International Journal of Psychophysiology 91 (2014) 186–193

Contents lists available at ScienceDirect

International Journal of Psychophysiology

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

Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis

Steven R. Green [a,1], Philip A. Kragel [b,c,1], Matthew E. Fecteau [c], Kevin S. LaBar [b,c,*]

[a] Department of Psychology, Indiana University, Bloomington, IN, United States
[b] Department of Psychology and Neuroscience, Duke University, Durham, NC, United States
[c] Center for Cognitive Neuroscience, Duke University, Durham, NC, United States

ARTICLE INFO

Article history:
Received 9 August 2013
Received in revised form 22 October 2013
Accepted 24 October 2013
Available online 31 October 2013

Keywords:
Autonomic nervous system
Electrodermal activity
Psychophysiology
Sympathetic arousal
Skin conductance response
Computer-assisted scoring

ABSTRACT

The skin conductance response (SCR) is increasingly being used as a measure of sympathetic activation concurrent with neuroscience measurements. We present a method of automated analysis of SCR data in the contexts of event-related cognitive tasks and nonspecific responding to complex stimuli. The primary goal of the method is to accurately measure the classical trough-to-peak amplitude of SCR in a fashion closely matching manual scoring. To validate the effectiveness of the method in event-related paradigms, three archived datasets were analyzed by two manual raters, the fully-automated method (Autonomate), and three alternative software packages. Further, the ability of the method to score non-specific responses to complex stimuli was validated against manual scoring. Results indicate high concordance between fully-automated and computer-assisted manual scoring methods. Given that manual scoring is error prone, subject to bias, and time consuming, the automated method may increase the efficiency and accuracy of SCR data analysis.

1. Introduction

Measures of electrodermal response from the human periphery, such as the skin conductance response (SCR), provide insight into activation levels of the sympathetic branch of the autonomic nervous system (Boucsein, 1992). SCRs are thought to reflect increases in the electrical conductivity of the skin caused by the release of sweat from eccrine sweat glands located on the palmar surface of the hand and foot. Because these glands are innervated by sympathetic sudomotor nerves, they provide a window into the activity of multiple brain structures, such as limbic regions, basal ganglia, and frontal cortex, that regulate the autonomic nervous system (Edelberg, 1972).

Due to the specificity of the measure, ease of setup, participant tolerance, and relatively low cost, SCRs have gained increasing popularity in clinical, neuroscientific, and psychological studies of emotion and decision-making, learning and conditioning, orienting and attention, and deception. While there are multiple parameters associated with a SCR, such as response latency, rise time, and half-recovery time, the most commonly used parameter is the amplitude of the SCR relative to a post-stimulus baseline. Traditional scoring of SCR amplitude consisted of manually measuring the distance from trough to peak of responses that fit a well-defined set of criteria pertaining to the amplitude, latency, and duration of the response (Barry, 1990; Levinson and Edelberg, 1985).

Manual scoring of skin conductance data has multiple benefits, making it a historically popular choice of data analysis. Primarily, close inspection of trial-by-trial data traces ensures that individual responses are physiological signals related to an event of interest. Manual scorers need only examine data close in time to an event in order to score a SCR. Unfortunately, manual scoring has several drawbacks, even with computer-assisted graphical interfaces. Chief among them is the amount of time needed to perform the analysis. Since traditional scoring requires a trained rater to inspect each event for a response, studies with many events are highly time-consuming. Another drawback is human bias wherein a rater may inadvertently vary the stringency of the criteria for including a response. Finally, manual analysis has been known to suffer from the *scale invariance problem* in which the detection of an inflection point depends on what scale the rater uses to inspect the data. For instance, viewing the electrodermal trace at low magnifications or poor viewing angles may lead to the misidentification of subtle changes in electrodermal data.

In an attempt to overcome some of the problems associated with manual scoring, computer-based algorithms have been previously implemented to detect SCRs (Trosiener and Kayser, 1993), although not in an event-related fashion, as response latency and duration are not utilized in detection analyses. Generally, these methods identify points in the skin conductance time-series with a slope of zero. If the change in skin conductance within this range is large enough, it is identified

[*] Corresponding author at: Duke University, Center for Cognitive Neuroscience, B203 LSRC Building, Research Dr. Box 90999, Durham, NC 27708-0999, United States. Tel.: +1 919 681 0664; fax: +1 919 681 0815.
E-mail address: klabar@duke.edu (K.S. LaBar).

[1] These authors contributed equally to the manuscript.

0167-8760/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ijpsycho.2013.10.015

as a SCR. While these methods can accurately extract increasing portions of a time series of skin conductance data, they do not filter out responses that are not plausibly event-related from a physiological perspective (that is, time-locked to the onset of a particular stimulus of interest). Other computer-based algorithms for peak detection have been implemented and compared to manual scoring, with favorable results for experimental designs with long inter-stimulus intervals (ISIs) that can accommodate temporal separation of individual SCR profiles from successive stimuli (Storm et al., 2000). While suitable when SCRs are distant in time and do not overlap, peak detection approaches based solely on the slope of the electrodermal trace are limited in their ability to isolate overlapping responses. If two SCRs occur within a short period of time, the skin conductance trace may not peak (have a slope of 0) before rising again.

Due to the increase in popularity of rapid, event-related experimental designs with shorter ISIs, additional methods have been developed to deal with the issue of overlapping SCRs. One graphical manual approach involves extending the baseline drift at stimulus onset to the time of a skin conductance peak, essentially linearly detrending the baseline drift (Barry et al., 1993). Approaches utilizing deconvolution (Alexander et al., 2005; Benedek and Kaernbach, 2010b; Lim et al., 1997) can be used to decompose skin conductance data into tonic and phasic activity, reducing the impact of overlapping responses. The goal of these methods is to more accurately measure SCRs by generating an estimate of phasic activity with a constant level of baseline activity. Alternatively, a general linear convolution model can be used to isolate event-related skin conductance activity (Bach et al., 2009). In solving a general linear model, this method generates parameter estimates that reflect the amplitude of task-related skin conductance activity. For researchers interested in experimental designs with short ISIs, these methods may be preferential for analyzing SCR data.

While methods estimating the SCR using mathematical models are attractive from a theoretical and procedural standpoint, one main issue complicates their use when compared to manual scoring: non-specific or spontaneous fluctuations. Changes in skin conductance that occur in the absence of stimuli can introduce error into models of electrodermal time-series. Spontaneous fluctuations have been successfully incorporated into generative models of skin conductance activity (Bach et al., 2010), although it remains unclear under what conditions assumptions about the occurrence and duration of these activations are valid. If assumptions concerning when spontaneous fluctuations are likely to occur are incorrect, the estimation of event related responses could be negatively impacted. We posit that, in the context of event-related analysis, focusing on data that is close in time to an event (i.e. the rise of the SCR) and is not dependent on characterizing spontaneous fluctuations will perform more consistently across a variety of experimental settings.

Here we present a traditional method of SCR data analysis in the context of event-related cognitive tasks that is fully-automated and does not depend on fitting data to a modeled response profile. The goal of our method is to automate manual scoring of stimulus-locked SCR amplitudes, while systematically dealing with overlapping SCRs and other common problems that introduce biases in manual scoring, such as consistency in applying response criteria. By design the software (called 'Autonomate') will apply the same criteria to each event to determine if a response occurred, thus avoiding the problem of manual raters inadvertently shifting their stringency of criteria as the data are analyzed. Furthermore, variation in the scale used to inspect the data by a manual rater (e.g. scoring under different magnifications) is not an issue for the software, as it is scale-invariant. To validate the new method in event-related paradigms, three archived datasets previously scored by two manual raters are analyzed using four software packages (Autonomate, and three methods which aim to address the issue of overlapping responses — AcqKnowledge, Ledalab, and SCRalyze), and the results are compared using standard metrics. To generalize the use of our method beyond event-related designs, we additionally validated Autonomate against manual scoring of non-specific SCRs in a fourth dataset of electrodermal responses to cinematic films. Complex datasets of this nature provide a challenging test of the software's utility as they contain more frequent and highly variable SCRs compared to event-related designs. By validating the software in a variety of experimental paradigms, we can more precisely determine under what conditions it is a suitable alternative to manual scoring.

2. Materials and methods

2.1. Automated method

Prior to analysis, data recorded at a sampling rate of 200 Hz were preprocessed using a 25 Hz finite impulse response low-pass filter and smoothed using a 3-sample moving average function. SCRs – one dimensional vectors of digitized data here denoted as $\mathbf{S}$ – were segmented into windows of $\mathbf{L}$ seconds following each stimulus. These data were down-sampled to 8 Hz using a Chebyshev Type I filter in order to reduce the effect of high frequency noise on subsequent analysis. The rises of candidate SCRs were found by searching for sections of the first order temporal derivative of the skin conductance data, $\mathbf{S}'$, that are above the threshold of $\mathbf{U}$ μS per second for a minimum duration of $\mathbf{w}$ seconds (Fig. 1A).[2] The start and end of candidate SCRs were determined by the zero crossings of $\mathbf{S}'$, and the amplitude was recorded as the difference in $\mathbf{S}$ between the second and first crossings.

Candidate SCRs were classified as being isolated or affected by neighboring responses by searching for patterns of inflection points (zero crossings in the second order temporal derivative). Inflection points were categorized based on whether the slope goes from increasing to decreasing (type $\mathbf{A}$) or decreasing to increasing (type $\mathbf{B}$) around the point. If there were three sequential inflection points within a SCR with a pattern $\mathbf{A}$–$\mathbf{B}$–$\mathbf{A}$, then the center point $\mathbf{B}$ was used to split the SCR into multiple candidate responses (Fig. 1B).

Each candidate SCR must meet a specific set of criteria used in hand scoring in order for the SCR amplitude to be regarded as time-locked to the stimuli and recorded. Consistent with our prior reports (e.g., Dunsmoor et al., 2009; Huff et al., 2009; Thomas and LaBar, 2008), the following response criteria were established as default: the latency between the eliciting stimulus onset and the rise of the response, or SCR latency, must be between 1 and 4 s; the time between response start and peak, or SCR duration, must occur between 0.5 and 5 s; and the response amplitude must be greater than 0.02 μS. While these default values are recommended, these criteria are free parameters in the Autonomate software (as are the variables $\mathbf{U}$, $\mathbf{L}$, and $\mathbf{w}$) and can be adapted by the user if further optimization is required. In the case that multiple SCRs meet all the criteria, the largest response within the window was recorded. For the analysis of spontaneous responses, the criterion of response latency was relaxed and all responses within a specified window were recorded. Once the final response was selected, the response amplitude was computed by finding the difference between local maxima and minima in the preprocessed (not down-sampled) data.

2.2. Validation of Autonomate software

To ensure the performance of our automated method closely matched that of the manual scoring on which it is based, subsamples of randomly selected subjects from three archived event-related studies were analyzed both manually (by two expert raters) and with the automated method. Studies using a range of stimuli and tasks were chosen to generate SCRs with variable amplitude, latency, and degree of overlap. For the event-related studies, electrodermal activity was recorded from the nondominant hand, using Ag–AgCl electrodes attached to the

[2] The term 'rise' is used here and throughout as a period of increase over baseline; not to be confused with the more specific term 'rise time,' which refers to the length of time from onset to peak for a SCR.

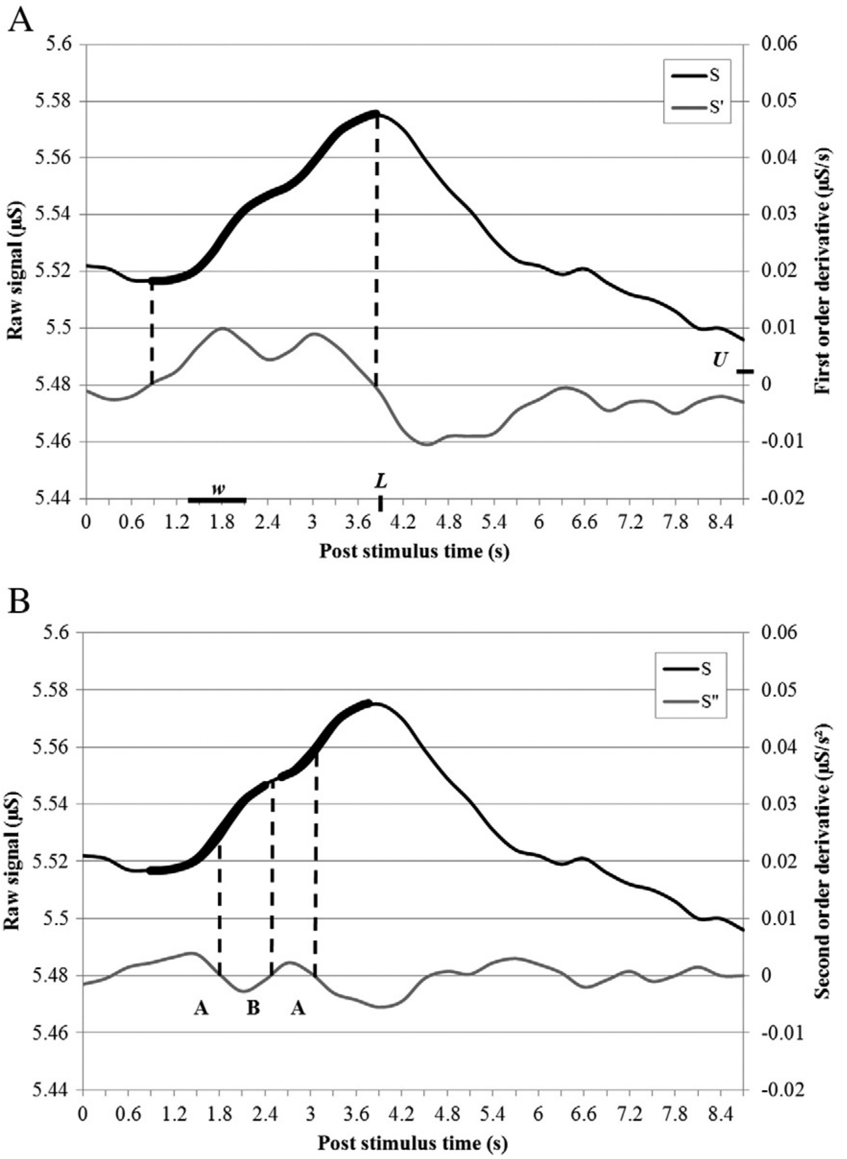

Fig. 1. Example of Autonomate's SCR amplitude estimation with overlapping responses. (A) Candidate SCR is detected by examining the first order temporal derivative **S′** up to time **L** for contiguous sections above threshold **U** μS/s for at least **w** s. (B) Candidate SCR is split by finding a sequence of zero crossings in second order temporal derivative **S″**. Crossings where **S′** is increasing are labeled **A** and crossings where **S′** is decreasing are labeled **B**. A pattern of **A**–**B**–**A** crossings indicates an inflection point where a candidate SCR should be split at point **B**. Candidate SCRs are indicated with thick black line superimposed on raw skin conductance signal **S**. Zero crossings are indicated with dashed lines.

middle phalanges of the second and third digits. In addition to the event-related studies, a study eliciting physiological responses with film clips was selected in order to validate the reliability of the method at estimating the amplitude of spontaneous responses. In this study electrodes were placed on the hypothenar eminence of the non-dominant hand. In all studies, participants were healthy adult volunteers (with no self-reported alcoholism, substance abuse, or current use of psychotropic medication) who were either students at Duke University or from the local community.

2.2.1. Study 1: context- and time-dependent fear renewal

This fear conditioning study (Huff et al., 2009) explored how the renewal of fear after extinction training is impacted by context shifts and the acquisition-to-extinction retention interval. Conditioned stimuli consisted of visual images of a snake and a spider, one of which was partially reinforced by electrical stimulation to the wrist during acquisition training. SCRs were measured in response to the 4-s presentation of the visual images during fear conditioning, extinction, and renewal testing. The mean ITI was 11 s. For validation purposes, SCRs to the visual stimuli were averaged across 40 trials during conditioning (Condition 1) and 32 trials for both the extinction (Condition 2) and test phases (Condition 3) for each subject.

2.2.2. Study 2: probabilistic classification learning with emotional and neutral outcomes

This event-related study (Thomas and LaBar, 2008) investigated how emotional outcomes in a probabilistic learning task impacted learning strategy and accuracy. Stimuli consisted of visual images of differently patterned cue cards that were probabilistically followed by associated visual outcomes. The outcomes were either relatively neutral (pictures of flowers and mushrooms) or negative (pictures of snakes and spiders). Prior to visual feedback, performance accuracy was indicated by auditory feedback lasting a half second on each trial, with presentation of a tone indicating correct responses or four 80-dB bursts of white noise indicating incorrect responses. The mean ITI was 5.5 s. For validation purposes, SCRs to the 4-s cue card onsets were averaged across 50 trials for each of two phases of the study for each subject in two separate days, yielding four cells for analysis (Conditions 1–4 ordered chronologically).

2.2.3. Study 3: generalization of fear along a gradient of facial expression intensity

This study (Dunsmoor et al., 2009) examined how fear conditioning to a moderately fearful face generalized to other faces along an emotional intensity gradient. Stimuli included five faces expressing fear that were morphed between neutral and fearful endpoints. During conditioning, the intermediate morph was paired with an electrical stimulation to the wrist whereas the most neutral face was explicitly unreinforced. The generalized stimuli consisted of the three other face values and were presented before conditioning (preconditioning) and during a post-conditioning generalization test. SCRs were measured in response to the 4-s duration presentation of the face stimuli. These three phases had mean inter-trial intervals (ITIs) of 6 s, 9 s, and 7 s, respectively. For validation purposes, SCRs to the generalized faces were pooled across all trials for each phase of the study for each subject. For each subject, 30 trials were averaged for the preconditioning phase (Condition 1), 20 for the fear conditioning phase (Condition 2), and 15 for each of three generalization tests (Conditions 3–5).

2.2.4. Study 4: psychophysiological responding during distinct emotional states

This study (Kragel and LaBar, 2013) examined how patterns of autonomic nervous system activity can be mapped to the experience of distinct emotions. Instrumental music and film clips were used to induce the experience of discrete emotions while autonomic nervous system activity was concurrently recorded. These stimuli are relatively complex and temporally-extended compared to those in the event-related studies, producing a large number of SCRs over a period of approximately 2 min. This aspect of the stimuli is critical to the validation of the software, because it provides the opportunity to score a large number of non-specific responses without a clear structure of events. For the purposes of the present work, we examined the amplitude of SCRs during the presentation of film clips intended to induce amusement and neutral states. For validation, the amplitude of SCRs was averaged across two trials for each of the two conditions for each subject.

2.3. Manual scoring

Researchers (M.E.F. and two student research assistants) were trained by a senior researcher to manually score SCRs using AcqKnowledge software (BIOPAC Systems, Goleta, CA). The training involved identification of various types of waveform artifacts, an explanation of candidate SCR criterion, and a dataset to apply what has been learned. Following completion of the practice dataset, the individual's scoring was compared to previous scoring completed in the lab to ensure consistency. AcqKnowledge software permits the rater to graphically view a segment of the SCR time series constituting a portion of a single trial to manually determine whether a candidate SCR meets the response criteria (see Section 2.1). If there are no responses that meet the criteria for a given trial, the data are scored as having 0 μS amplitude. The same set of preprocessing steps was used across the manual and automated scoring methods (see Section 2.1).

2.4. Concordance with manual scoring

For each of the four experiments, a series of analyses characterizing the agreement between manual and Autonomate amplitude estimation were performed. One-way random-effects average score intraclass correlation coefficients (ICCs, McGraw and Wong, 1996) were computed to assess the concordance of scoring across subjects within conditions specific to each study. For this computation, the scores from both manual raters were averaged and compared against those produced by the Autonomate software (the average ICC between manual raters was 0.982). In addition to examining concordance at the subject level, ICCs were additionally computed on a trial-by-trial basis for each subject to confirm that agreement at the subject level was driven by concurrence on individual trials. The trial-wise analysis was only performed for the event-related studies, because in the non-event related study there are multiple responses per trial and the number of responses can vary by rater.

Bland–Altman plots (Bland and Altman, 1986) were created using subject averages concatenated across conditions. This approach provides a graphical means of investigating the agreement between two measures by plotting the mean of two measures against their difference. Data with high agreement should fall along a horizontal line with little absolute difference. To produce these figures, data for the two computer assisted scorers were averaged, and the mean of manual and automated scoring was plotted against the difference of the two methods for each study. These plots yielded both the bias (the average difference of mean scores) between the two scoring systems and statistical outliers (points that fall outside a 95% confidence interval of the bias) that reveal significant differences in scoring between the two methods. In addition, correlations between the mean SCR amplitude and difference of SCR amplitudes were conducted to determine if discrepancies between manual and automated scoring are produced by a proportional bias (Ludbrook, 2010). For a similar approach to methodological validation in brain volumetry, see Morey et al. (2009).

Finally, a Pareto analysis (Gougeon, 2008) was performed to identify the factors that were most relevant to causing discrepancies between the two methods. For this analysis, data that had been identified as an outlier from the Bland–Altman plots (falling outside 2 S.D.'s from the mean difference) were subject to further investigation. This was accomplished by visually inspecting plots for discrepant data points that differed in amplitude by more than .1 μS between automated and computer assisted scoring. Raw traces for these trials were shown to the raters and they were asked to identify the cause of the deviation (such as issues resolving individual responses, presence of recording artifacts, etc.). This analysis highlights systematic sources of error which maximally contributed to discrepancies between Autonomate and manual scoring.

2.5. Comparing methods

In order to compare the overall effectiveness of Autonomate with other publicly available methods, the three event-related datasets were additionally analyzed with the software packages AcqKnowledge, Ledalab, and SCRalyze. The event-related EDA analysis routine from AcqKnowledge software version 4.1 (BIOPAC Systems Inc., Goleta, CA) was used to quantify SCR amplitudes. Continuous decomposition analysis (Benedek and 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. General linear models 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. The results for all methods were compared for each study using Bayes factors computed using the estimated error variance from a one-way analysis of variance (ANOVA) model that was created to test for the main effect of condition. For each of the three experiments, ANOVAs were performed using SPSS for Windows. In these models, *F*-tests examining the main effect of condition for each study were conducted in order to qualitatively compare effect sizes.

Using the error variance from these models, Bayes factors were computed for each pairwise combination of methods using the Bayesian information criterion (BIC) approximation (Schwarz, 1978). This method approximates the log of a Bayes factor as the difference of BIC between the two models (i.e. $\log(BF) = BIC_1 - BIC_2$). Bayes factors indicate how much evidence there is for one model relative to another, with values less than one favoring the reference model. A general guideline for

Table 1
Validation measures comparing Autonomate and manual scoring.

	Trial type	ICC(1,1)	F	Bias (μS)
Study 1	Condition 1	0.998	402.3*	0.003 (0.025)
	Condition 2	0.998	507.6*	−0.0001 (0.014)
	Condition 3	0.995	222.1*	0.003 (0.020)
Study 2	Condition 1	0.996	272.6*	0.028 (0.047)
	Condition 2	0.977	43.7*	0.025 (0.085)
	Condition 3	0.982	56.6*	0.022 (0.042)
	Condition 4	0.988	81.0*	0.015 (0.054)
Study 3	Condition 1	0.996	227.3*	0.002 (0.027)
	Condition 2	0.987	76.9*	0.008 (0.017)
	Condition 3	0.969	31.9*	0.010 (0.033)
	Condition 4	0.984	64.1*	0.008 (0.030)
	Condition 5	0.976	41.3*	0.004 (0.021)
Study 4	Condition 1	0.996	247.8*	−0.040 (0.056)
	Condition 2	0.983	58.1*	−0.117 (0.148)

Note. Bias indicates the mean and standard deviation in parenthesis of the difference between methods.

* $p < 0.001$.

interpreting Bayes factors suggests values between one and three indicates weak evidence, whereas increasing values indicate more definitive favor of one model over another (Jeffreys, 1961). This method of Bayesian analysis favors parsimonious models (Jefferys and Berger, 1992) and provides an alternative means of comparing methods beyond simple effect sizes.

3. Results

Reliability analyses indicated high levels of agreement between manual scoring and Autonomate. Table 1 shows ICCs for all four studies, corresponding F statistics, and within condition bias estimates comparing automated to manual scoring. Overall, there is a high degree of concordance in estimating response amplitude between manual and the automated scoring methods. Agreement of subject averages for each condition revealed excellent agreement between the two methods. Reliability analysis of individual responses similarly revealed excellent reproducibility, although at lower levels than subject averages. Examining trial by trial concordance, Study 1 ($N = 20$) had an average ICC of .872, Study 2 ($N = 20$) had an average ICC of .794, and Study 3 ($N = 20$) had an average ICC of .905.

Bland–Altman plots (Fig. 2A) show general agreement between Autonomate and manual scoring for all studies. In the event-related experiments (Studies 1–3), points in the plot fall evenly around the mean bias line, suggesting that Autonomate is not biased to over- or under-

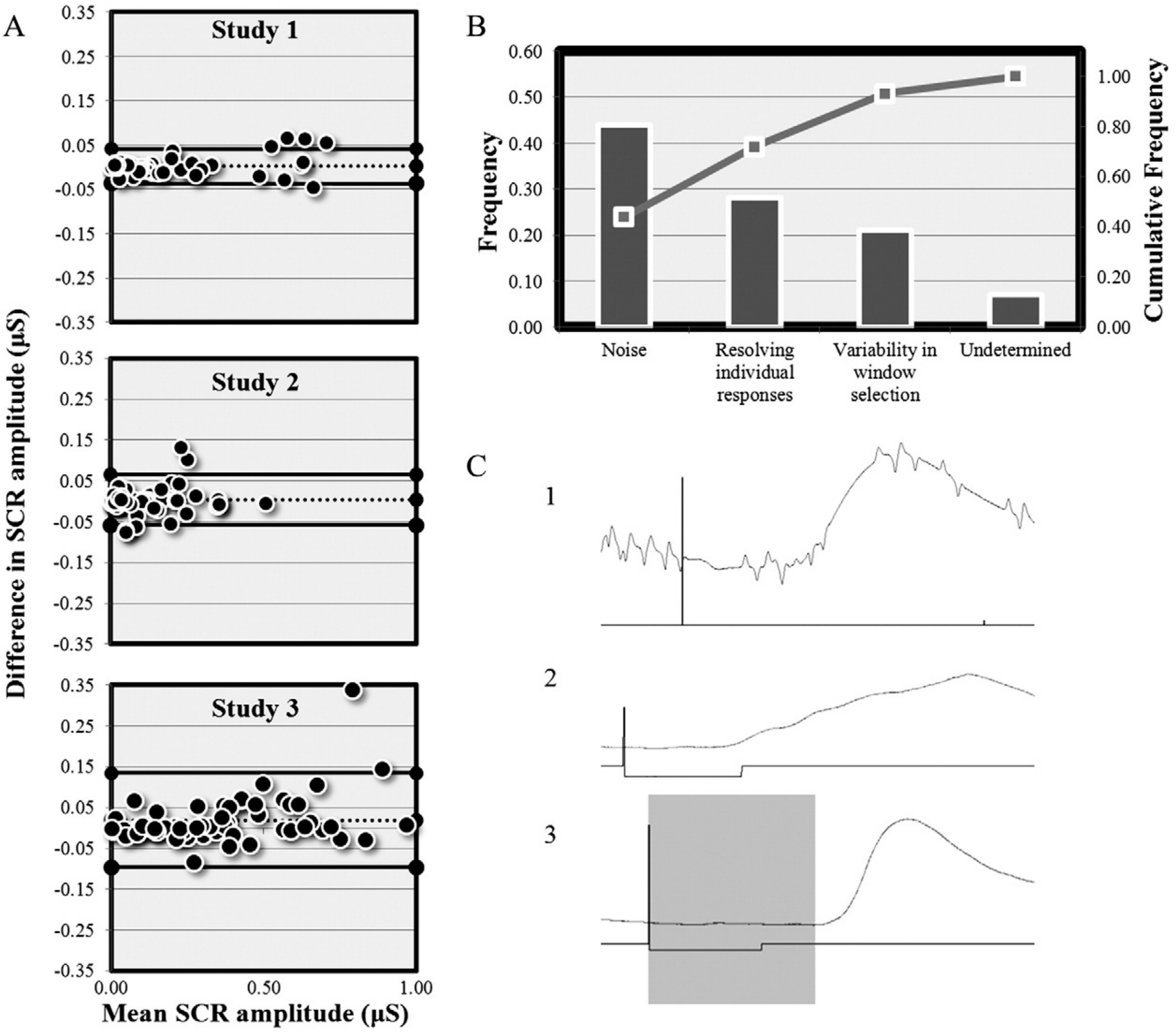

Fig. 2. Sources of variance between Autonomate and manual scoring methods. (A) Bland–Altman plots indicating excellent general agreement between methods, with a few outliers. (B) Pareto chart showing root causes of outliers. Bar plots indicate relative frequency (out of all outliers) while the scatterplot indicates cumulative frequency. (C) Exemplar trials in which (1) noise in data, (2) resolving individual responses, and (3) variability in window selection led to discrepancies in SCR amplitude measures between the two methods.

estimate SCR amplitude. Consistent with the event-related studies, Bland–Altman plots of Study 4 showed little difference between Autonomate and manual rating at typical SCR amplitudes (0 to 1 μS) although proportional bias was evident when the average SCR amplitude exceeded this range. Pareto analysis (Fig. 2B) revealed that these differences were caused primarily by differences in scoring noisy data, with discrepancies in resolving individual responses and variability in window selection (i.e. human error in selecting the start of an event-related window) accounting for most of the remainder of the outliers.

Descriptive statistics of SCR amplitude for the event-related studies show generally consistent results for the methods compared (Fig. 3). It is important to note that parameter estimates from SCRalyze are in arbitrary units and cannot be directly compared to the other measures. For all three studies, all SCR analysis methods showed the same trends across conditions, with the exception of SCRalyze in Study 2. Despite generally exhibiting similar changes in SCR amplitude across conditions, the AcqKnowledge software generally over-estimated responses in Study 3. The reconstructed SCR amplitudes from Ledalab were similar to those of computer assisted and automated scoring, but with overall larger estimates of SCR amplitude. These results are consistent with reports that scoring methods accounting for fluctuations in baseline activity yield larger amplitude estimates than classical SCR scoring methods (Benedek and Kaernbach, 2010a).

Statistical measures of predictive validity showed equivalent results regarding the effectiveness of computer based methods and manual scoring. *F*-tests for the main effect of condition (Table 2) showed that the methods had differing sensitivities to experimental conditions depending on the study. In Study 1, all methods yielded a significant

Table 2
Measures of predictive validity for all scoring methods.

	Rater type	*F*	p	Estimate of error variance	Bayes factor (Manual)	Bayes factor (Autonomate)
Study 1	Manual	7.249	0.002	0.034	–	–
	Autonomate	7.129	0.002	0.035	1.006	–
	SCRalyze	14.207	0.000	0.174	1.426	1.417
	Ledalab	12.233	0.000	0.061	1.135	1.128
	AcqKnowledge	9.290	0.000	0.020	0.891	0.886
Study 2	Manual	0.292	0.831	0.014	–	–
	Autonomate	0.368	0.776	0.015	1.015	–
	SCRalyze	2.808	0.046	0.115	1.580	1.556
	Ledalab	0.438	0.726	0.074	1.436	1.414
	AcqKnowledge	0.713	0.548	0.017	1.043	1.028
Study 3	Manual	7.453	0.000	0.039	–	–
	Autonomate	7.376	0.000	0.045	1.032	–
	SCRalyze	6.945	0.000	0.109	1.250	1.212
	Ledalab	2.667	0.037	0.182	1.397	1.354
	AcqKnowledge	5.414	0.001	0.071	1.139	1.104

Note. Bayes factors were computed as $BF = e^{\frac{\log \sigma_{e_1}^2 - \log \sigma_{e_2}^2}{2}}$ using manual scoring or Autonomate as the reference (model 2). Bayes factors greater than one favor the reference (second) model.

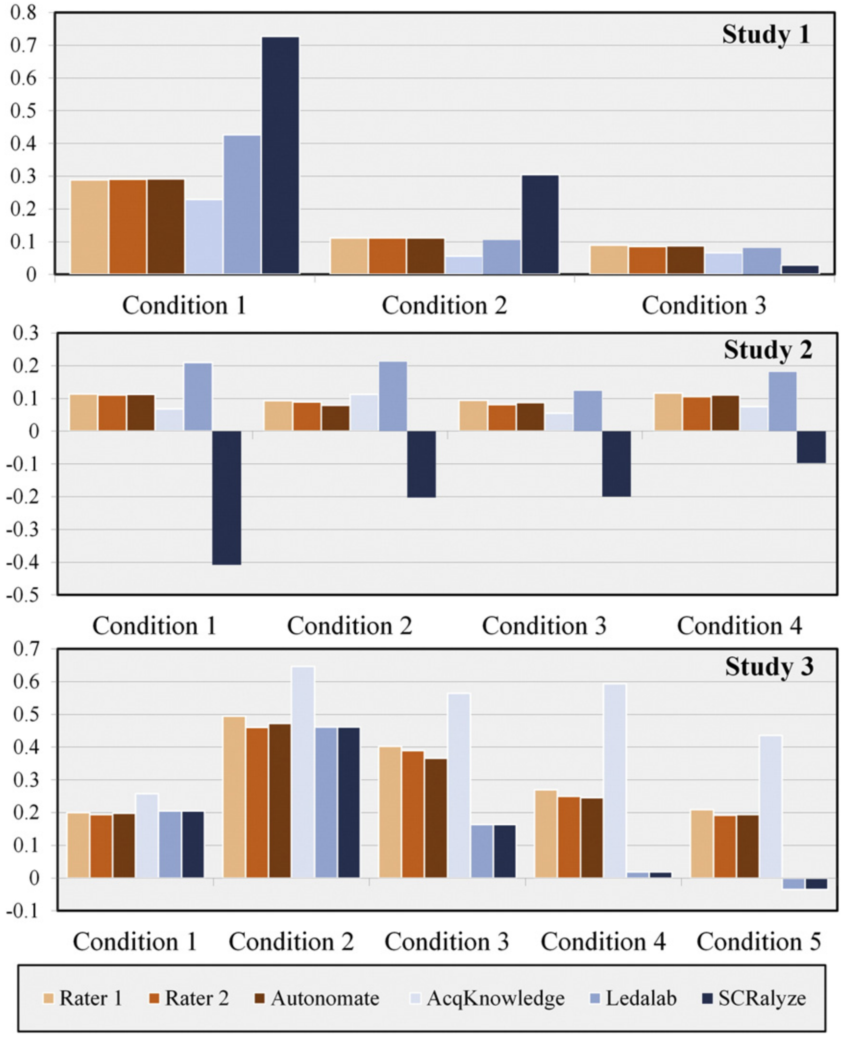

Fig. 3. Mean of SCR amplitude for each experimental condition in event-related studies. All measures reflect SCR amplitude in μS with the exception of SCRalyze, which indicate parameter estimates of the SCR in arbitrary units.

effect, with SCRalyze and Ledalab having larger effect sizes. Results for Study 2 indicated that only SCRalyze could detect a significant difference between conditions. Study 3 showed that all methods were sensitive to the main effect of condition, with Ledalab producing a weaker effect size. Generally, no method produced consistently larger effect sizes than the alternative methods. Comparing Bayes factors across studies for the automated method yielded largely similar results. Bayes factors close to one indicate nearly equal evidence for Autonomate and manual scoring. Bayes factors ranging from 1.1 to 1.6 show weak evidence favoring manual scoring over other software packages, with the exception of AcqKnowledge in Study 1.

4. Discussion

We developed an unsupervised, fully-automated method of analyzing SCR amplitude while accounting for overlap from neighboring responses, based on classical metrics of manually scoring electrodermal activity. The goal of this method is to provide a fully automated alternative to computer-assisted manual scoring that is consistent with the classical analysis of scoring SCRs. Manual scoring of SCRs is a demanding procedure that requires training and experience in order to be reliable and accurate. Additionally, the scoring of large datasets, especially event-related designs with high number of trials, requires a considerable amount of time and resources to complete. Given that manual scoring is subject to scorer bias, error prone, and time consuming, fully automated methods for analyzing SCRs have many advantages from a methodological and practical standpoint.

The proposed method of analyzing SCRs shows excellent absolute agreement with manual scoring methods and has little bias in measures of SCR amplitude. Across three event-related studies that had different stimuli, ISIs, number of trials, and ranges of SCR amplitudes, intraclass correlations indicate high concordance between the two trained computer assisted scorers and the automated method. In addition, similar high levels of concordance were observed when the approach was modified to score non-specific responses during the viewing of complex films. Bland–Altman plots revealed, however, a proportional bias in validation of non-specific responses, whereby subjects with larger SCRs produced lower averages when scored by manual raters. Despite this bias, the differences in estimated SCR amplitude were acceptable when SCRs were within the typical response range (0–1 μS). Given the proportional bias for larger SCRs, we recommend manual inspection of electrodermal data if Autonomate identifies SCRs in excess of 1 μS. We have included a graphical user interface which permits inspection and manual adjustment of Autonomate's scores. This interface allows researchers flexibility in checking the results of the software, either for validation purposes or to identify recording artifacts. These results suggest that our method is generalizable to different study designs and can produce SCR estimates equivalent to manual scoring, particularly in the case of event-related designs.

While ICC coefficients and Bland–Altman plots suggest excellent agreement between automated and manual scoring, identifying the sources of discrepancy between the two scoring methods is important to determine that systematic bias is not being introduced to SCR scoring. Root cause analysis showed that the majority of differences in the proposed method and manual scoring stem from variability in human scoring rather than limitations of the automated analysis. Determining when a response starts rising and when it stops is critical in assessing if a SCR meets the latency and duration criteria. Further, the sensitivity with which inflection points are identified is crucial to separating the effect of overlapping responses. Unlike human raters, the automated method is able to consistently use the same algorithm for determining these characteristics of a SCR. While having discrepancies in the scoring of noisy data is a concern for automating manual scoring, they contributed to approximately three outliers in a sample of 140 data points. The noisy data that is similar to skin conductance signals of interest is a problematic issue for any analysis schema, making it imperative to ensure good recording to generate accurate data, including maintaining good contact between the electrode and skin and minimizing motion artifacts. Considering the ICC, Bland–Altman, and root cause results in conjunction, we conclude that the proposed method is a suitable alternative to computer assisted scoring as far as validity is concerned.

Beyond our primary goal of automating manual scoring of SCRs in a computer-based implementation, we were additionally interested in examining how the automated method performs relative to other SCR analysis methods under different experimental conditions. While all of the tested methods hold very similar assumptions about the timing and shape of the skin conductance response, they are implemented using different algorithms, which may impact their performance in different contexts. All methods showed the same trends in all three studies, with the exception of SCRalyze in Study 2 and AcqKnowledge in Study 3. The parameter estimates from SCRalyze were negative for all conditions, indicating that the baseline level of skin conductance activity was greater than that during the event. Modeling events relative to constant baseline is not a feature of any of the other methods. Despite responses being below the implicitly modeled baseline, SCRalyze was the only method that differentiated the four conditions in Study 2. In Study 3, AcqKnowledge generally yielded larger responses than other methods, potentially due to overlapping responses in the event-related design. Together, these results show promise for the method in experimental settings with short stimulus durations and ITIs.

Using *F* statistics as a measure of predictive validity, there was little evidence for a method that was consistently superior across all three studies. In Study 1, which had the longest average ITI, Bayes factors weakly favored the AcqKnowledge software relative to other methods. In Studies 2 and 3, which had shorter ITIs, Bayes factors weakly favored the simpler of the methods tested: manual scoring and Autonomate. These two approaches involve fewer parameters in their analysis of the data and are more attractive than more complicated methods in this light. Considering consistency across studies and Bayes factors, these results favor Autonomate for researchers interested in directly comparing results from studies using both rapid and sparse event-related experimental designs. More broadly, these findings suggest that while there are a number of suitable methods and software packages for the analysis of SCRs, there is a need to determine under what experimental contexts (e.g. electrode placement, participant population, or medication status) different methods are advantageous on a larger scale.

One advantage with the present approach relative to modeling methods is its parsimony. However, modeling approaches may be warranted when the stimulus durations and ITIs are smaller than those used for the present analysis. In particular, we focused on minimum stimulus durations of 4 s and ITIs of 5 s. In experimental contexts where model-based approaches may prove more effective, the presented method is well-suited to serve as a reference to test model assumptions. Additionally, the current automated method facilitates the development and validation of new methods by easing the time and resource burden associated with manually scoring skin conductance data. With these experimental design parameters in mind, the outcome of the present study provides initial validation of the accuracy and precision of the proposed method, which makes it an attractive alternative to manual scoring methods and a useful tool for future developments in analyzing skin conductance data.

Acknowledgments

This work was supported by a Duke Institute for Brain Sciences Research Incubator award, NSF grant 0745919, and NIH grants 2 P01 NS041328, RC1 MH088678, and R01 DA027802. We thank Allison Bullock and Korrina Duffy for assistance with manual scoring of the data. For inquiries and information about the Autonomate software, go to http://www.labaratory.org.

References

Alexander, D.M., Trengove, C., Johnston, P., Cooper, T., August, J.P., Gordon, E., 2005. Separating individual skin conductance responses in a short interstimulus-interval paradigm. J. Neurosci. Methods 146, 116–123.

Bach, D.R., Flandin, G., Friston, K.J., Dolan, R.J., 2009. Time-series analysis for rapid event-related skin conductance responses. J. Neurosci. Methods 184, 224–234.

Bach, D.R., Daunizeau, J., Friston, K.J., Dolan, R.J., 2010. Dynamic causal modelling of anticipatory skin conductance responses. Biol. Psychol. 85, 163–170.

Barry, R.J., 1990. Scoring criteria for response latency and habituation in electrodermal research: a study in the context of the orienting response. Psychophysiology 27, 94–100.

Barry, R.J., Feldmann, S., Gordon, E., Cocker, K.I., Rennie, C., 1993. Elicitation and habituation of the electrodermal orienting response in a short interstimulus interval paradigm. Int. J. Psychophysiol. 15, 247–253.

Benedek, M., Kaernbach, C., 2010a. A continuous measure of phasic electrodermal activity. J. Neurosci. Methods 190, 80–91.

Benedek, M., Kaernbach, C., 2010b. Decomposition of skin conductance data by means of nonnegative deconvolution. Psychophysiology 47, 647–658.

Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310.

Boucsein, W., 1992. Electrodermal Activity. Plenum Press, New York.

Dunsmoor, J.E., Mitroff, S.R., LaBar, K.S., 2009. Generalization of conditioned fear along a dimension of increasing fear intensity. Learn. Mem. 16, 460–469.

Edelberg, R., 1972. Electrodermal recovery rate, goal-orientation, and aversion. Psychophysiology 9, 512–520.

Gougeon, D.J., 2008. The desk reference of statistical quality methods. Choice: Current Reviews for Academic Libraries, 45 283–285.

Huff, N.C., Hernandez, J.A., Blanding, N.Q., LaBar, K.S., 2009. Delayed extinction attenuates conditioned fear renewal and spontaneous recovery in humans. Behav. Neurosci. 123, 834–843.

Jefferys, W.H., Berger, J.O., 1992. Ockham Razor and Bayesian-Analysis. Am. Sci. 80, 64–72.

Jeffreys, H., 1961. Theory of Probability, 3d ed. Clarendon Press, Oxford.

Kragel, P.A., LaBar, K.S., 2013. Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions. Emotion 13, 681–690.

Levinson, D.F., Edelberg, R., 1985. Scoring criteria for response latency and habituation in electrodermal research — a critique. Psychophysiology 22, 417–426.

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. Int. J. Psychophysiol. 25, 97–109.

Ludbrook, J., 2010. Confidence in Altman–Bland plots: a critical review of the method of differences. Clin. Exp. Pharmacol. Physiol. 37, 143–149.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1, 30–46.

Morey, R.A., Petty, C.M., Xu, Y., Hayes, J.P., Wagner II, H.R., Lewis, D.V., LaBar, K.S., Styner, M., McCarthy, G., 2009. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. Neuroimage 45, 855–866.

Schwarz, G., 1978. Estimating dimension of a model. Ann. Stat. 6, 461–464.

Storm, H., Fremming, A., Odegaard, S., Martinsen, O.G., Morkrid, L., 2000. The development of a software program for analyzing spontaneous and externally elicited skin conductance changes in infants and adults. Clin. Neurophysiol. 111, 1889–1898.

Thomas, L.A., LaBar, K.S., 2008. Fear relevancy, strategy use, and probabilistic learning of cue-outcome associations. Learn. Mem. 15, 777–784.

Trosiener, H.-J., Kayser, J., 1993. Analysis and interactive control of electrodermal and cardiac activity: a computer program for IBM-PC's and compatibles. J. Psychophysiol. 171.