/10.1037/a0016511

Kołodziej, A., Magnuski, M., Ruban, A., & Brzezicka, A. (2021). No relationship between frontal alpha asymmetry and depressive disorders in a multiverse analysis of five studies. *eLife*, *10*, e60595. https://doi.org/10.7554/eLife.60595

Kragel P. A., & LaBar K. S. (2013). Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions.. *Emotion*, *13*(4), 681–690. http://dx.doi.org/10.1037/a0031820

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

Kuhn, M., Gerlicher, A. M. V., & Lonsdorf, T. B. (2022). Navigating the manyverse of skin conductance response quantification approaches – a direct comparison of Trough-to-Peak, Baseline-correction and model-based approaches in Ledalab and PsP. *Psychophysiology*. https://doi.org/10.1111/psyp.14058

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*, *1*(3), 389–402. 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*, *25*(2), 97–109. https://doi.org/10.1016/s0167-8760(96)00713-1

Lipp, O. V. (2006). Human fear learning: Contemporary procedures and measurement. In M. G. Craske, D. Hermans, & D. Vansteenwegen (Eds.), *Fear and learning: From basic processes to clinical implications* (pp. 37–51). American Psychological Association. https://doi.org/10.1037/11474-002

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. (2022). Multiverse analyses in fear conditioning research. *Behaviour Research and Therapy*, 104072. http://dx.doi.org/10.1016/j.brat.2022.104072

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

Morey, R. D., & Rouder, J. N. (2015). BayesFactor 0.9.11-1. Comprehensive R Archive Network.

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

Nikolin, S., Chand, N., Martin, D., Rushby, J., Loo, C. K., & Boonstra, T. W. (2022). Little evidence for a reduced late positive potential to unpleasant stimuli in major depressive disorder. *Neuroimage: Reports*, *2*(1), 100077. https://doi.org/10.1016/j.ynirp.2022.100077

Ohman, A. (1972). Factor analytically derived components of orienting, defensive, and conditioned behavior in electrodermal conditioning. *Psychophysiology*, *9*(2), 199–209. https://doi.org/10.1111/j.1469-8986.1972.tb00754.x

Ojala K. E., & Bach D. R. (2020). Measuring learning in human classical threat conditioning: Translational, cognitive and methodological considerations. *Neuroscience & Biobehavioral Reviews*, *114*, 96–112. http://dx.doi.org/10.1016/j.neubiorev.2020.04.019

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

Prokasy, W. F., & Ebel, H. C. (1967). Three components of the classically conditioned GSR in human subjects. *Journal of Experimental Psychology*, *73*(2), 247–256. https://doi.org/10.1037/h0024108

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Sjouwerman, R., Illius, S., Kuhn, M., & Lonsdorf, T. (2021). A data multiverse analysis investigating non-model based SCR quantification approaches. *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

Stewart, M. A., Stern, J. A., Winokur, G., & Fredman, S. (1961). An analysis of GSR conditioning. *Psychological Review*, *68*(1), 60–67. https://doi.org/10.1037/h0048816

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

Thomas L. A., & LaBar K. S. (2008). Fear relevancy, strategy use, and probabilistic learning of cue-outcome associations. *Learning & Memory*, *15*(10), 777–784. http://dx.doi.org/10.1101/lm.1048808

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 I. Martin & P. H. Venables (Eds.), *Thechniques in psychophysiology*. Wiley.

Vogel S., Klumpers F., Kroes M. C.W., Oplaat K. T., Krugers H. J., Oitzl M. S., Joëls M., & Fernández G. (2015). A Stress-Induced Shift From Trace to Delay Conditioning Depends on the Mineralocorticoid Receptor. *Biological Psychiatry*, *78*(12), 830–839. http://dx.doi.org/10.1016/j.biopsych.2015.02.014

Wacker J. (2017). Increasing the reproducibility of science through close cooperation and forking path analysis. *Frontiers in Psychology*, *8*, 1–4. http://dx.doi.org/10.3389/fpsyg.2017.01332

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

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

FIGURE S1 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)

FIGURE S2 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)

FIGURE S3 Average CS discrimination (±standard errors) 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)

FIGURE S4 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)

TABLE S1 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

How to cite this article: Kuhn, M., Gerlicher, A. M. & Lonsdorf, T. B. (2022). Navigating the manyverse of skin conductance response quantification approaches – A direct comparison of trough-to-peak, baseline correction, and model-based approaches in Ledalab and PsPM. *Psychophysiology*, *59*, e14058. https://doi.org/10.1111/psyp.14058