Stability and reliability of error-related electromyography over the corrugator supercilii with increasing trials

Nathaniel Elkins-Brown[1], Blair Saunders[1], Frank He[1], & Michael Inzlicht[1,2]

[1]Department of Psychology, University of Toronto

[2]Rotman School of Management, University of Toronto

Please send correspondence to:
Nathaniel Elkins-Brown
University of Toronto, Department of Psychology
1265 Military Trail
Toronto, Ontario M1C 1A4, Canada
Email: nat.brown@mail.utoronto.ca
Telephone: 647-710-4950

Abstract

EMG activity over the *corrugator supercilii* (cEMG), the primary facial muscle involved in negative emotions, is increased during the commission of errors on speeded reaction-time tasks. In the present paper, data from two previously published studies were re-analyzed to investigate the reliability and stability of error-related, correct-related, and difference cEMG across increasing numbers of trials. We found that for a modified go/no-go and a flanker task, error-related cEMG was highly stable and reliable in 14 trials, and correct-related cEMG between 56 and 82 trials, respectively. Given the typical number of trials used in studies of cognitive control, these findings suggest that many investigations of error monitoring are already sufficient to obtain acceptable error- and correct-related cEMG signals. Error-related cEMG activity is relatively easy to measure, and as such, it shows great promise for future research investigating in the cognitive and affective mechanisms of error monitoring.

Key words: stability; reliability; corrugator; EMG; error monitoring

Error monitoring encompasses the physiological, psychological, and behavioral processes that respond to performance errors during goal-directed behavior. The ability to continually monitor for errors in performance is critical for behaviors like driving vehicles, operating machinery, and performing surgery, where errors have hazardous consequences.

Since error-related ERPs were discovered more than two decades ago, there has been rapid growth in psychophysiological research on error monitoring. As an example of this, one of the original papers on neural error monitoring, "A neural system for error detection and compensation" (Gehring, Goss, Coles, Meyer, & Donchin, 1993), remains at present the third most cited paper in the history of *Psychological Science*[1]. Much of the ensuant literature has focused on the error-related negativity (ERN/Ne), the error positivity (Pe), and error-related engagement of the peripheral nervous system.

More recently, researchers have also found error-related increases in electromyographic activity over the *corrugator supercilii* (cEMG), the principle facial muscle involved in frowning. As this muscle has been historically associated with negative affect, the exertion of effort, and goal obstruction, its error-related increases have great potential in illuminating the cognitive and emotional underpinnings of error monitoring and their contributions to cognitive control. However, we know very little about the psychometric properties of this error-related corrugator response, and an assessment of its stability and reliability should precede its measurement in future studies. Here, we investigated error-related cEMG to determine its stability and reliability across increasing numbers of trials.

The Corrugator Supercilii

The corrugator is a small, narrow, pyramidal muscle located on the medial end of each eyebrow that primarily functions to draw the eyebrows down and inward. As such, the

[1] As assessed by Google Scholar as of March 2017.

corrugator supercilii is the principle muscle involved in frowning (Janis et al., 2007). Upper facial muscles like the corrugator are innervated by specific subdivisions of the facial motor nucleus, which receive projections primarily from the supplementary motor and rostral cingulate motor cortices (Morecraft, Stillwell-Morecraft, & Rossing, 2004). These cortical regions are thought to be a key anatomic entry point for both subcortical and prefrontal inputs into the cortical motor system, including those from the medial prefrontal and orbitofrontal cortices, basal ganglia, amygdala, thalamus, hypothalamus, and other regions (Damasio, 1994; Morecraft et al., 2007; Morecraft & Van Hoesen, 1993; 1998; 2003). This suggests that contraction of the corrugator may reflect the downstream activity of a wide variety of different cognitive and affective processes mediated by frontal and subcortical networks.

Accordingly, facial EMG over the corrugator muscle in humans has been previously associated with the experience and expression of negative affect (Cacioppo, Petty, Losch, & Kim, 1986; Lang, Greenwald, Bradley, & Hamm, 1993; Larsen, Norris, & Cacioppo, 2003), the exertion of mental and physical effort (van Boxtel & Jessurun, 1993; de Morree & Marcora, 2012), the perception of goal obstructions (Pope & Smith, 1994; Schacht, Nigbur, & Sommer, 2009), and the encoding of cognitively disfluent stimuli (Topolinski, Likowski, Weyers, & Strack, 2009; Topolinski & Strack, 2015). While this array of constructs may at first seem challenging to reconcile under any single theory of corrugator reactivity, there is considerable overlap between them. For example, it has been argued that mental effort itself is aversive (Botvinick, 2007; Inzlicht, Bartholow, & Hirsh, 2015; Kool, McGuire, Rosen, & Botvinick, 2010), and that affect is an integral component of processing fluency and perceptual predictions (Chetverikov & Kristjánsson, 2016; Topolinski, Erle, & Reber, 2015). It may be that corrugator is specifically responding to changes in negative affect that accompany effort and processing

fluency. However, prolonged corrugator activity can co-occur with reductions in skin conductance, pupil diameter, and start-blink response (Schacht, Nigbur, & Sommer, 2009; Schacht, Dimigen, & Sommer, 2010), which are all concomitant measures of arousal or affect. Thus, it remains unclear what precise cognitive or affective processes lead to changes in EMG activity over the corrugator, or whether diverse processes uniquely contribute to such changes.

Corrugator and Monitoring

Recently, researchers have found engagement of the corrugator supercilii in response to errors during inhibitory control tasks, and that these error-related increases were positively associated with behavioral (Lindström et al., 2013) and neurophysiological (Elkins-Brown, Saunders, & Inzlicht, 2016) markers of error monitoring.

Lindström et al. (2013) found that error-related facial EMG over the corrugator shared a number of physiological and behavioral characteristics with the ERN. Using a modified Go/No-Go task, the authors found that error-related cEMG was increased between 0 and 100 ms of a response, was positively correlated with the slowing of reaction times following errors (Gehring et al., 1993), and was sensitive to perceived punishment risk (i.e., the anticipation of electric shocks). The authors interpreted these findings as evidence that cEMG indexes avoidance-motivated control, based on the theory that anterior mid-cingulate cortex (aMCC) integrates information about negative affect, pain, and cognitive control to promote aversively-motivated behavior (Shackman et al. 2012).

Elkins-Brown et al. (2016) followed up this study by concurrently measuring facial EMG and EEG during an inhibitory control task with performance-contingent punishment. Although the punishment manipulation did not influence any behavioral or physiological measures, they found error-related increases in cEMG between 0 and 300 ms, with the largest increases between

0 and 100 ms. Both between- and within-subjects, the later part of this signal was related to increased Pe amplitudes. Larger ERNs also predicted greater error-related cEMG, but this effect did not survive correction for multiple comparisons. The authors interpreted these findings as evidence that prolonged error-related cEMG might signify orienting or error awareness, or represent processes that contribute to such orienting or error awareness.

These two studies provide a compelling argument for the continued investigation of error-related cEMG in psychophysiological research. Despite differences in methodology and greater variability in EMG, error-related cEMG emerged as significantly different from its correct-related counterpart in both studies. Given cEMG's putative neural architecture, the frequency of its association with cognitive and affective variables, and the similarity of its error-related form to ERPs, error-related cEMG has great potential to be a useful psychophysiological correlate of error monitoring in experimental and clinical research. Like other early response-related ERPs, error-related cEMG could reflect a number of different internally-generated monitoring responses to errors, such as negative affect or defensive reactivity (Weinberg, Riesel, & Hajcak, 2012), reward prediction error (Holroyd & Coles, 2002), motor irridation from response conflict (de Morree & Marcora, 2010), and so forth.

Before researchers begin exploring the functional significance of error-related cEMG, it would be prudent to first assess the signal's psychometric properties. Fifteen years passed between the discovery of the ERN and the first formal investigation of its stability and reliability (Olvet & Hajcak, 2009), a period of time in which the minimum number of trials necessary for a satisfactory signal remained unclear. Because EMG has higher inter- and intra-individual variability compared to EEG (Gerdle, Karlsson, Day, & Djupsjöbacka, 1999; Mathiassen, Winkel, & Hagg, 1995), such a time-lag might be particularly inappropriate for studies involving

cEMG. If error-related cEMG requires substantially more trials than error-related ERPs, knowledge of a precise minimum will be essential for studies involving this measure. This is especially true in the case of clinical studies, which are often carried out in sub-optimal conditions, under tight methodological constraints, and with limited samples sizes (Marco-Pallares, Cucurell, Munte, Strien, & Rodriguez-Fornells, 2011).

The Current Study

Here, we investigated two data sets (N > 45) to assess the reliability and stability of error-related cEMG in two common cognitive control paradigms: A modified, two-choice go/no-go task (Study 1) and the classic Eriksen flanker task (Study 2; Eriksen & Eriksen, 1974). We used existing methods from previous investigations (i.e. Cohen & Polich, 1997; Kaye, Bradford, & Curtin, 2016; Marco-Pallares et al., 2011; Meyer, Riesel, & Proudfit, 2013; Moran, Jendrusina, & Moser, 2013; Olvet & Hajcak, 2009) to assess the stability and reliability of cEMG in both studies. First, to explore the stability of the mean EMG signal, we randomly sampled and averaged even sets of error and correct trials (i.e., 2, 4, 6, up to 18) for each participant 500 times, and plotted the grand means of these resampled sets. Next, to assess the internal reliability of the signal, we calculated the split-half reliabilities of error and correct trial averages—from 2 to 18 trials for errors, and from 2 to 100 trials for corrects—and graphed the grand mean reliabilities of 100 resampled sets. Lastly, to assess the quality of the signal, we calculated the signal to noise ratios (SNRs) of error and correct averages—again from 2 to 18 trials for errors and from 2 to 100 trials for corrects—and graphed the grand means of the SNRs of 100 resampled sets. We also accompanied all error and correct data with their corresponding difference wave measures. Using this variety of analytic methods, we sought to characterize the stability, reliability, and signal quality of error- and correct-related cEMG in two common and

established cognitive control paradigms.

Method

All measures, conditions, data exclusions, and the determination of sample size are reported below, as recommended by the Open Science Framework (see osf.io/hadz3). Data and materials are available for Study 1 at (osf.io/c9tkd), and for Study 2 at (osf.io/mtrys).

Participants

Response-related data was extracted and reanalyzed from two studies already published from our laboratory at the University of Toronto Scarborough (Study 1: Elkins-Brown et al., 2016; Study 2: Saunders et al., 2015). Seventy-two students (47 female, mean age = 19.1 years, *SD* = 1.8) participated in Study 1 for course credit in their psychology class. Sixty-one students (32 females; mean age = 18.4, *SD* = 1.4) participated in Study 2, also for course credit. Four participants in Study 1 and one participant in Study 2 were excluded because of equipment malfunction; one participant in Study 2 was excluded for misunderstanding directions. For both studies, only participants with at least 18 errors were included in the analyses; this cutoff number was chosen in order to keep the final sample of participants as large and varied as possible, while retaining enough trials for error-related cEMG to stabilize and become reliable. As the ERN and Pe are thought to become reliable and stable much earlier than 18 trials, this number for cEMG is reasonable. After removing outliers, this left us with a sample of 49 participants for Study 1, and 56 participants for Study 2.

The sample size for Study 1 was not explicitly determined *a priori*; the investigators collected data until the end of the undergraduate term, with the expectation that at least 60 participants were a conservative estimate of sufficient power to test the hypotheses for the original study. The sample size in Study 2 was determined *a priori* based on the sample and

effect size of a previous study relevant to the original hypotheses (DeSteno, Li, Dickens, & Lerner, 2014). Importantly, the sample sizes for both of sets of data analyzed in the present study are sufficiently powered for conducting within-subject psychometric analyses, given the effect sizes for error-related cEMG results observed in previous studies of error-related cEMG (Elkins-Brown et al., 2016; Lindström et al., 2013).

Procedure

Study 1. Electrophysiological data was measured via facial EMG while participants completed a modified Go/No-Go task (see Figure 1). In this task, participants responded using keys on a DirectIN PCB keyboard (Empirisoft, New York, NY) in response to two stimuli, the letters "M" and "W". The presentation probability for each stimulus was asymmetric, giving a correspondingly asymmetric response ratio of 80:20. This manipulation was based on the standard Go/No-Go task (Simmonds, Pekar, & Mostofsky, 2008), where the high probability target induces a pre-potent tendency to respond, which has to be inhibited for low probability targets. Our task modifies the task demands of this manipulation by requiring participants to make an alternative and infrequent response to low-probability stimuli, rather than withholding a response on the "no-go" trial.

On each trial, participants were required to press the "Z" key when they saw the frequent, "M" stimulus (low-conflict), and to press the "/" key when they saw the infrequent, "W" stimulus (high-conflict). Participants were encouraged to respond both quickly and accurately. The low- and high-conflict stimuli were presented in a yellow or purple font depending upon punishment condition (see below). Each trial began with a fixation cross presented for 600ms, followed by either a low- or high-conflict stimulus that remained on screen until the participant responded or until a maximum of 1500 ms had passed. Correct responses

were followed by 400ms of a blank screen before the fixation cross of the next trial appeared, providing a total response to target interval of 1000 ms. Incorrect responses had a 50% chance of producing a 3500 Hz punishment tone. For incorrect responses that were punished, the incorrect key press was followed by 1000 ms of a blank screen, and then 1000 ms of the tone. Tones were followed by an additional delay of 1500 ms of blank screen before the onset of the next trial, producing a total response to target interval of 2500 ms for punished, incorrect trials. For the 50% of incorrect key presses that were not punished, the trial continued as if the participant had pressed the correct key, producing a total response to target interval of 1000 ms for unpunished, incorrect trials. Tones were presented from desktop speakers located approximately 3ft in front of the participant. The intensity of the sound was manipulated blockwise and within participants, where the punishment condition corresponded to a volume of 95 dB (roughly equivalent to a motorcycle engine at 5 m) and the unpunished condition to a volume of 20 dB (roughly equivalent to whispered conversation). Participants first completed a practice block of 20 trials, and then completed 12 normal blocks of 70 trials each (840 in total). Of these 12 blocks, 6 were punished and 6 were unpunished, and participants alternated between punished and unpunished blocks in counterbalanced sets of three.

The punishment manipulation had no effect on any physiological or behavioural measure in the previous study. This includes error rates, reaction times, ERN amplitudes, Pe amplitudes, or cEMG amplitudes across the entire epoch including the baseline (-200 to -100 ms). Thus, we do not discuss it further.

Study 2. Facial EMG was measured while participants completed a Flanker task (see Figure 2). In this task, participants identified the central letter in a five-letter string. All flanker arrays consisted of the letters H and S. Participants were instructed to press the “Z” key on their

keyboard if the target was S and the "/" key if the target was H. Arrays contained either flanker letters that were identical with (e.g., "HHHHH") or conflicted with ("SSHSS") the target letter.

Trials commenced with a central fixation cross lasting 250 ms. Flanker letters then appeared for 100 ms prior to the onset of the central target stimulus, and remained present with the target stimulus until participants responded with a key or for a maximum time of 1500 ms. This stimulus-onset asynchrony was used to increase the amount of response priming caused by the flanker stimuli relative to the target. Participant responses were followed by a blank screen lasting 1000 ms. Participants were encouraged to respond to target stimuli both quickly and accurately. Participants first completed 20 practice trials, followed by 500 experimental trials divided equally into 5 blocks. Compatible and incompatible flanker trials were presented with equal probability, and participants were allowed to take short, self-paced breaks between blocks.

Immediately following the flanker task, participants performed an autobiographical recall task that aimed to induce either feelings of happiness or gratitude (see DeSteno et al., 2014). However, because this induction task had no influence on any physiological measure, we did not account for it in the current analyses (see Saunders et al., 2015 for details). Following this autobiographical recall task, participants completed another post-induction Flanker task that followed a procedure identical to the pre-induction Flanker task, with 500 experimental trials divided equally into 5 blocks of 100 trials. Data from both the pre- and post-induction flanker tasks were combined into a single set of data.

Electrophysiological and signal pre-processing

For both Study 1 and Study 2, continuous EMG activity over the left corrugator supercilii muscle was recorded with two miniature EMG Ag/AgCl electrodes (Cacioppo et al., 1986). This signal was amplified using an ANT Refa8 TMSi (Advanced Neuro Technology, Enschede, The

Netherlands) device, and grounded with an electrode on the forehead. Of the two electrodes placed over the corrugator supercilii, the one placed more laterally over the eyebrow served as a reference for the one placed more medially (see Cacioppo, Tassinary, & Berntson, 2007). Impedances were brought below 5 kΩ before all recordings began. Impedances were not checked again over the course of either experiment for Study 1 or Study 2. For Study 1, recordings were digitized for the first 19 participants at 512 Hz using Advanced Source Analysis (ASA) 4.7.11 software. The sampling rate was increased to 1024 Hz for the remaining 53 participants in order to obtain greater temporal resolution for the EMG signal. Prior to filtering and analysis, the data for the first 19 participants were up-sampled offline to 1024 Hz using a spline interpolation procedure in Brain Vision Analyzer 2.0 (Brain Products GmbH, Gilching, Germany).

In Brain Vision Analyzer, the raw EMG signal was filtered offline using a 60 Hz notch filter, a 28-250 Hz IIR bandpass for the first 19 participants, a 28-500 Hz IIR bandpass for the remaining participants. To ensure that the loss of data in the 250-500 Hz range for the first 19 participants did not strongly influence their data, we compared data of the remaining 53 participants at both 250 and 500 Hz cutoffs. We found only small changes in amplitude (< 0.3 μV for trials and < 0.1 μV for averages) between these approaches, suggesting that the contributions of activity from 250-500 Hz are small. We also reran all analyses conducted in the present study after dropping the first 19 participants, and found only slightly lower split-half reliability compared to that of the full sample. Thus, the data from these two groups of participants are likely comparable for the purposes of the present study.

All data were rectified and then smoothed using a moving average procedure with a time constant of 20ms (Cacioppo et al., 2007). Automatic procedures were then used to reject EMG artifacts with voltages above 100 μV and below -100 μV. For Study 1, 21 participants had less

than 0.05% of their data removed this way, and 3 participants had less than 2% of their data removed this way. For Study 2, 25 participants had less than 0.05% of their data removed this way, and 5 participants had less than 3% of their data removed this way.

Data was then divided into trial epochs commencing 200 ms before the response and lasting up to 1000 ms after the response. Epochs were then baseline corrected by subtracting average voltages 200 ms to 100 ms before the response.

In MATLAB (R2015a, Version 8.5), all data points in every trial above 10 μV and below -10 μV were then removed as outliers, and any trial with more than 25% of its data removed this way was removed entirely from that participant's data. This range of exclusion was chosen because EMG over the corrugator is unlikely to show meaningful changes of such magnitude in under a second. For Study 1, participants had on average 8.23 (*SD* = 12.87) trials removed this way, excluding one participant who had 265 trials removed this way. For Study 2, participants had on average 20.84 (*SD* = 33.64) trials removed this way. Only 1 participant had their data entirely removed from Study 1 because of outliers.

For all analyses except split-half reliability, raw trial data was then averaged within-participants and per trial type, producing correct and error epoch averages.

Because of interindividual differences in facial perspiration and temperature, in the number and orientation of motor units in facial muscles, and in blood flow and the amount of tissue between electrodes and muscle fibres, raw EMG averages tend to have high interindividual variability (Halaki & Ginn, 2012). To reduce this variability, researchers commonly standardize (i.e., z-transform) data across epochs (Lindström et al., 2013; Schacht et al., 2009; 2010) or in reference to a maximal voluntary contraction (van Boxtel, 2010; Cacioppo, et al., 2007; Halaki et al., 2012). We standardized data across the -200 to 1000 ms epoch separately for error and

correct conditions. To standardize data, all raw data points in each error and correct average for a participant were subtracted by the corresponding raw epoch mean of that participant's average separately for error and correct trials, and then divided by the standard deviation of the corresponding epoch mean of that participant's average. Because standardization can create substantial changes in the morphology and magnitude of EMG activity, all primary analyses conducted in the present study were also performed on raw data, and these results are available in the Supplementary Materials for comparison.

It should be noted that when standardizing averages using this approach, Z-scores will be larger when standard deviations are lower for the entire epoch, and smaller when standard deviations are higher for the entire epoch. This means that standardized averages consisting of different numbers of trials can have different magnitudes even when their raw equivalents are the same, because standard deviations are negatively correlated with trial number. Thus, until the number of trials added to an unstandardized average no longer reduces its standard deviation, standardized averages will systemically underestimate the grand mean.

For the split-half reliability analysis, individual trials—rather than averages—were drawn separately from error and correct conditions and standardized using the same procedure described above. This trial level procedure was used for the split-half reliability analysis in order to approximate the standardization of averages used for all other data in the present study, because standardizing over split-half averages does not alter their split-half reliability. When averaged, unstandardized trials have comparable morphology but reduced magnitudes compared to standardized averages (see the Supplementary Materials).

For all analyses, cEMG was operationalized as the mean EMG activity between 0 and 100 ms post-response. To the extent that cEMG putatively represents an internally-generated

error monitoring signal—rather than a stimulus-related or response-insensitive somatic reflex—this operationalization is theoretically appropriate. This choice of time bin is supported by consistent findings from the two previous studies that explored error-related cEMG in the context of error-related ERPs like the ERN (Lindström et al., 2013; Elkins-Brown et al., 2016). However, researchers should consider investigating other time-windows of error-related cEMG in the future, such as the decrease in activity that appears to start around 200 ms post-response.

Signal Assessment and Analysis

Prior to investigating the psychometric properties of cEMG, we first conducted a number of comparisons to ensure that cEMG data across all participants was uniquely increased to errors and not correct responses, and to determine whether cEMG habituated or sensitized throughout the task. To compare error and correct averages, the MIXED function in SPSS (Version 23) was used to calculate Type III analyses of variance, effect-coded such that correct responses = -1, and error responses = 1. These models used a restricted maximum likelihood approach for fitting, and an unstructured correlations covariance matrix to estimate a random-intercept for the fixed effect of trial type (error/correct). The effect size for these calculations was denoted with semipartial R^2 (R^2_β; Edwards, Muller, Wolfinger, Qaqish, & Schabenberger, 2008).

To compare cEMG responses across time and trial type, we divided all correct and error trials for each participant into four equal sections, representing the first, second, third, and fourth quarters of the task. These sections were then averaged and standardized within each quarter and participant. Two repeated-measures factorial ANOVAs were conducted to see whether the mean cEMG increased or decreased over time. One factorial ANOVA was conducted for each study, with means predicted by the within-subjects factor of time (first, second, third, and fourth quadrant) and trial type (error vs. correct). Effect sizes for these calculations were denoted using

partial η^2 (η_p^2; Cohen, 1973). For all ANOVAs in the present study in which sphericity was violated, a Greenhouse-Gessier correction was used when Mauchley's epsilon was lower than or equal to .75. If epsilon was higher than .75, a Huynh-Felt correction was used. Bonferroni corrections were used for all pairwise comparisons.

Assessment of Stability. Next, we explored the stability of our signal using a resampling and averaging procedure. This resampling approach was used throughout the present investigation in order to reduce the high variability of small samples and of singular sampling (Abelson, 1995). We operationalized stability as the point in which trial means approximate the grand average, and the point when correct and error trials become consistently distinguished from one another. For each participant, we randomly sampled and averaged trials to create 500[2] trial averages for each set of even trial numbers between 2 and 18 trials. This created 9 sets of 500 trial averages for each participant, such that each set contained 500 averages made up of 2 trials each, or 4 trials each, or 6 trials each, and so forth up to 18 trials each. These 500 trial averages within each set and participant were then standardized, and the average activity from 0-100 ms was extracted from each of the 500 standardized averages. To create a set of data for difference waves, correct time bins were then subtracted from error time bins. Next, these 500 were then averaged together within participant and set, producing 9 average values (and 9 standard errors) for each participant and condition.

We then plotted the grand means of these 9 values averaged across participants, and compared error and correct data using 95% confidence intervals based on the within-subject

[2] We resampled only 500 times for the assessment of stability and 100 times for the assessment of reliability and signal power in order to reduce the calculation time in MATLAB. For all findings in the present study, results did not change very much past 50 resamples, and were often identical to two or three decimal places by 100 resamples.

standard deviations for each study. Difference grand averages for both studies were also plotted with 95% confidence intervals separately from error and correct data, for ease of visualization.

We expected that the magnitude of these grand averages would gradually increase for every cumulative set of trials, in line with the decreasing standard deviations of their constituent raw averages. This approach allowed us to see if higher numbers of standardized trial numbers were sufficient to approximate the actual grand average values for correct- and error-related cEMG, to see if the standard deviations of cEMG stabilized within 18 trials, and to see when correct- and error-related data could be statistically distinguished from one another.

Assessment of Reliability. To quantify the internal reliability of cEMG across increasing numbers of trials, we calculated split-half reliabilities of randomly sampled sets of averaged trials. For error and correct data, split-half reliabilities were calculated by splitting sets of randomly sampled, standardized trials into odd and even halves across all participants, averaging the halves separately, and then calculating the Pearson correlation between them. The resulting correlations ($r_{odd\text{-}even}$) were then subjected to a Spearman-Brown correction in order to compensate for the reduced number of trials per correlation due to splitting, such that $r_{SB} = 2 * (r_{odd\text{-}even}) / (1 + r_{odd\text{-}even})$.

The above calculation of reliability estimates were repeated using a resampling procedure that was similar to the one used for trial averages. For the error part of the reliability analysis, 9 reliability estimates (even trials from 2 to 18) were created for the entire sample of participants. Because of the low reliability and SNRs of correct trials, 50 these same reliability estimates were instead generated for correct reliability data, corresponding to sets of even trial numbers from 2 to 100 trials. Reliability estimates across the whole sample were then generated 100 times,

producing 100 reliability estimates for each set of trial numbers from 2 to 18 (for error trials) or from 2 to 100 (for correct trials). The grand means of these estimates values were then graphed.

After determining the minimum number of trials necessary for highly reliable error- and correct-related signals, we also calculated the internal reliability of the difference wave using a specific subtraction procedure. This procedure was used to provide a conservative approximation of the difference wave data that would be typical in an experiment using a modified Go/No-Go or flanker tasks, where there are often many more correct than error trials. From increasing sets odd and even error trial averages from 2 to 18, we subtracted the odd and even trial averages of the correct number of trials that had been determined to be highly reliable in the previous analysis. In the case of Study 1, this number was 56 correct trials, and in the case of Study 2, this number was 82 trials. These subtractions were then corrected, repeated 100 times, and then graphed.

Assessment of Signal Power. Lastly, to quantify the signal power of cEMG across increasing numbers of trials, we calculated signal-to-noise ratios (SNR) for randomly sampled sets of averaged trials. SNRs were estimated by dividing the root mean square of a signal's average by its standard deviation (Marco-Pallares et al., 2011). To calculate the SNR of the difference wave, we used a procedure that was comparable to the analysis of difference wave reliability. We subtracted a 56-trial correct average from increasing sets of modified Go/No-Go error averages, and an 82-trial correct average from increasing sets of flanker error averages. Error and correct data were standardized prior to subtraction. The root mean squares of these difference scores were then divided by their standard deviations.

The resampling procedure used for SNRs was similar to both that of trial averages and that of the reliability estimates. 100 SNRs were generated for each set of even trial numbers for

each participant, and these 100 SNRs were then averaged to produce 9 error average values and 50 correct average values (along with their standard errors) within each participant. We then plotted the grand means of these average values across participants, and compared error and correct data using 95% within-subject confidence intervals.

Results

Grand Averages

Differences in the grand averages of cEMG between errors and correct trials are presented in Figure 3. For Study 1, cEMG was larger for errors (M = 0.47, SE = 0.12) than for correct responses (M = -0.09, SE = 0.10) (b = -0.56, SE = 0.15), $F(1, 60.518) = 13.884$, $p < .001$, $R^2_\beta = 0.18$. Similarly for Study 2, cEMG was larger for errors (M = 0.72, SE = 0.09) than for correct responses (M = 0.13, SE = 0.09) (b = -0.59, SE = 0.10), $F(1, 57.000) = 38.219$, $p < .001$, $R^2_\beta = 0.40$.

Stability of Averages

Grand averages of each quarter are presented in Figure 4. For both Study 1 and Study 2, there were no significant effects of time or significant time*trial type interactions, all Fs < 2, ps > 0.10, suggesting that error- and correct-related cEMG likely does not habituate or sensitize over time. There were, however, significant main effects of trial type, indicating that error trials were larger than correct trials for both Study 1, $F(1,48) = 6.529$, $p = 0.012$, $\eta_p^2 = 0.12$, and Study 2, $F(1, 55) = 13.713$, $p < 0.001$, $\eta_p^2 = 0.20$.

Randomly sampled averages from 2 trials to 18 trials, along with the grand average of the entire sample, are presented in Figure 5. For error data in both Study 1 and Study 2, standardized error averages increased as sets of trials increased, reflecting the decreasing standard deviations of cumulative sets of trials. For Study 1, error data emerged as significantly larger than correct

data in as few as 4 trials, and for Study 2, error data emerged as significantly larger than correct data in as few as 2 trials. For both Study 1 and Study 2, difference data was significantly larger than zero at only 2 trials. This suggests that error and correct data is distinguishable even at just 2 to 4 trials per participant.

However, for both Study 1 and Study 2, all error trial averages underestimated the grand average; no 95% within-subject confidence interval for each set of trial averages crossed over with the actual grand average. More than 18 trials in a normalized average are likely necessary for a given participant's standardized average to accurately approximate the grand mean of all participants for both inhibitory control and flanker tasks. This means that participants who make less than 18 errors in either task will still have error averages with magnitudes that are lower than participants with more than 18 errors. Conversely, the grand average of all participants' correct data for both Study 1 and Study 2 was adequately approximated in as few as 2 trials. When subtracting correct averages from error averages, 18 trials was sufficient to approximate the grand difference wave in the flanker task, but 18 trials was still not sufficient to approximate the grand difference wave in the Go/No-Go task.

Split-half Reliability

Spearman-Brown split-half reliability estimates for correct, error, and difference data are depicted in Figure 6. Reliability values exceeding .90 are thought to indicate excellent reliability, between .70 and .90 to indicate high reliability, between .50 and .70 to indicate moderate reliability, and below .50 to indicate low reliability (Hinton, Brownlow, McMurray, & Cozen, 2004). Figure 6 suggests that high reliability for error-related cEMG was obtained by 14 trials for both Study 1 (r = 0.73, SE = 0.06) and Study 2 (r = 0.71, SE = 0.07). For correct-related cEMG, high reliability was obtained in 56 trials for Study 1 (r = 0.72, SE = 0.07) and 82 trials for Study

2 (r = 0.71, SE = 0.07). When subtracting reliable correct averages from error ones, high reliability of the difference wave was obtained in only 10 trials for Study 1 (r = 0.71, SE = 0.07), but did not quite reach high reliability by 18 trials for Study 2 (r = 0.67, SE = 0.08).

These findings show that cEMG error data can become highly reliable within 14 trials, but excellent reliability cannot be obtained in 18 trials. At least 56 trials for a modified Go/No-Go task and 82 trials for a flanker task are necessary to obtain a highly reliable correct-related cEMG signal. When subtracting a correct trial average of the lowest acceptable reliability from increasing sets of error trials, the reliability of the modified Go/No-Go difference wave is slightly higher than that of the error data, but slightly lower than that of the error data for the Flanker task. These difference wave findings likely reflect differences in the reliability of the correct data between the two tasks, rather than differences in the error data.

Signal-to-Noise Ratios (SNRs)

SNR estimates for sets of trial numbers of correct, error, and difference data are depicted in Figure 7. For each study and condition, we conducted a one-way repeated measures ANOVA of SNRs with trial number as a factor to examine the differences between SNRs as a function of trial number. For all significant main effects of trial number, pairwise comparisons were then used to compare the means of different trial numbers. Because these comparisons were corrected using a Bonferroni procedure, all significant differences between trial numbers should be considered highly conservative estimates.

There was a significant effect of trial number for both for Study 1 and Study 2 error data (both ps < 0.001, both Fs > 18). Pairwise comparisons for sets of trial numbers revealed that for Study 1, there were significant differences in SNRs every 2 to 4 trials, while for Study 2, there were significant differences for every adjacent set of trials. This suggests that 2 to 4 trials is a

conservative estimate of the number of trials needed to increase SNRs in a modified Go/No-Go task, while only 2 trials is a similarly conservative estimate for gains in SNRs in the flanker task. The range of grand mean SNRs in Study 1 error data was from 1.28 to 1.68, while for Study 2 was from 1.27 to 1.78.

In contrast to error data, increases in signal power for correct data were much more modest for increasing numbers of trial sets, although still clearly linear. There were significant effects of trial number for both Study 1 and Study 2 correct data (both ps < 0.001, both Fs > 6). Pairwise comparisons for sets of trial numbers revealed that for both Study 1 and 2, gains in SNRs across trial numbers were highly inconsistent, requiring anywhere from 22 to 46 more trials in order to obtain significant differences. This suggests that 46 trials are a conservative estimate of the number of trials needed to increase SNRs in either a modified Go/No-Go or flanker tasks. The range of grand mean SNR correct data in Study 1 was from 1.18 to 1.49, while for Study 2 was from 1.16 to 1.41.

When subtracting reliable correct averages from increasing sets of error trials, difference cEMG had increases in SNRs that were worse than error trials and better than correct trials. There was a significant effect of trial number for both Study 1 and Study 2 (all ps < 0.005, all Fs > 8). For Study 1, more than 18 trials were necessary for significant gains in SNR, after correcting for multiple comparisons. For Study 2, there were significant gains in SNRs every 4 to 8 trials, after correcting for multiple comparisons. The range of grand mean SNR difference data in Study 1 was from 1.28 to 1.68 and for Study 2 was from 1.27 to 1.78.

These results suggest that there are meaningful increases in SNRs for every 2 to 4 error trials added to an average for both tasks, and for every 46 or fewer correct trials added to an average for both tasks. When reliable correct data is subtracted from error data, meaningful

increases in SNRs may require more than 18 trials for a modified Go/No-Go task, but only 4 to 8 trials for a flanker task. When comparing correct and error trials, correct data at 100 trials still does not obtain the SNRs of highly reliable error data at 14 trials. It is less likely that this difference is due to greater noise in the correct signal compared to the error signal—as the standard deviations of each condition are relatively similar at the same number of trials—but is likely the result of substantially lower signal power for correct trials compared to error trials. For all data, increases in signal power were linear, suggesting that further gains in power are probable for trial numbers that go beyond the maximum number of trials investigated in the present study.

In summary, our analyses suggest that error-related cEMG is stable and reliable psychophysiological correlate of performance monitoring. For both tasks investigated in the present study, 14 error trials were sufficient to obtain a highly reliable error-related signal that could be distinguished from correct data, but 18 error trials still underestimated the grand mean. For correct-related data, 56 trials were necessary for the modified Go/No-Go and 82 trials were necessary for the flanker task in order to obtain high reliability. For difference data, 18 trials were sufficient to approximate the grand difference wave of the flanker task, but were still not sufficient to approximate the grand difference wave of the modified Go/No-Go task. The SNRs of cEMG in both tasks were low, and it is likely that significant gains in signal power are still possible past the number of trials investigated in the present study.

Discussion

The goal of the present study was to characterize response-related cEMG across increasing numbers of trials in two different speeded response tasks, and to determine how many of those trials were necessary to obtain a stable and reliable signal. We found that the

psychometric properties of error-related cEMG permit it to be measured feasibly in most psychophysiological experiments, as only 14 error trials were necessary for high reliability in both the modified Go/No-Go and the flanker task, and correct and error trials could be consistently distinguished in only 2 to 4 trials. Standardized error trial averages still slightly underestimated the grand mean at 18 trials, however, meaning that there will still be slight differences in magnitude between participants with many errors and those with 18 or less. Additionally, although signal power for error-related cEMG was low in general, there were significant gains in SNR for every 2 to 4 trials added to the average, and it appeared that further gains in signal power were likely still possible past 18 trials.

In general, the reliability of error-related cEMG responses is comparable to response-locked ERP components, such as the ERN and Pe (Olvet & Hajcak, 2009; Pontifex et al., 2010), where between 6 and 8 trials produced components with high stability and reliability across different age groups and contexts. Error-related cEMG may only require slightly more trials to reach levels of stability and reliability that are comparable to the ERN and Pe. This implies that many study designs that assess error-related ERPs may be sufficient for stable and reliable error-related cEMG. However, in sub-optimal measuring conditions and in smaller samples, more than 18 error trials may be necessary in order to maximize signal power and obtain stable standard deviations across conditions.

In contrast to error-related responses, correct-related responses could approximate the grand mean in only 2 trials, but had lower SNRs and reliability. To obtain high reliability, the modified Go/No-Go task required 56 trials and the flanker task required 82 trials. Because both error- and correct-related cEMG have similar levels of noise, it is probable that EMG over the corrugator simply has low sensitivity to correct responses in general, at least within the 0-100 ms

time bin investigated in the present study. This interesting distinction between correct and error responses is not shared by other response-locked ERPs, where correct-related responses can sometimes have higher reliability than their error-related counterparts (e.g., Xu & Inzlicht, 2015).

Lastly, difference wave cEMG had psychometric properties between that of error- and correct-related responses. For both tasks, only 2 trials were necessary for the difference wave to be significantly greater than zero. For the Go/No-Go task, high reliability was obtained in only 10 trials, gains in SNRs across trial numbers from 2 to 18 were modest, and 18 trials was still not sufficient to approximate the difference of the grand mean. For the flanker task, high reliability was still not quite obtained in 18 trials, there were gains in SNRs every 4 to 6 trials, and 18 trials were sufficient to approximate the difference of the grand mean.

The future of error-related cEMG. Our findings open up new avenues for future error monitoring research. Given the corrugator's underlying neurophysiology and its historical association affect and cognition, error-related cEMG may reflect the downstream activation of a number of internally-generated monitoring signals that detect, evaluate, and evoke peripheral nervous system responses to errors. Researchers may be able to disentangle the functional sensitivity of error-related cEMG through a number of study design approaches, such as affective priming (e.g., Aarts, De Houwer, & Pourtois, 2012), within-participant measures of experience (Saunders, Milyavskaya, & Inzlicht, 2015), and inductions of emotion and effort (Codispoti, Ferrari, & Bradley, 2007; Codispoti, Mazzetti, & Bradley, 2009; Schüpbach, Gendolla, & Silvestrini, 2014). If error-related cEMG is also present in other error monitoring tasks—such as the Simon, Stroop, or stop-signal tasks—researchers will also have greater flexibility in assessing its functional significance and its potential to differentiate between competing accounts of error

monitoring processes. Lastly, because abnormal neural monitoring and altered facial responses are common to clinical populations (e.g., Weinberg et al., 2012; de Wied, van Boxtel, Zaalberg, Goudena, & Matthys, 2006), error-related cEMG may also have a place in psychopathological investigations of performance monitoring.

These findings should be considered in light of a number of limitations. Firstly, both of our studies were conducted in primarily healthy, young adults. As the psychometric properties of ERPs and the magnitude of EMG over the corrugator can be moderated by age (e.g., Marco-Pallares et al., 2011; Smith, Hillman, & Duley, 2005) and clinical status (e.g., Foti, Kotov, & Hajcak, 2013; Matzke, Herpetz, Berger, Fleischer, & Domes, 2013), our results may not generalize to other populations. Future researchers should consider investigating the psychometric properties of cEMG in younger, older, and clinical cohorts to assess the generalizability of our findings to different populations.

Secondly, in order to maintain a moderate sample size, our analysis of error-related cEMG only went up to 18 trials for error-related cEMG, and 100 trials for correct-related cEMG. Our data clearly show that significant gains in signal power are possible past 18 trials, and that more than 18 trials are needed for averages to adequately estimate the grand average. As future studies may require a signal with excellent psychometric properties in order to compensate for a small sample and suboptimal measurement conditions, future work should explore the psychometric properties of error-related cEMG with greater trial numbers. This is especially important when the number of trials or standard deviations across conditions or participants are likely to be unequal. Otherwise, comparisons across conditions will be hindered for standardized data, as was the case in the present study. Because it may be practically impossible to hold trial

number constant across conditions for certain paradigms, researchers should alternatively consider investigating other forms of standardization that are not confounded by trial number.

Thirdly, our analyses did not address the within-trial temporal stability, test-retest stability, or test-retest reliability of error-related cEMG, as has been done in previous studies (e.g., Larson, Baldwin, Good, & Fair, 2010; Weinberg & Hajcak, 2011). In a recent psychometric assessment of corrugator responses during threat-of-shock and picture-viewing paradigms, corrugator potentiation had either poor or adequate psychometric properties, respectively (Kaye et al., 2016). Although error-related cEMG may reflect a different process than cEMG modulations over several seconds in time, both suffer from the high inter- and intra-individual variability that characterize electromyographic measurements. As such, assessing these temporal metrics will be critically important for comprehensively evaluating whether error-related cEMG is a fundamentally reliable and stable signal. Because error-related cEMG appears to both increase and decrease over the course of the post-response epoch, researchers should also consider pinpointing the timing of these changes across conditions and in different tasks using mass univariate approaches (Groppe, Urbach, & Kutas, 2011; Maris & Oostenveld, 2007).

The data presented here constitute an exciting first step in the description of a novel error-related psychophysiological response. We find that this signal becomes stable and reliable within the range of trial numbers studied here, providing researchers with new directions to explore the physiological and psychological foundations of error monitoring.

References

Aarts, K., De Houwer, J., & Pourtois, G. (2013). Erroneous and Correct Actions Have a Difference Affective Valence: Evidence from ERPs. *Emotion*, *13*(5), 960-973. doi: 10.1037/a0032808

Abelson, R. P. (1995). Statistics as principled argument. Hillsdale, NJ: Erlbaum.

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience*, 7, 356–366.

van Boxtel, A. (2010). Facial EMG as a tool for inferring affective states. In *Proceedings of Measuring Behavior*. Wageningen: Noldus Information Technology.

van Boxtel, A., & Jessurun, M. (1993). Amplitude and bilateral coherency of facial and jaw-elevator EMG activity as an index of effort during a two-choice serial reaction task. *Psychophysiology*, *30*(6), 589-604. doi: 10.1111/j.1469-8986.1993.tb02085.x

Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology*, *50*(2), 260. doi: 10.1037/0022-3514.50.2.260

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (Eds.). (2007). *Handbook of Psychophysiology*. Cambridge, UK: Cambridge University Press.

Chetverikov, A., & Kristjánsson, Á (2016). On the joys of perceiving: Affect as feedback for perceptual predictions. *Acta Psychologica*, 169: 1-10. doi: 10.1016/j.actpsy.2016.05.005

Codispoti, M., Ferrari, V., & Bradley, M. M. (2007). Repetition and event-related potentials: distinguishing early and late processes in affective picture perception. *Journal of Cognitive Neuroscience, 19*(4), 577-586. doi: 10.1162/jocn.2007.19.4.577

Codispoti, M., Mazzetti, M., & Bradley, M. M. (2009). Unmasking emotion: Exposure duration and emotional engagement. *Psychophysiology, 46*(4), 731-738. doi: 10.1111/j.1469-8986.2009.00804.x

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33,* 107-112. doi: 10.1177/001316447303300111

Cohen, J., & Polich, J. (1997). On the number of trials needed for P300. *International Journal of Psychophysiology, 25*(3), 249-255. doi: 10.1016/s0167-8760(96)00743-x

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in quantitative methods for psychology*, *1*(1), 42-45. doi:10.20982/tqmp.01.1.p042

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam, 140-142.

DeSteno, D., Li, Y., Dickens, L., & Lerner, J. (2014). Gratitude: A Tool for Reducing Economic Impatience. *Psychological Science*, 25, 1262–1267. doi: 10.1177/0956797614529979.

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., & Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine*, 27, 6137–6157. doi: 10.1002/sim.3429

Elkins-Brown, N., Saunders, B., & Inzlicht, M. (2016). Error-related electromyographic activity over the corrugator supercilii is associated with neural performance monitoring. *Psychophysiology*, *53*(2), 159-170. doi: 10.1111/psyp.12556

Eriksen, B.A., & Eriksen, C.W. (1974). Effects of noise letters upon identification of a target letter in a non-search task. *Perception and Psychophysics*, 16, 143–149. doi: 10.3758/bf03203267.

Foti, D., Kotov, R., & Hajcak, G. (2013). Psychometric considerations in using error-related brain activity as a biomarker in psychotic disorders. *Journal of Abnormal Psychology*, *122*(2), 520. doi: 10.1037/a0032618

Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385-390. doi: 10.1111/j.1467-9280.1993.tb00586.x.

Gerdle, B., Karlsson, S., Day, S., & Djupsjöbacka, M. (1999). Acquisition, processing and analysis of the surface electromyogram. In U. Windhorst & H. Johansson (Eds.), *Modern Techniques in Neuroscience Research* (pp. 705-755). Heidelberg, Germany: Springer Berlin Heidelberg.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711-1725. doi: 10.1111/j.1469-8986.2011.01273.x

Halaki, M., & Ginn, K. (2012). Normalization of EMG Signals: To Normalize or Not to Normalize and What to Normalize to? In G.R. Naik (Ed.),*Computational Intelligence in Electromyography Analysis - A Perspective on Current Applications and Future Challenges*. Rijeka, Croatia: Intech. doi:10.5772/49957

Heller, A. S., Greischar, L. L., Honor, A., Anderle, M. J., & Davidson, R. J. (2011). Simultaneous acquisition of corrugator electromyography and functional magnetic resonance imaging: A new method for objectively measuring affect and neural activity concurrently. *Neuroimage*, *58*(3), 930-934. doi: 10.1016/j.neuroimage.2011.06.057

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679. doi:10.1037/0033-295x.109.4.679

Janis, J. E., Ghavami, A., Lemmon, J. A., Leedy, J. E., & Guyuron, B. (2007). Anatomy of the corrugator supercilii muscle: Part I. Corrugator topography. *Plastic and Reconstructive Surgery, 120*(6), 1647-1653. doi: 10.1097/01.prs.0000282725.61640.e1

Kaye, J. T., Bradford, D. E., & Curtin, J. J. (2016). *Psychophysiology*. doi: 10.1111/psyp.12663.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*, 665–682. doi: 10.1037/a0020198

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30, 261–273. doi: 10.1111/j.1469-8986.1993.tb03352.x

Larson, M. J., Baldwin, S. A., Good, D. A., & Fair, J. E. (2010). Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): The role of number of trials. Psychophysiology, *47*(6), 1167-1171. doi:10.1111/j.1469-8986.2010.01022.x

Lindström, B. R., Mattsson-Mårn, I. B., Golkar, A., & Olsson, A. (2013). In your face: risk of punishment enhances cognitive control and error-related activity in the corrugator supercilii muscle. *PloS One*, *8*(6), e65692.

Maidhof, C., Rieger, M., Prinz, W., & Koelsch, S. (2009). Nobody is perfect: ERP effects prior to performance errors in musicians indicate fast monitoring processes. *PLoS One, 4*(4), e5032-e5032.

Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, *48*(6), 852-860. doi: 10.1111/j.1469-8986.2010.01152.x

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177-190. doi:10.1016/j.jneumeth.2007.03.024

Mathiassen, S. E., Winkel, J., & Hägg, G. M. (1995). Normalization of surface EMG amplitude from the upper trapezius muscle in ergonomic studies—a review. *Journal of Electromyography and Kinesiology, 5*(4), 197-226. doi: 10.1016/1050-6411(94)00014-x

Matzke, B., Herpertz, S. C., Berger, C., Fleischer, M., & Domes, G. (2013). Facial reactions during emotion recognition in borderline personality disorder: a facial electromyography study. *Psychopathology*, *47*(2), 101-110. doi: 10.1159/000351122

Meyer, D. E., Riesel, A., & Proudfit, G. H. (2013). Reliability of ERN across multiple tasks as a function of increasing errors. *Psychophysiology*, *50*, 1220-1225. doi: 10.1111/psyp.12132

Moran, T. P., Jendrusina, A. A., & Moser, J. S. (2013). The psychometric properties of the late positive potential during emotion processing and regulation. *Brain Research*, 1516, 66–75. doi:10.1016/j.brainres.2013.04.018

Morecraft, R.J., McNeal, D.W., Stillwell-Morecraft, K.S., Gedney, M., Ge, J., Schroeder, C.M., & Van Hoesen, G.W. (2007). Amygdala Interconnections with the Cingulate Motor Cortex

in the Rhesus Monkey. *The Journal of Comparative Neurology*, *500*(1)*:134-165.* doi:10.1002/cne.21165

Morecraft, R. J., Stilwell–Morecraft, K. S., & Rossing, W. R. (2004). The Motor Cortex and Facial Expression: New Insights From Neuroscience. *The Neurologist*, *10*(5), 235-249. doi:10.1097/01.nrl.0000138734.45742.8d

Morecraft, R. J., & Van Hoesen, G. W. (1993). Frontal granular cortex input to the cingulate (M3), supplementary (M2) and primary (M1) motor cortices in the rhesus monkey. Journal of Comparative Neurology, 337(4), 669–689. doi:10.1002/cne.903370411

Morecraft, R. J., & Van Hoesen, G. W. (1998). Convergence of Limbic Input to the Cingulate Motor Cortex in the Rhesus Monkey. Brain Research Bulletin, *45*(2), 209–232. doi:10.1016/s0361-9230(97)00344-4

Morecraft, R. J., & Van Hoesen, G. W. (2003). Functional neuroanatomy of limbic structures and some relationships with prefrontal cortex. In R.B. Schiffer, S.M. Rao, & B.S. Fogel (Eds.), *Neuropsychiatry* (pp. 294-327). Baltimore, USA: Lippincott, Williams and Wilkins.

de Morree, H.M., & Marcora, S.M. (2010). The face of effort; frowning muscle activity reflects effort during a physical task. *Biological Psychology*, *85*(3), 377-382. doi: 10.1016/j.biopsycho.2010.08.009

de Morree, H. M., & Marcora, S. M. (2012). Frowning muscle activity and perception of effort during constant-workload cycling. *European Journal of Applied Physiology*, 112, 1967–1972. doi: 10.1007/s00421-011-2138-2

Olvet, D. M., & Hajcak, G. (2009). The stability of error‐related brain activity with increasing trials. *Psychophysiology*, *46*(5), 957-961. doi: 10.1111/j.1469-8986.2009.00848.x

Pontifex, M. B., Scudder, M. R., Brown, M. L., O'Leary, K. C., Wu, C. T., Themanson, J. R., & Hillman, C. H. (2010). On the number of trials necessary for stabilization of error‐related brain activity across the life span. *Psychophysiology*, *47*(4), 767-773. doi: 10.1111/j.1469-8986.2010.00974.x

Pope, L. K., & Smith, C. A. (1994). On the distinct meanings of smiles and frowns. *Cognition & Emotion*, *8*(1), 65-72. doi: 10.1080/02699939408408929

Proulx, T., Inzlicht, M., & Harmon-Jones, E. (2012). Understanding all inconsistency compensation as a palliative response to violated expectations. *Trends in Cognitive Sciences*, *16*(5), 285–291. doi:10.1016/j.tics.2012.04.002

Saunders, B., He, F., & Inzlicht, M. (2015). Does gratitude improve self-regulation? No evidence that gratefulness enhances neural performance monitoring or conflict-driven control. *PLoS One*, *10*(12): e0143312. doi: 10.1371/journal.pone.0143312

Saunders, B., Milyavskaya, M., & Inzlicht, M., (2015). What does cognitive control feel like? Effective and ineffective cognitive control is associated with divergent phenomenology. *Psychophysiology, 52,* 1205-1217. doi: 10.1111/psyp.12454

Schacht, A., Dimigen, O., & Sommer, W. (2010). Emotions in cognitive conflicts are not aversive but are task specific. *Cognitive, Affective, & Behavioral Neuroscience*, *10*(3), 349-356. doi: 10.3758/CABN.10.3.349

Schacht, A., Nigbur, R., & Sommer, W. (2009). Emotions in go/nogo conflicts. *Psychological Research*, 73, 843–856. doi:10.1007/s00426-008-0192-0

Schüpbach, R. L., Gendolla, G. H. E., Schüpbach, R. L., Gendolla, G. H., & Silvestrini, N. (2014). Contrasting the effects of suboptimally versus optimally presented affect primes on effort-related cardiac response. *Motivation and Emotion*, *38*(6), 748-758. doi:

10.1007/s11031-014-9438-x

Simmonds, D. J., Pekar, J. J., & Mostofsky, S. H. (2008). Meta-analysis of Go/No-go tasks demonstrating that fMRI activation associated with response inhibition is task-dependent. *Neuropsychologia*, *46*(1), 224-232. doi: 10.1016/j.neuropsychologia.2007.07.015

Smith, D. P., Hillman, C. H., & Duley, A. R. (2005). Influences of age on emotional reactivity during picture processing. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *60*(1), P49-P56. doi: 10.1093/geronb/60.1.p49

Topolinski, S., Erle, T. M., & Reber, R. (2015). Necker's smile: Immediate affective consequences of early perceptual processes. *Cognition*, *140*, 1-13. doi:10.1016/j.cognition.2015.03.004

Topolinski, S., Likowski, K. U., Weyers, P., & Strack, F. (2009). The face of fluency: Semantic coherence automatically elicits a specific pattern of facial muscle reactions. *Cognition & Emotion*, *23*(2), 260–271. doi:10.1080/02699930801994112

Topolinski, S., & Strack, F. (2015). Corrugator activity confirms immediate negative affect in surprise. *Frontiers in Psychology*, 6. doi:10.3389/fpsyg.2015.00134

Weinberg, A., & Hajcak, G. (2011). Longer term test–retest reliability of error-related brain activity. *Psychophysiology*, *48*(10), 1420-1425. doi: 0.1111/j.1469-8986.2011.01206.x

Weinberg, A., Riesel, A., & Hajcak, G. (2012). Integrating multiple perspectives on error-related brain activity: the ERN as a neural indicator of trait defensive reactivity. *Motivation and Emotion*, *36*(1), 84-100. doi: 10.1007/s11031-011-9269-y

de Wied, M., van Boxtel, A., Zaalberg, R., Goudena, P. P., & Matthys, W. (2006). Facial EMG responses to dynamic emotional facial expressions in boys with disruptive behavior

disorders. *Journal of Psychiatric Research*, 40(2), 112-121. doi:
101016/j.jpsychires.2005.08.003

Figures

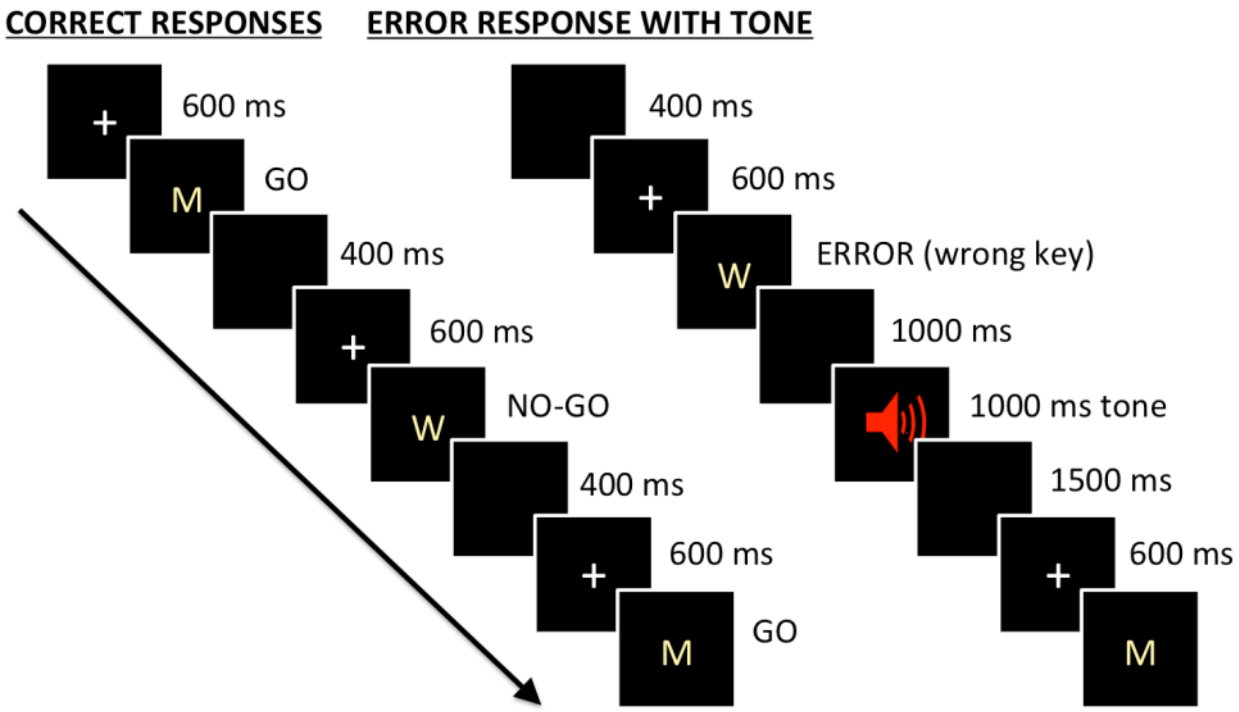

Figure 1. Depiction of modified Go-NoGo task. Participants were instructed to respond to the letters "M" and "W" with different keys on the keyboard. When participants pressed a wrong key, they had a 50% chance to be presented with a 1000 ms, 3500 Hz tone. Depending upon the block, this tone was either loud (95 dB) or quiet (20 dB). On trials when participants pressed a wrong key and were not presented with the tone, the task continued as if the participant had pressed the correct key.

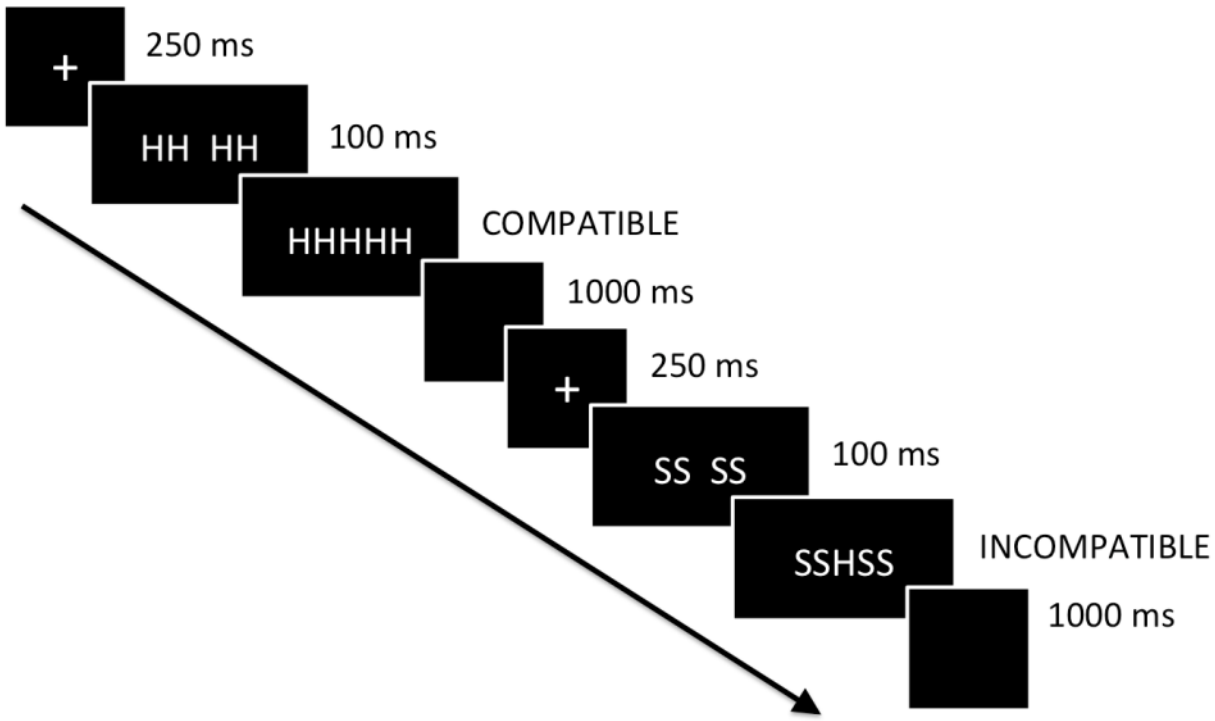

Figure 2. Depiction of flanker task. Participants were instructed to respond to letter in the center of the array. Following the presentation of the fixation cross, participants were primed with the flankers ("HH_HH" or "SS_SS", with no underscore) for 100 ms. This was followed by the presentation of the full array that included the center letter, which was either compatible ("HHHHH"/"SSSSS") or incompatible ("HHSHH"/"SSHSS") with the adjacent flanker letters.

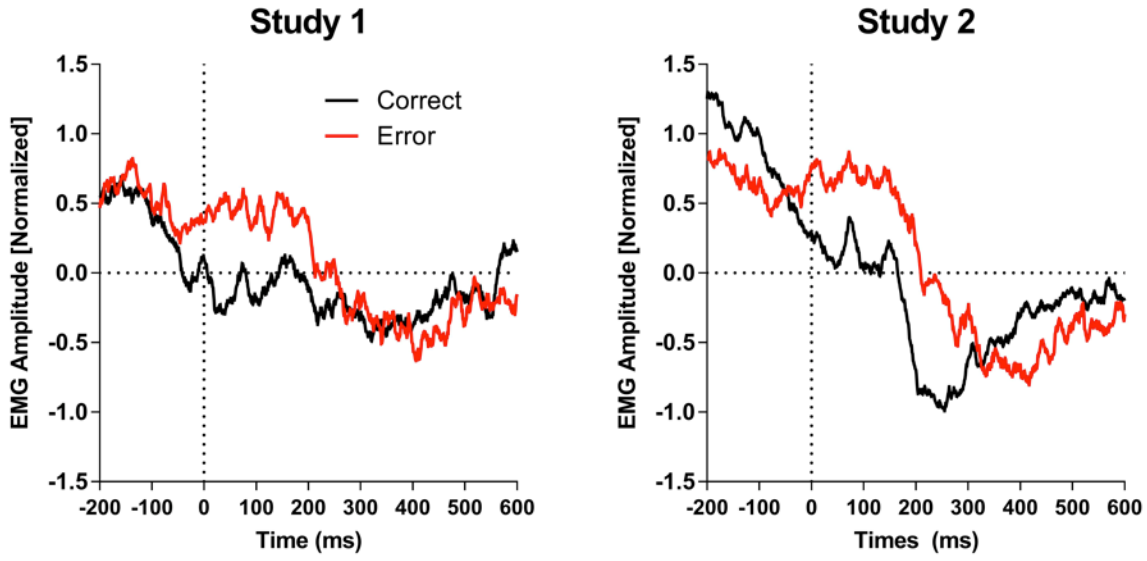

Figure 3. Grand average, standardized cEMG for correct and error responses for Study 1 and 2. In the present study, cEMG was defined as EMG activity between 0 and 100 ms post-response. Pre-response differences between conditions are not indicative of an unstable baseline; this is an effect of standardized the raw data separately by condition across the epoch.

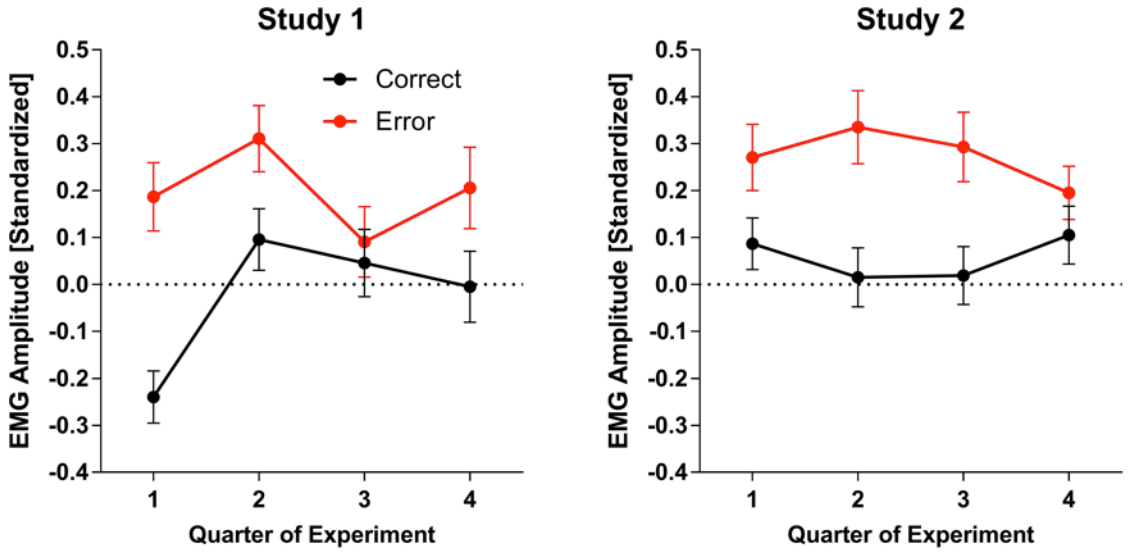

Figure 4. Grand average cEMG for correct and error responses for Study 1 and 2, across four sequential quarters of the experiment. Error bars depict within-participant standard errors.

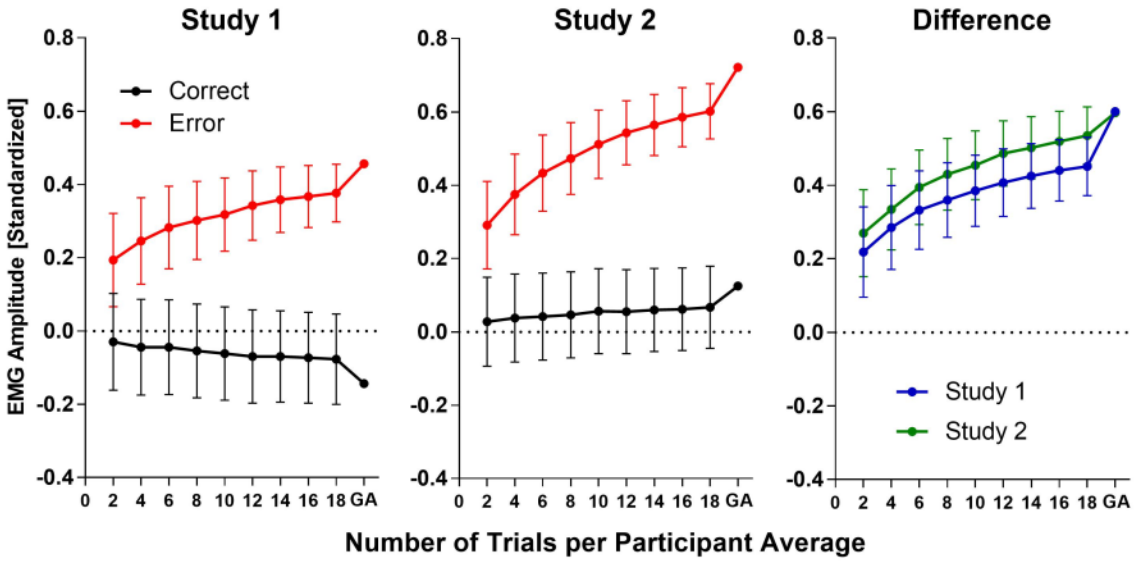

Figure 5. Standardized cEMG error, correct, and difference averages from Study 1 and 2. Each average represents the grand mean of resampled participant averages. Error bars depict 95% within-subject confidence intervals based on the average of within-participant standard deviations.

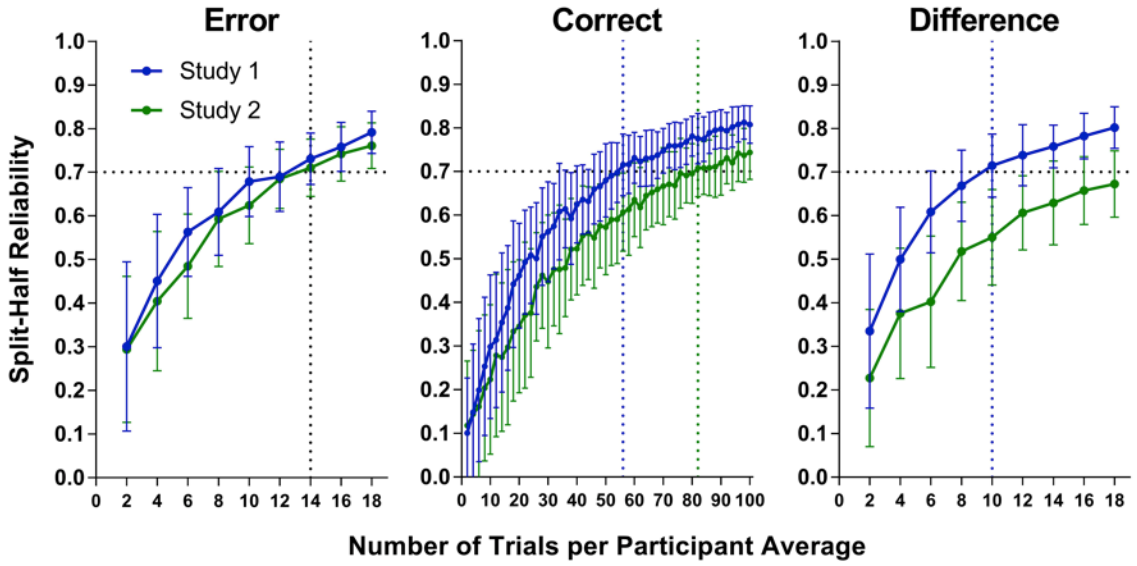

Figure 6. Spearman-Brown corrected split-half reliabilities of standardized cEMG trials from error, correct, and difference data, where each split-half reliability represents the grand mean of resampled participant *r* values. For error data, both Study 1 and 2 required at least 14 trials for high reliability ($r \geq .70$). For correct data, Study 1 required 56 trials (blue-dotted line) and Study 2 required 82 trials (green-dotted line) for high reliability. For difference data, high reliability was obtained in 10 trials for Study 1, but was not quite obtained for Study 2 in 18 trials ($r = .68$). Error bars depict between-participant standard errors.

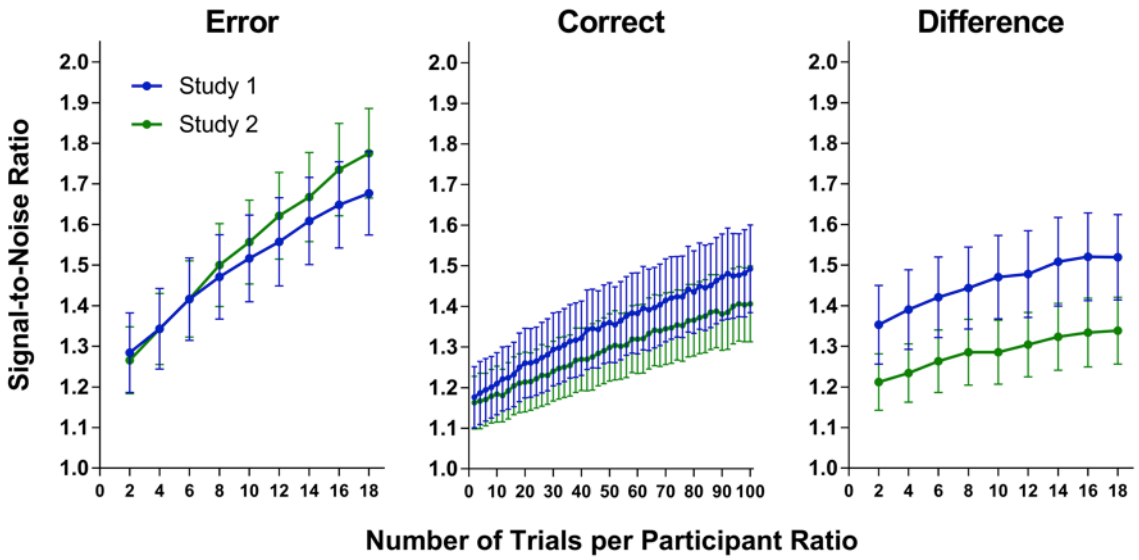

Figure 7. Signal-to-noise ratios of standardized cEMG averages for correct, error, and difference data, where each SNR represents the grand mean of resampled and averaged participant SNRs. For error and difference data, these sets of averages ranged from 2 to 18 trials; for correct trials, these sets of averages ranged from 2 to 100. Error bars depict the between-subject average of within-participant standard deviations.