**Title:** Filtering and model-based analysis independently improve skin-conductance response measures in the fMRI environment: validation in a sample of women with PTSD

**Authors:** Anthony A. Privratsky[a], Keith A. Bush[a], Dominik R. Bach[b, c], Emily M. Hahn[d], Josh M. Cisler[e]

**Affiliations:** [a]Brain Imaging Research Center, Department of Psychiatry, University of Arkansas for Medical Sciences, Little Rock, AR USA 72205
[b]Wellcome Centre for Human Neuroimaging and Max-Planck UCL Centre for Computational Psychiatry and Ageing, 12 Queen Square, University College London, London WC1N 3BG, United Kingdom
[c]Computational Psychiatry Research, Department of Psychiatry, Psychotherapy, and Psychosomatics, Psychiatric Hospital, University of Zurich, Lenggstrasse 31, 8032 Zurich, Switzerland
[d]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, 149 Thirteenth St, Charlestown, MA USA 02129
[e]Department of Psychiatry, University of Wisconsin-Madison, Madison, WI USA 53726

**Funding sources:** NIMH R21MH108753 (Cisler, all data)
NIDA T32DA022981-09 (Privratsky)
NSF BCS-1735820 (Bush)
ERC-2018 CoG-816564 (DRB)
E.M.H. does not have funding to disclose.

**Acknowledgements:** DRB is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. ERC-2018 CoG-816564 ActionContraThreat).

**Declarations of interest:** none

**Corresponding Author/Permanent Address:**
Anthony A. Privratsky
Brain Imaging Research Center
Department of Psychiatry
University of Arkansas for Medical Sciences
4301 W. Markham Street, Little Rock, AR 72205
Slot #554
Email: aaprivratsky@uams.edu
Phone: (832) 656-9485

**ABSTRACT**

Numerous methods exist for the pre-processing and analysis of skin-conductance response (SCR) data, but there is incomplete consensus on suitability and implementation, particularly with regard to signal filtering in conventional peak score (PS) analysis. This is particularly relevant when SCRs are measured during fMRI, which introduces additional noise and signal variability. Using SCR-fMRI data (n = 65 women) from a fear conditioning experiment, we compare the impact of three nested data processing methods on analysis using conventional PS as well as psychophysiological modeling. To evaluate the different methods, we quantify effect size to recover a benchmark contrast of interest, namely, discriminating SCR magnitude to a conditioned stimulus (CS+) relative to a CS not followed by reinforcement (CS-). Findings suggest that low-pass filtering reduces PS sensitivity ($\Delta d$ = -20%), while band-pass filtering improves PS sensitivity ($\Delta d$ = +27%). We also replicate previous findings that a psychophysiological modeling approach yields superior sensitivity to detect contrasts of interest than even the most sensitive PS method ($\Delta d$ = +110%). Furthermore, we present preliminary evidence that filtering differences may account for a portion of exclusions made on commonly applied metrics, such as below zero discrimination. Despite some limitations of our sample and experimental design, it appears that SCR processing pipelines that include band-pass filtering, ideally with model-based SCR quantification, may increase the validity of SCR response measures, maximize research productivity, and decrease sampling bias by reducing data exclusion.

**Key words:** *electrophysiology; skin-conductance; SCR; GSR; EDA; fMRI*

## 1. INTRODUCTION

The skin-conductance response (SCR) has remained a commonly used index of sympathetic arousal in psychophysiological research for more than 100 years due to its accessibility, objectivity, and ability to be continuously and passively collected while subjects engage in experimental procedures (Boucsein, 2012). This appeal has persisted despite that the SCR signal is prone to many sources of interference and while it has moved into increasingly noise-prone environments. Recording SCRs in the fMRI environment is particularly challenging. In addition to introducing artifacts by direct electromagnetic induction of recording cables, the scanner bore significantly restricts optimal positioning of electrodes, electrode leads, and participants, all of which may increase the likelihood of noise artifacts (Gray et al., 2009; Lagopoulos et al., 2005). Along with the normal, tonic variation in the skin-conductance level (SCL) throughout experiments, these factors make valid data pre-processing, analysis, and interpretation challenging.

To underline these considerations, we performed a systematic review of 95 fear conditioning articles (containing 100 experimental reports) utilizing SCR between 2019-2020, consistent with PRISMA guidelines (Moher et al., 2015) and optimal database coverage recommendations (Bramer et al., 2017) (see Supplementary Material for search terms, inclusion criteria, and data). We found that poor data quality or data processing challenges were cited as sources of lost or excluded data in 18% of all reports, and in 27% of reports utilizing concurrent fMRI (n = 22). Adding exclusions of "non-responders" based on quantitative metrics that are potentially related to signal quality (e.g. differential SCRs or minimum thresholds), these numbers rose to 61% for all experiments and 68% for concurrent fMRI-SCR. Adding exclusions for inconsistently defined technical issues, these numbers both rose to 73%. Because different methods of participant exclusions in fear conditioning research can lead to different conclusions from a given data set (Lonsdorf and Merz, 2017; Lonsdorf et al., 2019, 2019), such high exclusion rates call for more optimal pre-processing methods that allow retaining more participants.

Electrodermal recording hardware manufacturers, psychophysiology manuals, and consensus guidelines (Boucsein et al., 2012; Braithwaite et al., 2013; Figner et al., 2011) suggest both online and offline filtering methods to avoid or remove SCR artifacts. However, to our knowledge, the implementation, effect, and importance of these methods have not been quantitatively compared in peer-reviewed literature, nor has their appropriateness to SCR collected in the fMRI environment been tested. Prior work suggested that in a psychophysiological modeling (PsPM) approach, a combination of high- and low-pass filtering yields the most sensitive response measures (Bach et al., 2013). This approach is based on a general linear convolution model (GLCM) to quantify magnitude of phasic SCRs (Bach et al., 2009). However, the explicit effects of filtering on peak-score (PS) analysis, a far more common phasic SCR quantification method, have not been quantitatively reported. Indeed, despite existing filtering recommendations (e.g. Boucsein et al., 2012), our literature review suggests that the implementation of filtering strategies for data quality control is not standard practice. Just over half (56%) of SCR-utilizing studies reported any signal filtering; just 25% reported high-pass filtering to remove low-frequency artifacts. We speculate that these omissions may be due to lack of awareness of suitable filtering methods, to the belief that filtering is not necessary to derive accurate phasic SCR measurements, or to concerns that filtering may affect signal of interest. Popular methods of filtering are optionally implemented via hardware (BIOPAC, USA; Spike2, CED, UK) or software (PsPM (Bach et al., 2018); Ledalab (Benedek and Kaernbach, 2010); AcqKnowledge [BIOPAC, USA]) packages; the need to assess the appropriateness of filtering and its effect on data is therefore largely left to individual investigators. This problem is highlighted by our finding that of 61 experimental reports that cited difficulty with data quality control or that excluded participants for poor data (i.e. qualitative or quantitative signal metrics), only 54% reported any filtering methodology and only 23% used high-pass or equivalent. In some studies, researchers omitted analysis of collected SCR data due to processing or noise concerns despite its planned integral role in hypothesis testing.

Among those studies in our review that implemented filtering or similar artifact suppression methods (e.g. smoothing), there is currently no standard practice. Of the 56 studies reporting any type of filtering, researchers report low-pass filter parameters that differ by 3 orders of magnitude (e.g. 0.1 Hz, 5 Hz, and 200 Hz), and 35 did not fully describe the filter type and parameters used. Moreover, of the remaining 44 studies that did not implement filtering, 16 (36%) cited visual inspection, manual artifact removal, and/or interpolation as the sole method/s for quality assessment and data exclusion, while the remaining 28 (64%) cited either no quality control or only minimum response thresholding. These methods are inherently subjective, and risk biasing samples by excluding subjects with greater noise artifacts. As many fMRI-SCR artifacts are likely due to participant motion, this biasing may particularly impact psychological research, as subject motion has been linked to neural genotypes and psychopathology (Couvy-Duchesne et al., 2016; Engelhardt et al., 2017). The ability to remove noise from SCR data using a systematized, automated, and validated methodology is therefore warranted.

Considering the aforementioned inconsistencies, we sought to explicitly demonstrate the effect of three nested SCR processing methods on the quality of response measures extracted from the same dataset. These methods were specifically chosen to span the full range of filtering intensities, namely, no filtering, low-pass filtering to remove only high-frequency artifacts, and low- and high-pass filtering to remove both high and low frequency artifacts. To quantitatively evaluate each measurement method, we assessed the sensitivity of SCR measures to recover an experimentally induced psychological variable (i.e. the discrimination of aversive associations). This approach has previously been termed "predictive validity" (Bach and Friston, 2013), or perhaps more appropriately, “retrodictive validity” (Bach et al., 2020) and allows an evaluation of how well the CS-US association can be measured with a given method. It has previously been used to assess and compare the precision of psychophysiological methods and models (Bach, 2014; Bach et al., 2013; Staib et al., 2015). Here, we quantify retrodictive validity in two ways: (1) by the effect size of the contrast (Cohen, 1960), and (2) using Akaike Information Criteria (Akaike, 1998), which additionally allows a statement on whether retrodictive validity of two methods is decisively different. We hypothesized that response measures derived from unprocessed SCR data would yield lower effect sizes in contrasts of interest, whether those measures are derived via psychophysiological modelling or the traditional PS approach. Because of the widespread utility of SCR measures in psychophysiological research, the results of these tests could impact research in a wide variety of areas.

As there are also discrepant methods and stages for inter-subject standardization of SCRs for both PS and PSPM analysis, we further tested the effects of all processing methods on contrasts from data standardized both before and after response calculation with a variety of previously reported rescaling methods, namely, rescaling (raw data, response measures, or average response measures) by a maximum or standard deviation (z-scoring).

An additional source of inconsistency in PS data analysis arises from the use of disparate methods for calculating peak-scores. Specifically, a variety of post-stimulus windows are used to constrain identification of peak responses for PS calculations. In the PS studies reviewed, 1-4 s post-stimulus onset is the most commonly used window for finding peak maxima, although this window is commonly as short as 1-3 or 1-3.5 s and as long as 1-8 s. Given that several guidelines for SCR data processing and analysis have noted that tonic signal drift as well as high-frequency artifacts may reduce accurate SCR measurement (Boucsein et al., 2012; Braithwaite et al., 2013), we hypothesize that these differences in peak-score estimation will lead to systematic differences in response measures for two reasons. Firstly, short windows may miss maxima while long windows may capture CS-unrelated responses or artifacts (Braithwaite et al., 2013; Lim et al., 1997). Secondly, and relatedly, these windows should display unequal susceptibility to nonlinear drift. For example, previous reports have suggested that drift (typically observed in a downward direction over an experiment) may lower amplitude estimates (Braithwaite et al., 2013), therefore a 1-3 s window consistently missing maxima in unfiltered data will underestimate SCRs to a greater extent than the same window in filtered data. Similarly, as underlying tonic drift will skew PSs in proportion to the time required from onset to peak latency, accurately identified peaks within a

4 second window in unfiltered data should be proportionally more accurate than those accurately identified in a 5 second window in unfiltered data. Stated differently, filtering should improve measurement of SCRs with a 5 second peak latency more greatly than those with a 4 second peak latency. We therefore tested the effect of filtering on peak-scores calculated with a variety of post-stimulus ranges.

In an additional exploratory analysis, we investigated how these different pre-processing and analysis strategies impact on "non-responder" or "non-learner" exclusions, as performed in many previous studies. Although our literature review cannot yield conclusive evidence on this point, particularly given reporting inconsistencies, we note that "non-responder" exclusions due to quantitative signal metrics were more common in studies that did not include filtering (10% or N = 276/2656 participants) than those did (5%, N = 235/4438). Exclusion criteria in our review were based on below zero or sub-threshold (e.g. 0.05 μS) discrimination (11 studies) or SCRs (as an average or proportion) below 0.01, 0.02, 0.03, 0.04, 0.05, and 0.1 μS. Given that filtering and rescaling may affect absolute scales of SC data, we analyzed whether filtering may reduce the number of exclusions based on the common within-subject criterion of negative discrimination (i.e. lower SCR for CS+ than CS-).

Hence, this manuscript intends to demonstrate the effect of different filtering and standardization methodologies on response measures and contrast effect sizes in conventional PS (at several post-stimulus scoring windows) as well as psychophysiological modeling based analyses. We hypothesize that response measures will vary with the presence of phasic and tonic distortions in SC signal, and that these distortions may be reduced by low- and high-pass filtering, respectively, resulting in enhanced signal recovery. Furthermore, we hypothesize that even in comparisons of the most effectively processed data, psychophysiological modeling will yield greater signal recovery than peak-scoring, in line with previous reports (Bach et al. 2010, Staib et al. 2015).

## 2. METHODS

### *2.1. Participants*

The final sample of 65 women (ages 21-50, *M* = 33.7, *SD* = 9) with a diagnosis of PTSD (Clinician Administered PTSD Scale V (Blake et al., 1995) participated in this research as part of a larger, two-day study designed to assess a novel medication for the treatment of PTSD. All analyses are derived from participant data on Day 1, before study medication administration, therefore this should not be considered a confounding factor. Data were acquired at two locations, the University of Arkansas for Medical Sciences (n=35) and the University of Wisconsin-Madison (n=30). Participant race was roughly representative of the average of the study regions (69% Caucasian, 23% African descent, 2% Hispanic, 2% Native American, and 4% mixed African/Caucasian). Exclusion criteria included internal ferromagnetic material, pregnancy, positive pre-scan drug screens, current substance use disorders, or any past or current psychotic disorders, as assessed with the SCID-IV-NP (Spitzer et al., 2002). After exclusion for these factors, 69 women began scanning procedures. We excluded four of these participants for aborted scans due to claustrophobia (n = 2), beginning the task in the incorrect sequence (n = 1), or for missing log files (n = 1). Participants were not excluded based on any qualitative or quantitative SCR metrics. We intentionally included 12 participants despite visually-identified SCR that had few identifiable signal changes (n = 6), biologically implausible noise structure (n = 4), and seemingly only SCRs to the first presentations of the unconditioned stimulus (n = 2). Medications in the final sample included oral contraceptives (n = 6; normally cycling n = 49; postmenopausal or oophorectomy n = 10), thyroid medications (n = 4), SSRIs, SNRIs, or TCAs (25), trazodone or prazosin (15), mood stabilizers (10), antipsychotics (8), stimulants (6), NDRIs (5), benzodiazepines (4), and beta-blockers (3). Lifetime diagnoses were positive for depression (40), generalized anxiety disorder (37), substance use disorder (31), alcohol dependence (23), bipolar disorder (14), and panic disorder (6).

### *2.2. Ethics statement*

All participants provided written informed consent after receiving written and verbal descriptions of the study procedures, risks, and benefits. Study procedures and informed consent documents were approved in accordance with Institutional Review Board policies at UAMS and UW-Madison.

### *2.3. Experimental design*

#### *2.3.1. Task*

Participants underwent a same-day fear conditioning and extinction task, consisting of pseudorandomly alternating, 3 s presentations of two geometric stimuli (CS+ and CS-) over a colored background with a jittered inter-trial interval (ITI) of 2-6 s (thus allowing a minimum of 5-9 seconds for SCRs post-stimulus onset). Acquisition and extinction contexts were signaled by background colors blue and yellow, respectively (n=47), or vice versa (n=18). CS+ and CS- were a triangle and circle (n=47), or vice versa (n=18). Following a baseline phase containing 6 presentations each of CS+ and CS- with no electrical stimulations, participants underwent acquisition (18 presentations each of CS+ and CS-), wherein electrical shocks (unconditioned stimulus, US) were administered 2.5 s after the onset of 50% of CS+ presentations. This resulted in 9 CS+ no shock (CS+ns) presentations. Participants then underwent extinction (18 presentations each of CS+ and CS-, with no US). In order to encourage consistent levels of attention, participants were instructed to use their right hand to press either of two buttons for each stimulus presentation to indicate the observed shape, as quickly and accurately as possible. Participants were told that the task was designed to study attention and distraction in PTSD, and that they would receive, in a double-blind, either a placebo or one of two dosages of L-DOPA following the experiment. Participants were not informed about stimulus contingencies, but only that they may periodically receive electrical stimulations. After baseline stimuli, and after each half of acquisition and extinction, participants were asked to rate the likelihood that the shock would occur with each shape on a 0-10 point scale (corresponding to 0-100% chance, n = 5 rating periods). In total, the task contained 98 events (84 CS, 9 US, and 5 rating periods).

#### *2.3.2. Electrical stimulations*

Shocks were administered via the BIOPAC STM100C module using pre-gelled electrodes (EL508) placed on the skin of the fleshy portion of the mediolateral, left lower leg, directly over the tibialis anterior muscle. Participants calibrated the intensity of the administered shock prior to the experiment by setting it to a level that was “as unpleasant as possible without being painful,” and were encouraged to aim for a level that corresponded to a 7 on a scale of 0-10 (0, no unpleasantness; 10, painful). Amperage on the stimulation adapter was pre-set to the maximum (50 mA) to allow the greatest range of intensity selections. Stimulations consisted of a 2 ms square-wave pulse. After making their intensity selection, participants rated the unpleasantness of their chosen setting on a scale of 0-10. Average participant ratings were 6.7 (UAMS, *SD* = 0.75) and 6.9 (UW, *SD* = 0.38) and did not differ between sites ($t(63) = -1.38$, $p = .17$).

### *2.4. Data Collection*

#### *2.4.1. SCR data acquisition*

At both sites, SCR data were acquired on a BIOPAC MP150 Data Acquisition System using the EDA100C module with MECMRI-TRANS cable system (BIOPAC, USA). Data were acquired directly into BIOPAC AcqKnowledge 4.3 software at 2000 Hz (Arkansas site) or 1000 Hz (Wisconsin site). No online hardware or software filtering was applied. Two SCR electrodes (EL509, isotonic recording GEL 101) were placed on the medial portions of the thenar and hypothenar eminences of the left hand for recording and one electrode was placed on the ventral surface of the left wrist for ground.

#### *2.4.2. fMRI scanning environments*

fMRI scanning occurred concurrently with SCR-recording during the fear conditioning task; we report scan parameters for the purpose of reporting noise variables which may contribute to artifacts. At the Arkansas site, fMRI data were acquired on a Philips Achieva 3T X-series scanner using a 32-channel headcoil. Echo planar imaging sequences were used to collect the functional images using the following sequence parameters: TR/TE/FA = 2000 ms/30 ms/90°, FOV = 240 × 240 mm, matrix = 80 × 80, 37 axial slices (parallel to AC–PC plane to minimize OFC signal artifact), slice thickness = 2.5 mm, and final resolution of 3 × 3 × 3 mm.

At the UW-Madison site, fMRI data were acquired on a GE MR750 3T scanner using an 8-channel headcoil. EPI sequences used to collect the functional images used the following parameters: TR/TE/FA = 2000ms/ 25 ms/ 60°, FOV = 24 cm, matrix = 64 x 64, 40 sagittal slices, slice thickness = 4 mm, original resolution was 4 x 3.75 x 3.75 mm.

*2.5. Data analysis*

We tested the effects of three nested processing methods for data quality control, (1) no filtering (downsampling only), (2) low-pass filtering at 5 Hz, and (3) low- and high-pass filtering (at 7 different cutoff frequencies), on two methods for deriving phasic SCR measures: (1) peak-scoring, and (2) psychophysiological modeling using a canonical SCR response function (GLCM-based regression). In the peak-scoring approach, we tested the effect of filtering on peak-scores calculated from four different lengths of post-stimulus maximum-response range (MR) windows (section 2.5.5). We additionally tested the effect of filtering on contrasts derived under different methods of subject-wise rescaling (section 2.5.4).

In order to evaluate the effect of these processing methods on experimental contrasts, we calculated subject-level contrasts for discrimination during fear conditioning (acquisition CS+ (no shock) – acquisition CS-) for each processing method, followed by group-level effect sizes for each contrast in each processing method. In order to assess statistically decisive differences in processing methods, we performed linear mixed-effects regressions (PsPM's pspm_predval.m function; where subject intercept effects and CS response estimates act as predictors and the known psychological state act as the outcome variable) and calculated Akaike Information Criterion (AIC) values for each contrast within each processing method according to previously described methods (Akaike, 1998; Bach et al., 2013) (further detailed in section 2.5.7). These processing methods resulted in 252 total permutations of peak-score derivation ([9 filtering methods] × [7 subject-wise rescaling methods] × [4 MRs]) and 45 permutations of GLCM parameter derivation ([9 filters] × [5 subject-wise rescaling methods]).

*2.5.1. Processing method 1 (PM1): no filtering*

To reflect the minimal processing performed in reports that we reviewed that do not report any temporal filtering or detrending strategy, we conducted analyses on raw data, downsampled to 10 Hz for computational efficiency. Downsampling to this frequency is generally considered to not affect SCR scoring, as SCRs are regarded to have a frequency of less than 0.5 Hz (Fahrenberg et al., 1983; Lim et al., 1997). Response measures (RMs) were then derived from these downsampled data. Though we did not implement anti-aliasing, this would not be expected to influence our processing method performance metric (effect size/retrodictive validity), as 1) stimuli presentation intervals are not regular, and 2) visible 60 Hz fluctuations (presumably from mains noise) exhibited amplitudes of just 0.0015 microsiemens (μS), or 0.009% of the mean raw SCR peak-score across our sample (0.17 μS).

*2.5.2. Processing method 2 (PM2): low-pass filtering*

To reflect filtering methodologies that remove only high frequency artifacts, we low-pass filtered raw, full-resolution SCR data using MATLAB (Mathworks, R2016b, USA). Consistent with prior reports (Staib et al., 2015), this consisted of (1) a 10 ms median filter followed by a (2) bidirectional first order, low-pass Butterworth filter (MATLAB functions butter.m and filter.m) with a cut-off frequency of 5 Hz. Notably, a low-pass filter frequency cutoff greater than 0.5-0.6 Hz is generally considered to preserve SCRs (Gerster et al., 2017; Lim et al., 1997). The median filter is effectively an additional low-pass filter, capable of removing high frequency outliers

without appreciably affecting signal on the timescales of the SCR waveform. Data were then downsampled to 10 Hz as in PM1. The Butterworth filter was applied bidirectionally using MATLAB's filtfilt.m function in order to offset the small forward shift in the time (~ 300 ms) of individual peaks induced by unidirectional filtering.

*2.5.3. Processing method 3 (PM3): low-pass and high-pass filtering*

Our final filtering methodology was performed according to a previously validated filtering methodology for SCR data in GLCM-based analyses (Bach et al., 2013). This filter consisted of the addition of a high-pass filter to the methodology of PM2. This resulted in a (1) 10 ms median filter followed by (2) a bidirectional, first-order Butterworth filter using a high-pass with variable cutoff frequencies and low-pass of 5 Hz. The range of high-pass cutoff frequencies tested were increasingly conservative (0.01, 0.0159, 0.02, 0.03, 0.05, 0.07, and 0.10 Hz), spanning the range of cutoffs used in reviewed studies. Data were then down-sampled to 10 Hz.

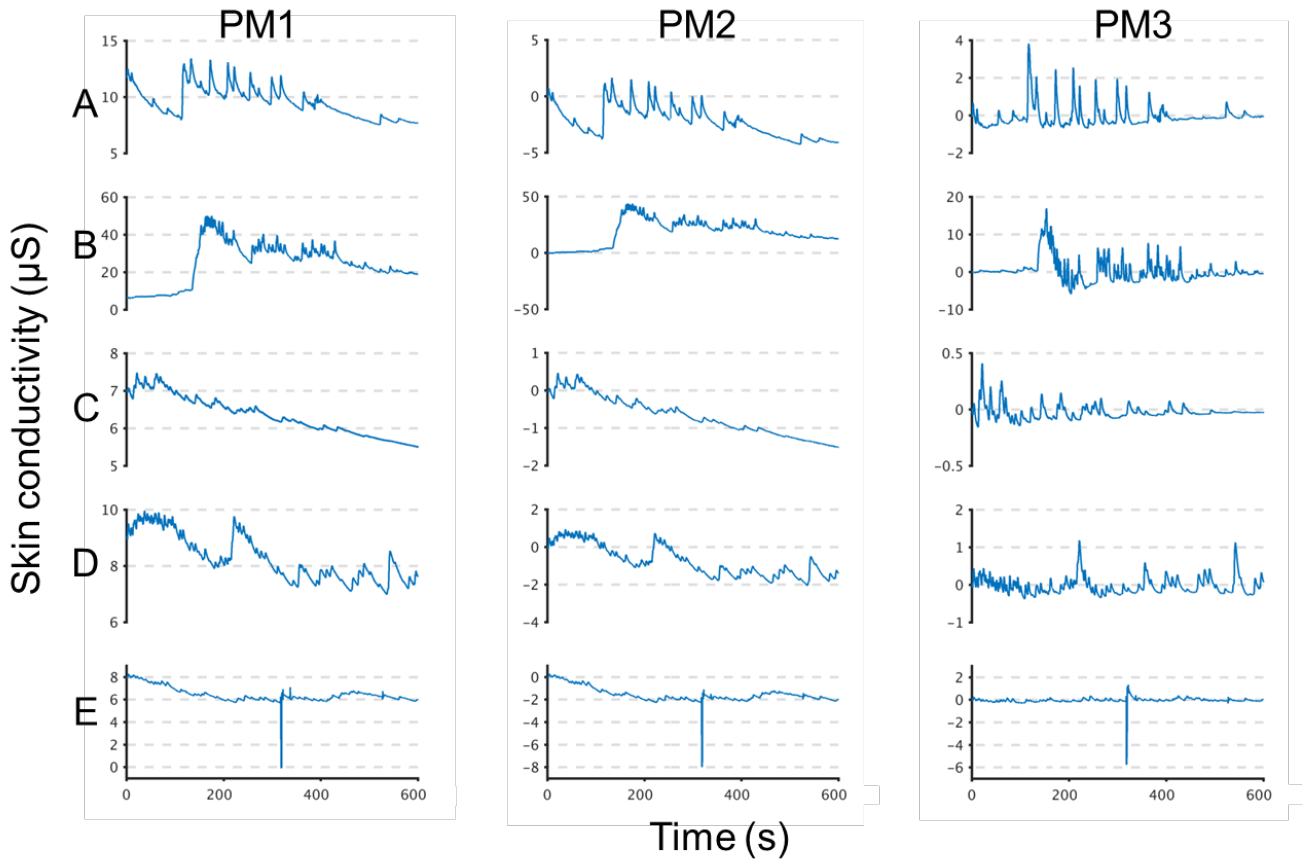



**Figure 1. Illustration of processing method effects.** Five subject datasets (A-E) were chosen for exemplary noise properties. Subjects B-E had particularly poor data quality. PM = processing method. PM3 high-pass cutoff frequency = 0.02 Hz. µS = microsiemens. s = seconds.

*2.5.4. Subject data rescaling*

It is widely considered necessary to remove inter-subject variance in baseline SCR properties before performing group analyses (Lykken and Venables, 1971). Rescaling may be performed directly on processed (continuous) SCR data (thousands of data points), on the means of response measures (derived from processed data) from each stimulus category (8 data points), or on all response measures derived from processed data (98 data points). In all cases, we performed rescaling post-processing, however it is unclear which stage and metric is best for rescaling. We therefore derived contrast effect sizes from: (1) unscaled data, (2) z-scored continuous data, (3) continuous data scaled by the maximum recorded value (Ben-Shakhar, 1985; Lykken and Venables,

1971), (4) z-scored mean responses by stimulus category, (5) mean responses scaled by the maximum mean response, (6) all responses z-scored and (7) all responses scaled to the highest response.

*2.5.5. Response measure method 1: Peak-scoring (PS)*

To quantify SCR amplitudes, we identified the maximum SCR value occurring within variable maximum-response ranges (MRs), 1-3.5 s, 1-4 s, 1-4.5 s, and 1-5 s after each stimulus onset. We subtracted the minimum value identified in the 0-2 s pre-stimulus data (baseline) from this maximum to derive a peak score.

*2.5.6. Response measure method 2: General linear convolutional model-based regression (GLCM)*

We implemented an experimentally validated GLCM of the phasic skin-conductance response (Bach et al., 2009; Bach et al., 2010; Bach et al., 2013; Bach, 2014; Gerster et al., 2017). This analytic model is performed similarly to the standard convolutional HRF linear regression model in fMRI BOLD signal analyses, and has been shown to be more sensitive than both peak-scoring and other model-based approaches in several contrasts (Bach, 2014). The use of this model, which assumes (almost) constant neural burst latency after a CS, was motivated by the short CS-US interval of 2.5 s which is considerably shorter than the 3.5 s that were used in the context of an alternative PsPM in which the neural burst latency is explicitly estimated from the data (Bach et al., 2010; Staib et al., 2015). Design matrices were created in MATLAB by convolving the canonical SCR function (SCRF; PsPM (Bach et al., 2010)) with the stimulus design matrix to create expected phasic SCR. We tested the effect of processing on response measures derived from 3 experimental design variants: (1) condition-wise multiple regression (1 regressor per stimulus type; 1 multiple regression model per subject), (2) trial-wise multiple regression with all regressors in one model (1 regressor per stimulus and 97 nuisance regressors for the remaining stimuli; 1 multiple regression model per subject), and (3) trial-wise multiple regression with separate models for each trial (1 regressor per stimulus and 1 nuisance regressor for the remaining 97 stimuli; 98 multiple regression models per subject). The latter two approaches, termed least-squares all (LSA) and least squares separate (LSS)(Mumford et al., 2012), are designed to reduce the effects of collinearity between neighboring responses in experiments with short ITIs. Unless otherwise stated, results in the main text are from the condition-wise approach; results from the other designs are available in the Supplementary Materials. For all GLCM methods, 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).

2.5.7. *Evaluation of pre-processing effects on signal sensitivity*

As described above, we assessed the validity of measures derived from each pre-processing method, for PSs and GLCMs, based upon their reliability to detect established within-subjects effects, namely, discrimination (greater group SCRs to conditioned stimuli (acquisition CS+ns) versus neutral stimuli (acquisition CS-)). Discrimination is a well-established subject effect in PTSD and women, despite an overall reduction compared to healthy controls, primarily driven by greater responses to CS- (Duits et al., 2015; Lonsdorf et al., 2015). Prior work suggests that band-pass filtered data yields the greatest sensitivity for SCRs in GLCM analyses (Bach et al., 2013), however we nonetheless sought to independently reproduce this finding in GLCMs, extend it to PSs, and to additionally demonstrate how filtering affects common metrics of signal quality for both methods, such as contrast effect size. Effect sizes (Cohen's *d)* were calculated by dividing the group mean difference by the standard deviation of subject differences. To test for decisive differences in the ability to detect these contrasts between methods (and hence for significant differences in filter quality), we modeled the ability of participant responses to each stimulus condition to predict dummy coded measures (i.e. for discrimination, CS+ns = 1, CS-acquisition = 0). The theoretical basis for this approach has been described at length (Bach and Friston, 2013; Bach et al., 2018). Importantly, the advantage of this approach over comparing effect sizes alone is that it allows a statement on whether the effect sizes/predictive validities of two measures are decisively different. Thus, the residual sum of squares from these regression models was used to calculate Akaike Information Criterion (AIC)

scores for direct comparison of methods' propensities for psychophysiological state estimation according to the following formula:

$$AIC = n * \log\left(\frac{RSS}{n}\right) + 2k$$

where *n* is the number of observations, and *k* is the number of free model parameters (which is the same for all models here such that AIC differences equal NLL or BIC differences). Lower AIC values indicate better model fit, where |ΔAIC| ≥ 6 between any pair corresponds to strong support for the lower AIC model (similar to $p < .05$ in frequentist statistics).

*2.5.8. Exploratory evaluation of filter effects on a common, quantitative metric for participant exclusion*

As filtering and rescaling may alter the absolute scale of raw SCR data (typically µS), we chose to examine the effect of filtering on a quantitative, within-subject measure commonly used to exclude participants (Lonsdorf et al., 2019). Namely, we examined how variable levels of processing affected the exclusion of participants for discrimination values below zero (SCRs to CS+ns < CS-). Furthermore, as rescaling does not affect the direction of discrimination contrasts within subjects, we performed these analyses only on unscaled PSs. We did not perform these analyses on GLCM contrasts, as exclusions based on GLCM parameters are not common practice.

**3. RESULTS**

*3.1. Filter effects on PS response measures*

*3.1.1. Mean peak-scores*

Filtering method significantly affected mean peak-score distributions (Fig. 2; for convenience PM effects are shown for the highest retrodictive validity PS processing method). Mean PS derived from raw data (PM1) and low-pass filtered (PM2) data showed significant discrimination at $p < .05$, ($t(63) = 2.55, 2.03$; $p = 0.01, 0.046$), while those from PM3 data showed significance at $p < .005$ ($t(63) = 3.24$; $p = 0.002$).

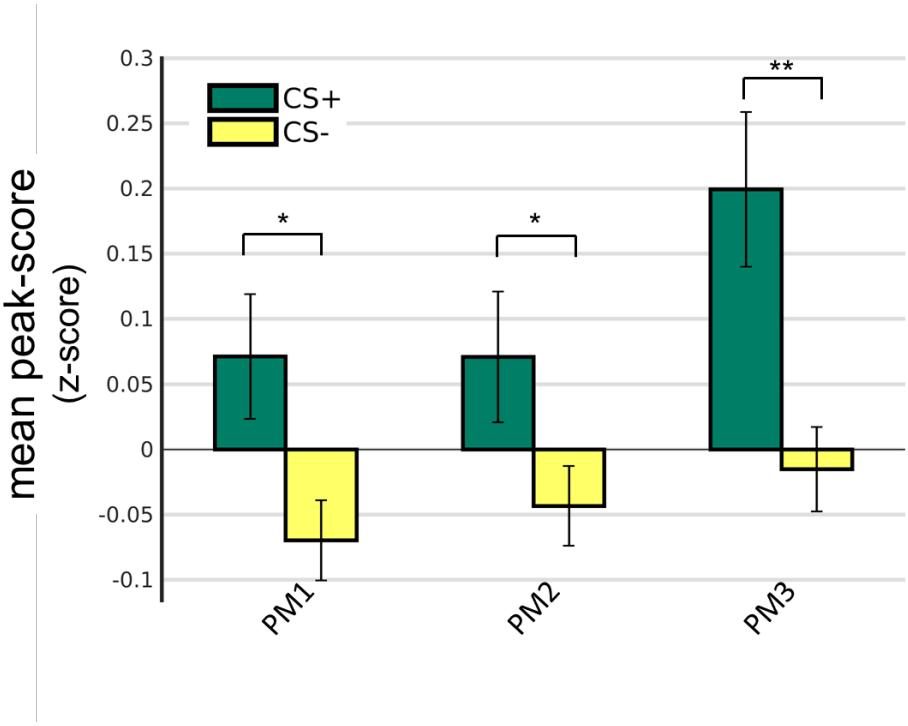



**Figure 2. Mean peak-scores by processing method.** For convenience, displayed data are those of only the highest effect size PS calculation window (maximum response range (MR) = 5 s), rescaling method (method 6: z-scoring all response measures) and high-pass filter (0.03 Hz). PM1 = processing method 1; PM2 = processing

method 2; PM3 = processing method 3 with 0.03 Hz high-pass filter. Displayed contrasts are paired t-tests for discrimination during fear conditioning (CS+ no shock – CS-). ns $p>0.05$, *$p<0.05$, **$p<0.005$.

*3.1.2. Discrimination contrast means and effect sizes (PS)*

Mean discrimination contrast values with their variability are shown in Fig. 3A and are translated to effect sizes in Fig 3B. Effect sizes increased with increasing MR and were maximized at MR = 5 s. At an MR = 5 s, PM2 decreased effect size relative to PM1 (Fig. 3B, PM1 $d$ = 0.3166, PM2 $d$ = 0.2521; Fig. 6A, ΔAIC = +4.52 (moderate support)). PM3 (high-pass filtering at cutoff of 0.03 Hz) increased effect sizes by 27% and 60% relative to PM1 and PM2 (Fig. 3B, MR = 5, PM3 $d$ = 0.4024; Fig. 6A, ΔAIC = -7.7, - 11.3, respectively).

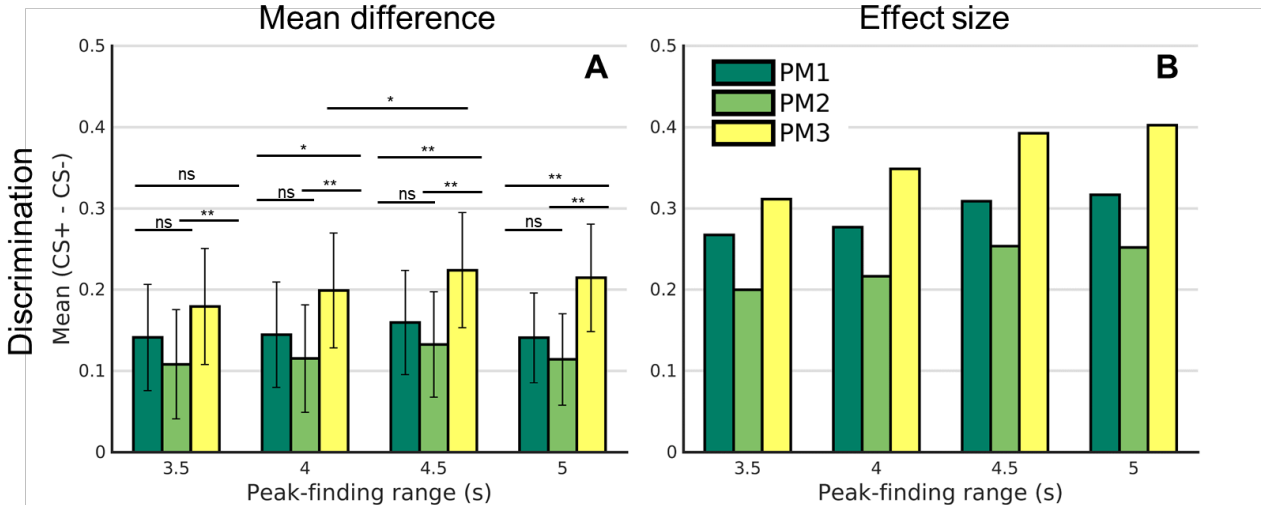



**Figure 3. Filter effects on peak-score (PS) discrimination means and effect sizes.** Low-pass filtering (PM2) had an insignificant effect on contrast means (panel A) and reduced signal recovery (panel B), while band-pass filtering (as in processing method 3 (PM3)) increased contrast means and improved signal recovery. Displayed mean contrasts and effect sizes are those resulting from the rescaling method and filter settings with the highest retrodictive validity for discrimination (rescaling method 6, z-scoring all response measures (RMs); high-pass filter cutoff (PM3) of 0.03 Hz; see Fig. 6A for all filter and rescaling method effects at MR = 5). Displayed statistical tests are 2-sample t-tests on group-mean contrast values between filtering methods. ns $p>0.05$, *$p<0.05$, **$p<0.005$.

*3.1.3. Summary effect of filtering on PS-based contrasts*

Low-pass filtering (PM2) reduced retrodictive validity for discrimination, the primary outcome contrast of interest (Fig. 6A, Supplementary Fig. 1). For all rescaling methods and across all MRs, PM3 band-pass filtering at a 0.03 Hz high-pass cutoff maximized retrodictive validity (Supplementary Fig. 1). Rescaling *all* PSs (i.e. prior to averaging within stimulus type) maximized discrimination contrasts effect sizes (Fig. 6A and Supplementary Fig. 1).

*3.2. Filter effects on GLCM response measures*

*3.2.1. Mean parameter estimates*

Processing method affected mean parameter estimates (Fig. 4). Notably, discrimination was not significant following PM1 and PM2 ($t(63)$ = -1. 39, -1.43, $p > 0.16$), but was following high-pass filtering ($t(63)$ = 6.82, $p = 3.9 \times 10^{-9}$).

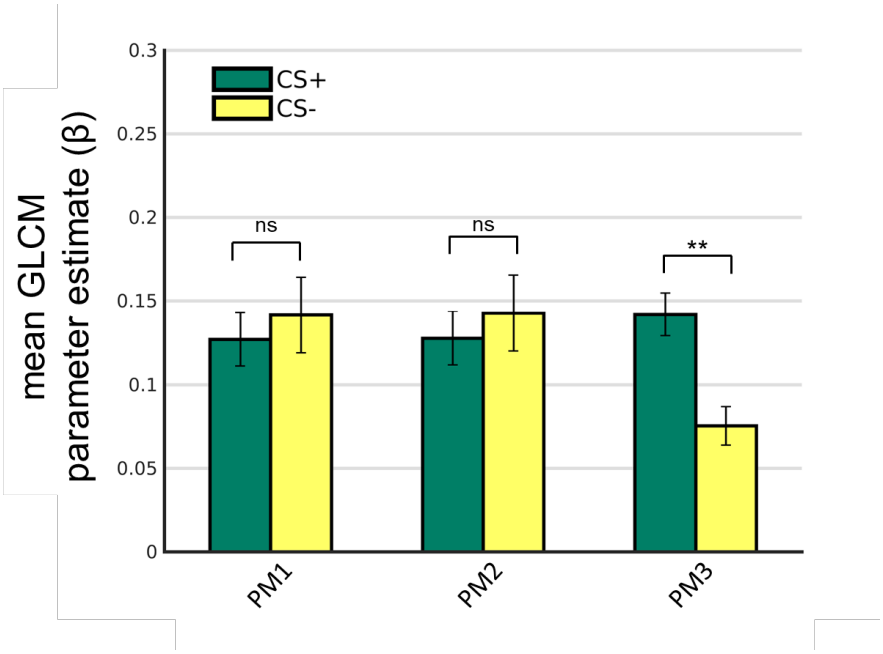



**Figure 4. Mean GLCM parameter estimates by processing method.** Parameter estimates are those of the highest retrodictive validity condition-wise GLCM methodology (rescaling method 2 (z-score SCR data prior to response fitting); PM3 high-pass cutoff 0.02 Hz). PM1 = processing method 1; PM2 = processing method 2; PM3 = processing method 3. Error bars are standard errors. Displayed contrasts are paired t-tests for group discrimination during fear conditioning (CS+ no shock – CS-). ns $p>.05$, *$p<.05$, **$p<.005$.

*3.2.2 Discrimination contrast means and effect sizes (GLCM)*

Unfiltered (PM1) and low-pass filtered SCR data (PM2) yielded inverted discrimination contrasts (Fig. 5, acquisition CS- responses tended to be greater than acquisition CS+ns responses). Band-pass filtered data (PM3) produced discrimination contrast means in the expected direction and with significantly greater effect magnitude than PM1 and PM2. Group effect sizes (Fig. 5B) were nearly equivalent for PM1 and PM2 ($d$ = - 0.1728, - 0.1773, respectively) and greater for PM3 ($d$ = 0.8456). AIC values strongly supported these differences of PM3 over PM1, but not between PM1 and PM2 (Fig. 6B, ΔAIC = - 67.1, - 0.2, respectively).

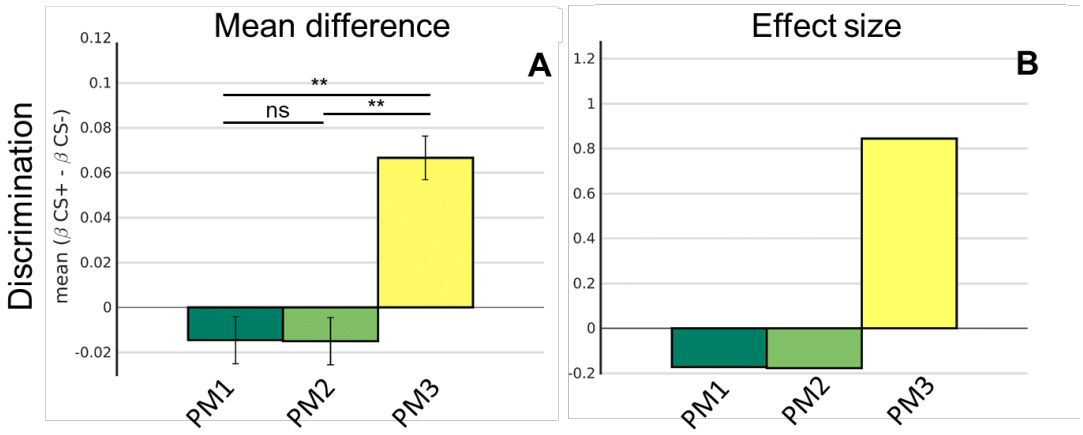



**Figure 5. Filter effects on GLCM contrasts and effect sizes.** Displayed contrasts and effect sizes are those of the highest validity condition-wise GLCM methodology (rescaling method 2 (z-score SCR data); PM3 high-pass cutoff 0.02 Hz). PM3 ($d$ = 0.8456) significantly increased effect sizes over PM2 and PM1 ($d$ = - 0.1773, - 0.1728; ΔAIC = - 66.9, - 67.1, respectively). PM2 did not significantly increase effect size over PM1 (ΔAIC = - 0.2). β = GLCM

parameter estimate. Error bars are standard errors. Displayed contrasts are 2-sample t-tests. ns $p>0.05$, *$p<0.05$, **$p<0.005$.

*3.2.3. Summary effect of filtering on GLCM-based contrasts*

A high-pass filter cutoff between 0.0159 and 0.03 Hz maximized discrimination contrast effect sizes irrespective of rescaling method (Fig. 6B; 0.03 Hz was optimum for rescaling methods 1 and 3; 0.02 Hz for rescaling method 2; 0.0159 Hz for rescaling method 4; 0.01 Hz for rescaling method 5). Note that rescaling methods 6 and 7 are not feasible on condition-wise GLCMs but are applicable to trial-wise LSS and LSA GLCMs (Supplementary Fig. 2).

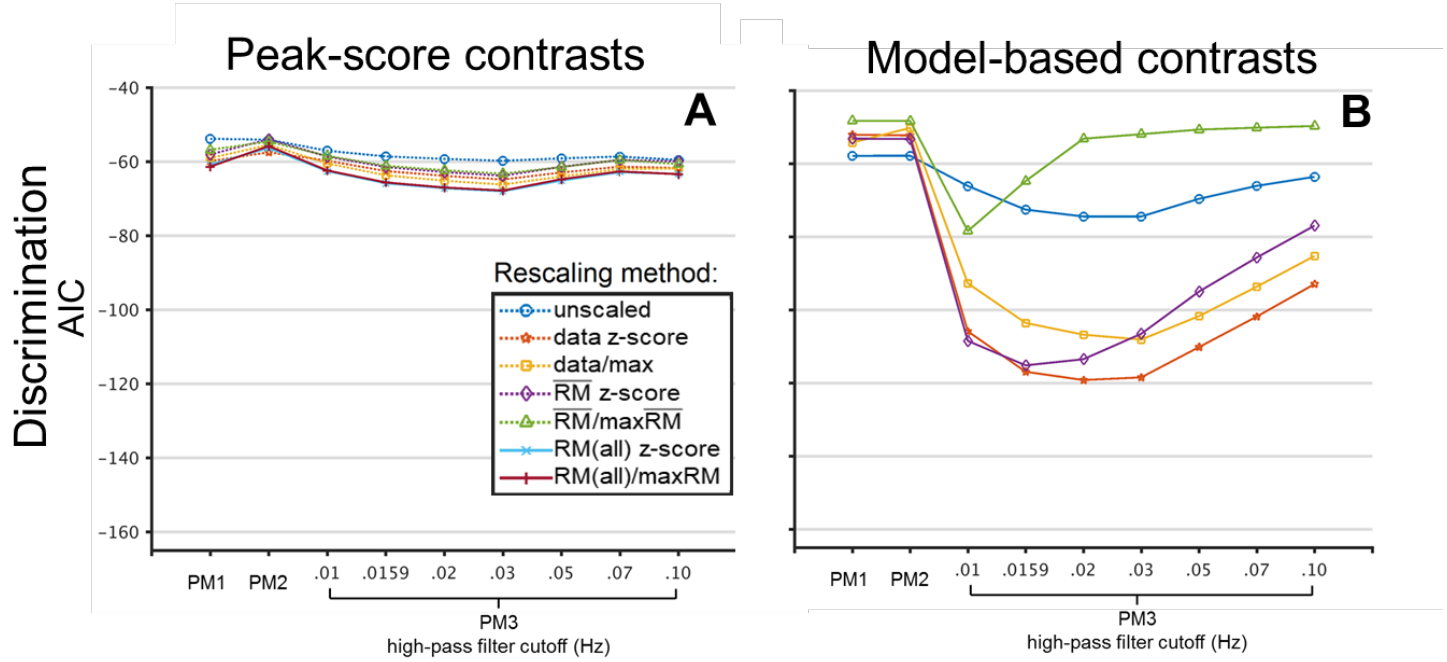



**Figure 6. Effects of filtering on retrodictive validity for known psychophysiological state.** Lower Akaike Information Criterion (AIC) values indicate better model fit. |ΔAIC| ≥ 6 between any pair corresponds to strong support for the model of lower value. The prediction model consisted of a linear mixed-effects regression where the hypothesized state acted as the response variable (dummy-coded, CS+ns = 1, CS- = 0), and subject response estimates and dummy-coded subject effects comprised the prediction matrix. Displayed peak-score (PS) AIC values (panel A) are those of the highest validity PS derivation method (maximum response range (MR) = 5 s); for a full report of PS AIC values at all MRs, see Supplementary Fig. 1. GLCM AIC values (panel B) show all tested processing methods for condition-wise GLCMs. RM = response measure.

*3.3. GLCM versus PS retrodictive validity*

The optimal condition-wise GLCM method (PM3, 0.02 Hz high-pass) resulted in an effect size increase of 110% over the optimal PS method ($d$ = 0.8456 vs. $d$ = 0.4024, ΔAIC = - 49.6). Alternative (condition-wise) GLCM methods also produced significantly greater effect sizes for discrimination than the optimal PS method (Supplementary Fig. 2, LSS d = 1.0477 (maximized by z-scoring average RMs), LSA d = 0.8520 (maximized by z-scoring all RMs).

*3.4. Exploratory analysis of the effect of filtering on a common quantitative participant exclusion metric*

Using the most effective processing method for PS analysis (MR = 5 s), PM1 and PM2 resulted in the exclusion of 32 and 31 participants, respectively, based on negative discrimination (CS+ns < CS-), whereas the most effective PM3 (high-pass cutoff of 0.02 Hz) resulted in the exclusion of 24 participants. For comparison, exclusions were greater, but still optimal at PM3, at MR = 3.5 s (PM1 n = 31, PM2 n = 35, PM3 n = 31), MR = 4 s (PM1 n = 30, PM2

n = 31, PM3 n = 26), and MR = 4.5 s (PM1 n = 29, PM2 n = 29, PM3 n = 24). For comparison, the most sensitive GLCM method (PM3 at 0.02 Hz high-pass) resulted in negative discrimination in 12 participants.

## 4. CONCLUSION

Temporal filtering of SCR, particularly of low-frequency drift artifacts, is cited in methodological texts as a useful option for removing high and low frequency artifacts (Boucsein et al., 2012; Braithwaite et al., 2013), but not universally employed in empirical studies. This is particularly important considering that SCR-fMRI studies appear to more frequently report challenges with data corruption and exclusion and because band-pass filtering is not standard practice in peak-score (PS) analysis. Previous explicit tests of filtering effects on response measures focused only on model-based regression methods on data collected outside of the fMRI environment (Bach et al., 2013). Here, we quantitatively describe the effects of basic filtering methods on experimentally-relevant SCR measures in a fMRI environment on traditional PS as well as model-based analysis, and find that band-pass filtering significantly improves signal detection for both methods.

### *4.1. Effect on PS measures*

Bandpass filtering increased effect size of PS contrasts relative to no filtering (+27%) and low-pass filtering (+60%). Low-pass filtering unexpectedly decreased PS effect sizes (Fig. 6A & Supplementary Fig. 1). We speculate that this may be due to the removal of high-frequency, motion-induced artifacts correlated with presentation of the CS+ns. This is important given that the majority of PS studies we reviewed reported using only low-pass filtering.

While we demonstrate that PS response measures yield contrast estimates in the appropriate direction without filtering, we stress that even this minimal processing method (downsampling to 10 Hz) removes many large, high-frequency artifacts (>4 *SD*), which differed in frequency and amplitude between runs and participants. We did not quantitatively evaluate the effect of downsampling on peak-scoring.

### *4.2. Effect on model-based measures*

Band-pass filtering universally increased the ability to detect discrimination contrasts. Low-pass filtering alone did not benefit model-based contrasts, and in some cases actually resulted in reduced signal recovery (i.e. when raw SCR data is rescaled to the largest data point). These findings lend independent support for the importance of previous recommendations to both low- and high-pass filter SCR data in psychophysiological model-based quantitation of SCRs (Bach et al., 2013). We speculate that nonlinear drift in unfiltered signal disrupts signal detection by skewing SCR values entering regressions. Many participants' data demonstrated such non-linear SCL (Fig. 1, subjects A, C-E). Importantly, we found in separate analyses (not shown) that linear detrending, square root transforming, and log transforming, of raw SC data or response measures, cannot adequately correct for these distortions.

### *4.3. Comparison of PS and model-based contrast detection*

The highest validity GLCM methods more than doubled discrimination contrast effect sizes over the highest validity PS measures (+110%, +160%, +112% for condition-wise, LSS, and LSA GLCMs, respectively). These results lend independent support for the superiority of model-based SCR measures over traditional PS analysis.

### *4.4. Effect of filtering on exclusions by quantitative signal metrics*

The results of our exploratory analysis on the effect of filtering on "non-learner" exclusions for negative discrimination suggest that in a PS approach, processing method can drastically influence exclusion numbers, on a range from 37% (n = 24/65) for optimally derived (MR = 5 s), band-pass filtered data (PM3 at a high-pass cutoff of 0.03 Hz), to 54% for the least effective method (PM2/low-pass filtered only) derived from a 1-3.5 s response window. For comparison, the most commonly implemented method of MR = 4 s on unfiltered or low-pass

filtered data would result in the exclusion of 46% (n = 30/65) and 48% (n = 31/65) of participants, respectively, for negative discrimination in our sample. This difference alone, of 9% between optimally filtered and unfiltered data using common methods, is greater than the difference in exclusions we found between filtered and unfiltered data across all quantitative exclusion thresholds (5%, as referenced in the Introduction, when including exclusions for negative discrimination, sub-threshold discrimination, or sub-threshold SCRs). This suggests that further work is warranted to understand how suboptimal processing methods alone may influence exclusions.

It is worth noting that some authors exclude trials rather than participants for negative or subthreshold responses. Alternatively, some authors set such trials to zero under the assumption that the true value of these responses is zero. If true, these responses might be considered noise. Consequently, excluding these responses might be thought to improve discrimination and reduce the number of participants excluded for below zero discrimination. Given that high-pass filtering reduces downward drift, it is reasonable to speculate that filtering may improve discrimination by reducing the number of negative trials (in effect offering no advantages over simply excluding negative trials in unfiltered data). However, in follow up analyses, again on unscaled, MR = 5 s peak score data, we found that filtering had a negligible effect on the average number of negative CS+ or CS- trials (CS+ PM1 = 5.2 [SD = 2.3], PM3 4.8 [SD = 4]; CS- PM1 = 10.4 [SD = 4], PM3 10.8 [SD = 3.8]). Moreover, we found that rounding negative trials to zero or rounding raw microsiemens values to the nearest third or second decimal did not reduce the number of participants excluded for negative discrimination (n = 32 in each case, the same number excluded using PM1 without zeroing or rounding of negative CRs). These results highlight the advantage of high-pass filtering to improve signal recovery and do not support the assumption that the true value of negative responses is zero.

*4.5. Implications*

Our results suggest that although it is commonly assumed that background SCL has little impact on SCR PS measures (Braithwaite et al., 2013), PS contrasts show considerable increases in sensitivity when SCL trends are removed. Moreover, these sensitivity increases were greater than those gained from low-pass filtering alone and in some cases, low-pass filtering may actually reduce effect size estimates (possibly as a result of removing noise correlated with CS+ns). These points are both significant considering that the vast majority of SCR studies do not report removal of low frequency drift with high-pass filtering, and many studies report using only low-pass filtering. This indicate that past studies have underutilized SCR, and some may have been more subject to Type 1 and/or Type 2 errors than previously thought. This may also be true of PSs calculated using a response window of less than 5 seconds. In the case of PSs calculated from unfiltered data or low-pass filtered data where maximal responses were derived from a 3.5 second post-stimulus window, for example, contrast effect sizes may be reduced by as much as 50% compared to those derived from high-pass filtered data using 5 s maximum windows. Thus, standardization of rigorous, validated band-pass filtering may reduce the number of excluded subjects and datasets, increase effect sizes of contrasts that critically inform hypotheses, and, accordingly, increase the power to detect effects at fixed sample sizes. Furthermore, our independent validation of the increased sensitivity of model-based over PS-derived scores suggests that wider implementation of model-based SCR methods is warranted. In addition, although we cannot draw conclusions regarding the source of our and others' anecdotal reports of true SCR "non-responders" in the fMRI environment, we believe that filtering and model-based analysis may reduce the exclusion of subjects for non-responding in some cases. Future work should examine whether the fMRI field truly impacts the rate of true SCR non-responding.

*4.6. Limitations*

We deliberately performed our comparison on noisy SCR data recorded in an fMRI environment, and the relative benefit of some filtering and analysis methods may be less pronounced for other experimental settings. Our analyses were limited to fMRI-concurrent SCR data, and thus we could not directly compare the effects of filtering on scanner-related versus scanner-unrelated artifacts. Furthermore, we did not allow for stabilization of

SCL prior to beginning experiments – experiments that allow for such a stabilization period would be less susceptible to drift-induced corruption. However, it is important to note that our review revealed that such baseline stabilization is not common practice, despite recommendations. Hence, we believe the demonstrated effects of filtering stand to crucially benefit future work with SCR in this environment and others.

Other limitations to the generalizability of our results may be our experiment's short ITI and restriction to women with PTSD, concurrent psychopathology, and presence of psychotropic medications. These behavioral variables may reduce discrimination due to inherently higher responses to CS- (Duits et al., 2015; Lonsdorf et al., 2015). Assuming that average higher amplitude SCRs exhibit a shorter average time constant than small SCRs, these reduced discrimination variables should favor more restrictive passbands (i.e. more restrictive/higher frequency high-pass filter thresholds). Similarly, experiments recovering signal from lower SCR experimental phases such as extinction or tests of fear recovery may favor more liberal passbands. Furthermore, although our SCRs have a minimum of 5-9 seconds for onset and recovery and our maximum response range extended only to 5 seconds, it is possible that some participants displayed peak latencies beyond 5 seconds. It is also possible that filtering particularly benefited our experiment by reducing the effect of SCR tails (particularly those elicited by aversive stimuli) on subsequent responses. As peak scores are inevitably more sensitive to tail overlap, it is expected that filtering will be of greater benefit to PS SCRs in short ITI experiments than to GLCM SCRs. However, it is notable that recent work utilizing direct nerve stimulation and recording along with concurrent SCR collection has suggested that stimulation frequencies as frequent as 0.6 Hz retain 95% of signal variance in skin-conductance data, when analyzed under LTI assumptions using a model-based approach (Gerster et al., 2017). This frequency corresponds to stimulations every 1.66 seconds, more frequent than experimental stimuli occurring most frequently every 5-9 s or CS-US intervals most frequently 2.5-6.5 s. For these reasons, we believe that the primary benefit of filtering was indeed through reducing the effects of tonic signal drift rather than by reducing SCR tail overlap. Despite this, it is imperative that future investigations explore the effect of filtering in long ITI experiments, and investigators should tailor their passband to one that maximizes signal recovery under their particular experimental conditions.

*4.7. Recommendations*

Given the aforementioned findings, we strongly recommend adoption of band-pass filtering regardless of the intended method for response quantification. For response quantification method, we strongly recommend psychophysiological modeling, as implemented by standardized software. Regarding filter frequencies, the implementation of the method (e.g. condition-wise, LSS, or LSA), type of rescaling, experimental design (e.g. ITI, balance of stimuli across time, allowance for an SCL stabilization period, etc.) and the contrast under question may influence the ideal choice. For example, rescaling subject data or responses by their maximum value often led to reduced benefit from filtering compared to other rescaling methods (Supplementary Fig.1 & 2) or, unexpectedly, even a reduction in overall sensitivity relative to less rigorously filtered data within the same rescaling method (Supplementary Fig. 2). In light of this consideration, we are comfortable with a strong recommendation to choose a high-pass filter frequency between 0.01 and 0.03 Hz for any model-based (or PS) analysis, provided that subject rescaling uses z-scoring, which scales on a multivariate measure of variance rather than a single value susceptible to outliers. While LSS-style GLCMs led to optimal signal recovery for discrimination contrasts (Supplementary Fig. 2), our ITIs are relatively short (2-6 s). As ITIs increase and the effects of collinearity and its interaction with SCL and stimuli imbalance decrease, LSA and LSS should asymptotically produce identical results. Given no information about stimuli balance and length of ITIs, we recommend using LSS-style GLCMs, as they may more accurately account for potential collinearity (Mumford et al., 2012).

If utilizing PSs, investigators using our CS-US interval of 2.5 s may tentatively consider using a 5 s post-stimulus onset to limit identification of peak maxima, a bandpass filter with cutoff frequencies of 0.02 to 5 Hz (bidirectional Butterworth filter), and z-scoring all derived PSs (i.e. across all trials, without averaging) by subject prior to analyses. However, we strongly urge investigators to consider the importance of allowing an SCR

stabilization period, and to consider that high-pass filtering may particularly benefit experiments with shorter ITIs. Given the growing complexity of experiments that take place during fMRI or that implement instrumental responses (e.g. stimulus identification, approach or avoidance responses, etc.), startle probes, and/or rating periods during CS presentation, however, we expect high-pass filtering to also benefit experiments with longer ITIs.

Implementation of these recommendations promises to increase the efficiency, productivity, and reproducibility of future research efforts. Moreover, implementation of filtering and model-based methods may decrease systematic sample biasing due to the exclusion of racial groups likely to have lower SCRs (Kredlow et al., 2017) or groups with psychological, neural, and genetic traits that correlate with motion artifacts (Couvy-Duchesne et al., 2016; Engelhardt et al., 2017). Importantly, researchers can implement the described filtering methods retrospectively. Moreover, universal implementation of both band-pass filtering and psychophysiological modeling of SCRs may increase the validity of SCR-study conclusions, increasing rigor and reproducibility. Finally, we hope that our explicit description of SCR filtering and scoring methods will reduce uncertainty and expedite efforts in SCR measurement.

**References:**


Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer, New York, NY. https://doi.org/10.1007/978-1-4612-1694-0_15

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). 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., Flandin, G., Friston, K. J., & Dolan, R. J. (2010). Modelling event-related skin conductance responses. *International Journal of Psychophysiology*, *75*(3), 349–356. https://doi.org/10.1016/j.ijpsycho.2010.01.005

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., Fleming, S. M., & Voelkle, M. (2020). *Retrodictive validity and true-score inference* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/bhdez

Benedek, M., & Kaernbach, C. (2010). 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

Ben-Shakhar, G. (1985). Standardization Within Individuals: A Simple Method to Neutralize Individual Differences in Skin Conductance. *Psychophysiology*, *22*(3), 292–299. https://doi.org/10.1111/j.1469-8986.1985.tb01603.x

Blake, D. D., Weathers, F. W., Nagy, L. M., Kaloupek, D. G., Gusman, F. D., Charney, D. S., & Keane, T. M. (1995). The development of a Clinician-Administered PTSD Scale. *Journal of Traumatic Stress*, *8*(1), 75–90. https://doi.org/10.1007/BF02105408

Boucsein, W. (2012). *Electrodermal activity*. Springer Science & Business Media.

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E., & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, *49*, 1017–1034.

Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., & Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Systematic Reviews*, *6*(1), 245. https://doi.org/10.1186/s13643-017-0644-y

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Couvy-Duchesne, B., Ebejer, J. L., Gillespie, N. A., Duffy, D. L., Hickie, I. B., Thompson, P. M., Martin, N. G., Zubicaray, G. I. de, McMahon, K. L., Medland, S. E., & Wright, M. J. (2016). Head Motion and Inattention/Hyperactivity Share Common Genetic Influences: Implications for fMRI Studies of ADHD. *PLOS ONE*, *11*(1), e0146271. https://doi.org/10.1371/journal.pone.0146271

Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., van den Hout, M. A., & Baas, J. M. P. (2015). UPDATED META-ANALYSIS OF CLASSICAL FEAR CONDITIONING IN THE ANXIETY DISORDERS: Review: Updated Meta-Analysis of Fear Conditioning in Anxiety Disorders. *Depression and Anxiety*, *32*(4), 239–253. https://doi.org/10.1002/da.22353

Engelhardt, L. E., Roe, M. A., Juranek, J., DeMaster, D., Harden, K. P., Tucker-Drob, E. M., & Church, J. A. (2017). Children's head motion during fMRI tasks is heritable and stable over time. *Developmental Cognitive Neuroscience*, *25*(Supplement C), 58–68. https://doi.org/10.1016/j.dcn.2017.01.011

Fahrenberg, J., Walschburger, P., Foerster, F., Myrtek, M., & Müller, W. (1983). An Evaluation of Trait, State, and Reaction Aspects of Activation Processes. *Psychophysiology*, *20*(2), 188–195. https://doi.org/10.1111/j.1469-8986.1983.tb03286.x

Figner, B., Murphy, R. O., & others. (2011). Using skin conductance in judgment and decision making research. *A Handbook of Process Tracing Methods for Decision Research*, 163–184.

Gerster, S., Namer, B., Elam, M., & Bach, D. R. (2017). Testing a linear time invariant model for skin conductance responses by intraneural recording and stimulation. *Psychophysiology*. https://doi.org/10.1111/psyp.12986

Gray, M. A., Minati, L., Harrison, N. A., Gianaros, P. J., Napadow, V., & Critchley, H. D. (2009). Physiological recordings: Basic concepts and implementation during functional magnetic resonance imaging. *NeuroImage*, *47*(3), 1105–1115. https://doi.org/10.1016/j.neuroimage.2009.05.033

Kredlow, M. M., Pineles, S. L., Inslicht, S. S., Marin, M.-F., Milad, M. R., Otto, M. W., & Orr, S. P. (2017). Assessment of skin conductance in African American and Non-African American participants in studies of conditioned fear. *Psychophysiology*, *54*(11), 1741–1754. https://doi.org/10.1111/psyp.12909

Lagopoulos, J., Malhi, G. S., & Shnier, R. C. (2005). A fiber-optic system for recording skin conductance in the MRI scanner. *Behavior Research Methods*, *37*(4), 657–664.

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

Lonsdorf, T. B., Haaker, J., Schümann, D., Sommer, T., Bayer, J., Brassen, S., Bunzeck, N., Gamer, M., & Kalisch, R. (2015). Sex differences in conditioned stimulus discrimination during context-dependent fear learning and its retrieval in humans: The role of biological sex, contraceptives and menstrual cycle phases. *Journal of Psychiatry & Neuroscience: JPN*, *40*(6), 368–375.

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., & Merz, C. J. (2017). More than just noise: Inter-individual differences in fear acquisition, extinction and return of fear in humans-Biological, experiential, temperamental factors, and methodological pitfalls. *Neuroscience & Biobehavioral Reviews*, *80*, 703–728.

Lykken, D. T., & Venables, P. H. (1971). Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology*, *8*(5), 656–672.

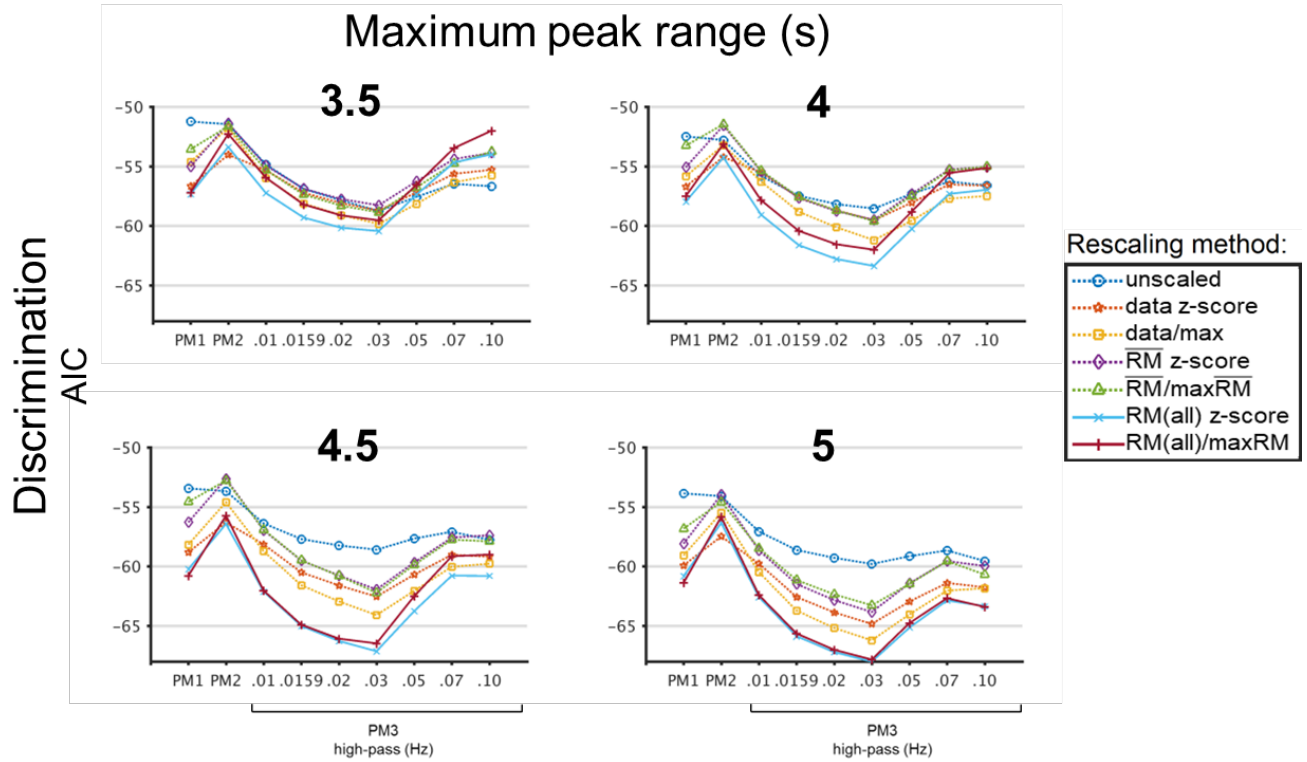
Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

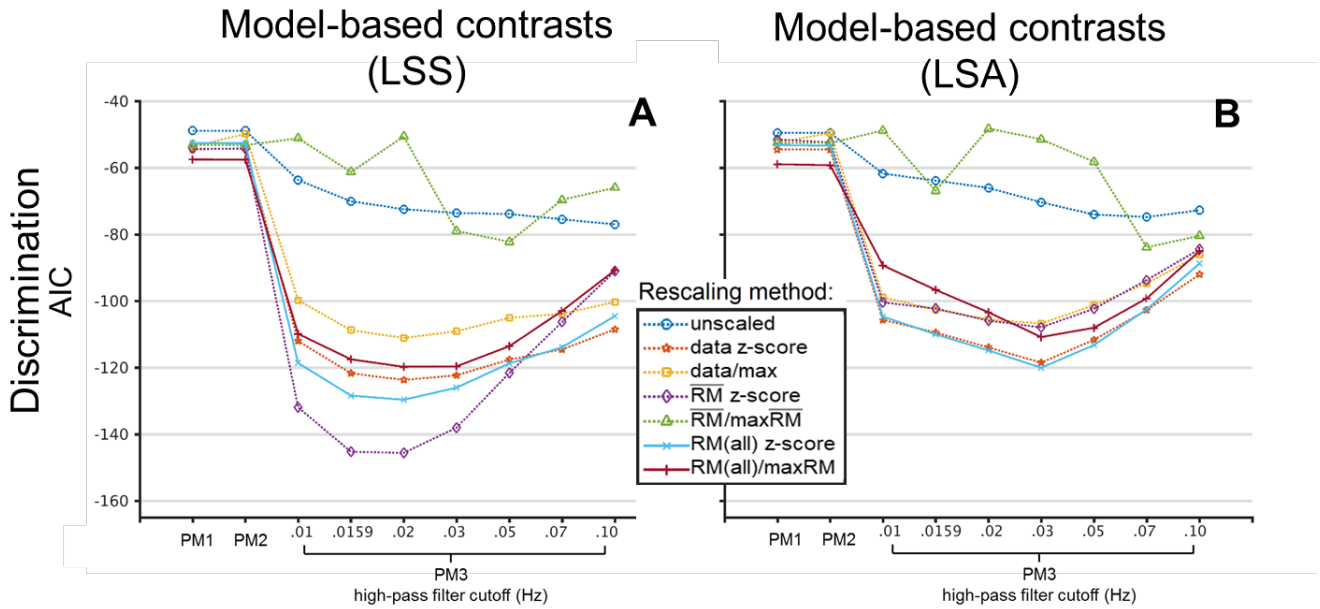
Spitzer, M., Robert, L., Gibbon, M., & Williams, J. (2002). Structured clinical interview for DSM-IV-TR axis I disorders, research version, non-patient edition (SCID-I/NP). *New York: Biometrics Research, New York State Psychiatric Institute*.

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

**Supplementary Materials:**

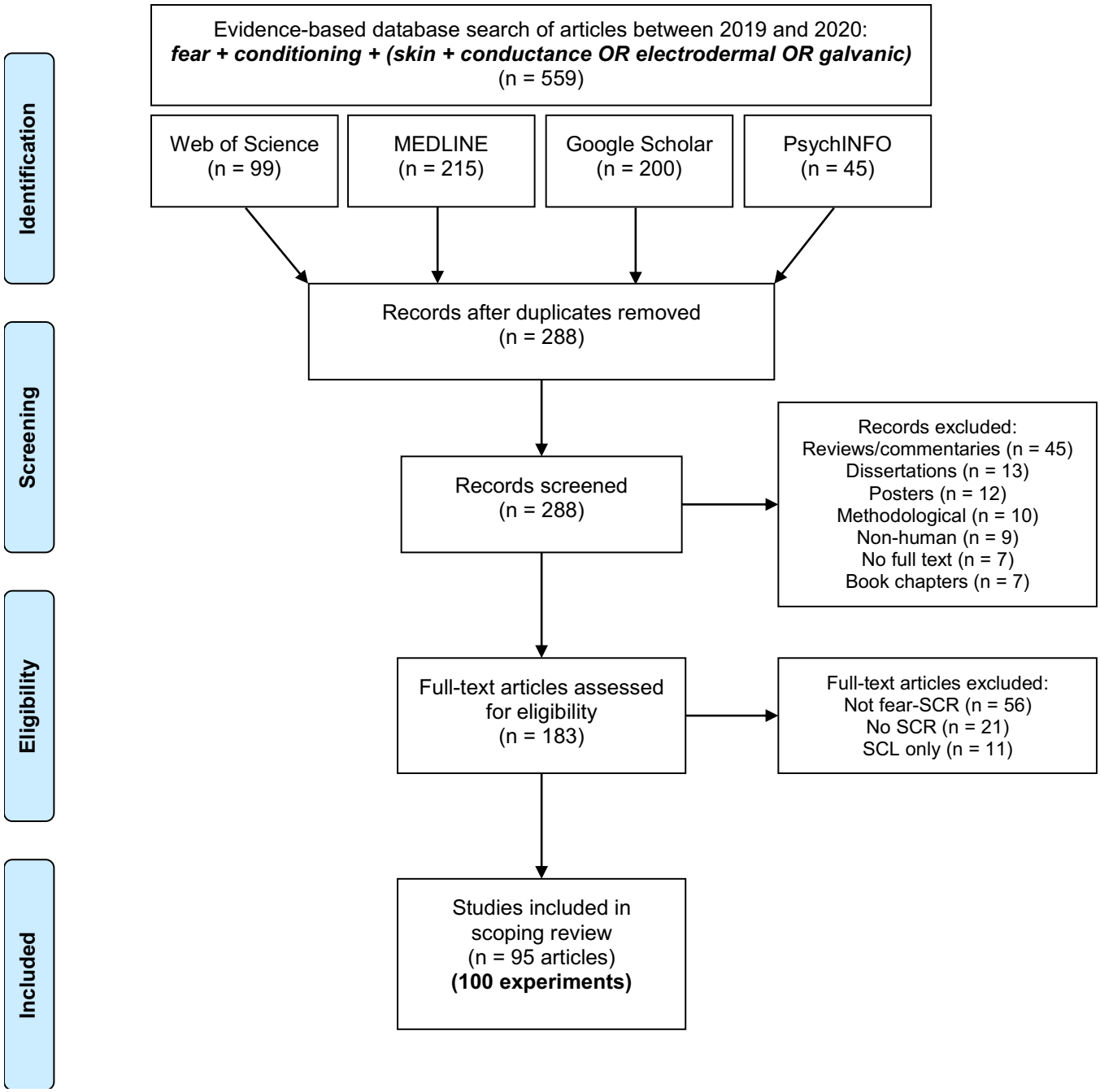



**Supplementary Figure 1. Retrodictive validity for discrimination (all PS methods).** Lower Akaike Information Criterion (AIC) values indicate better model fit. |ΔAIC| ≥ 6 between any pair corresponds to strong support for the model of lower value. The prediction model consisted of a linear mixed-effects regression where the hypothesized state acted as the response variable (dummy-coded, CS+ns = 1, CS- = 0), and subject response estimates and dummy-coded subject effects comprised the prediction matrix.




**Supplementary Figure 2. Retrodictive validity for discrimination (trial-wise GLCM methods).** Lower Akaike Information Criterion (AIC) values indicate better model fit. |ΔAIC| ≥ 6 between any pair corresponds to strong

support for the model of lower value. The prediction model consisted of a linear mixed-effects regression where the hypothesized state acted as the response variable (dummy-coded, CS+ns = 1, CS- = 0), and subject response estimates and dummy-coded subject effects comprised the prediction matrix.LSA = least-squares all (separate regressor for trial-SCR of interest and multiple nuisance regressors for all other trials). LSS = least-squares separate (separate regressor for trial-SCR of interest versus a single nuisance regressor for all other SCRs).




**Supplementary Figure 3. PRISMA Flow Diagram of systematic review of SCR-fear reports from 2019-2020.** Search methodology was consistent with recommendations for for systematic reviews in health science (Bramer et al. 2017). After exclusions (right side panels), we identified 95 eligible articles describing 100 experiments in fear conditioning utilizing SCR between 2019 and May of 2020. Article names, experimental variables, filtering methods, and exclusion criteria are listed in Supplementary Data Tables.