American Psychological Association

Journal of Abnormal Psychology

0021-843X/20/$12.00

2020, Vol. 129, No. 1, 29–37
http://dx.doi.org/10.1037/abn0000458

Methodological Choices in Event-Related Potential (ERP) Research and Their Impact on Internal Consistency Reliability and Individual Differences: An Examination of the Error-Related Negativity (ERN) and Anxiety

Julia Klawohn and Alexandria Meyer
Florida State University

Anna Weinberg
McGill University

Greg Hajcak
Florida State University

Researchers in clinical psychophysiology make several methodological decisions during the analysis of event-related potentials (ERPs). In the current study, we review these choices from the perspective of individual differences. We focus on baseline period and reference scheme (i.e., average, mastoid, current source density), as well as choices regarding where (i.e., single electrode site vs. pooling of sites), when (i.e., area, area around peak), and how (i.e., subtraction- or regression-based difference scores) to quantify ERPs. To illustrate the impact of these analytic pathways on internal consistency reliability and individual differences, we focus on the error-related negativity (ERN) and anxiety—and present data from 2 samples: 1st, in adults with diagnosed generalized anxiety disorder (GAD); 2nd, in relation to continuous self-reported symptoms of GAD in a large community sample of female adolescents. Results generally indicated similar internal consistency and between-subjects effect sizes across all evaluated methods. Nonetheless, some patterns of variation emerged, such as that, across both data sets, difference-based ERN measures, especially with mastoid reference, yielded more robust associations with GAD diagnosis and symptoms, despite somewhat lower internal consistency. The current analyses suggest that the association between ERN and anxiety is robust across a range of commonly used methodological choices. The present study is an example of how systematic analyses of analytic strategies on measures of internal consistency and between-subjects variability could help inform individual-differences ERP research.

General Scientific Summary
This study systematically explores the effects of different analytic choices on the internal consistency of a electroencephalogram neural measure and its relation to symptoms of anxiety in two independent data sets. Results indicate psychometric properties of the error-related negativity are robust across a range of commonly used methodological choices.

Keywords: ERN, event-related potentials, ERP, psychopathology, methods

Supplemental materials: http://dx.doi.org/10.1037/abn0000458.supp

Event-related potentials (ERPs) are measures of neural function that can be used to study distinct neural processes both within and across individuals; they are relatively inexpensive and easy to assess, as well as feasible and safe to assess across the life span. Moreover, ERPs can be used as neurocognitive measures to characterize and differentiate clinical groups and related traits and symptom dimensions (Weinberg, Dieterich, & Riesel, 2015). Further, ERPs can function as neurobiological risk markers that can

Julia Klawohn, Department of Biomedical Sciences and Department of Psychology, Florida State University; Alexandria Meyer, Department of Psychology, Florida State University; Anna Weinberg, Department of Psychology, McGill University; Greg Hajcak, Department of Biomedical Sciences and Department of Psychology, Florida State University.

Part of the results were presented at the 2018 Annual Meeting of the Society for Psychophysiological Research. Supported by National Institute of Mental Health Grants R01 MH097767 to Greg Hajcak, F31 MH091837 to Anna Weinberg, and F31 MH102880 to Alexandria Meyer. This study was approved by the Institutional Review Board at Florida State University (2018.26010).

Correspondence concerning this article should be addressed to Julia Klawohn, Department of Biomedical Sciences and Department of Psychology, Florida State University, 1107 West Call Street, Tallahassee, FL 32306. E-mail: julia.klawohn@gmail.com

track patterns of familial risk for psychopathology—and ERPs can be leveraged to prospectively predict the onset and course of psychiatric disorders (Hajcak, Klawohn, & Meyer, 2019).

Although ERPs directly reflect functional electrocortical activity, they are the end state that results from a number of data processing steps and analytic decisions. Although there are commonly accepted processing principles and recommendations (e.g., Keil et al., 2014; Luck, 2014), there are various suitable options to choose from for many of the processing steps. For instance, ERPs reflect the voltage, or electrical difference, between an electrode site and a reference site—an ERP waveform "at" one electrode really reflects the electrical potential between that electrode and the reference used. One of the most common reference schemes is to contrast electroencephalogram (EEG) channels against the averaged activity recorded from the left and right mastoid electrodes. Similar approaches use an electrode placed on the nose or both earlobes as the reference. On the other hand, the average reference scheme uses the mean activity of all channels as the reference for each individual EEG channel (for more detailed descriptions see Luck, 2014). Further, a form of *reference-free* data analysis is the current source density (CSD) transformation, which estimates radial current flow from the scalp-recorded EEG by using neighboring electrodes as the reference (Kayser & Tenke, 2015). Although all these referencing schemes are reasonable for most ERPs, they do alter the appearance of waveforms considerably (Hajcak, Weinberg, MacNamara, & Foti, 2012), and there is only scarce knowledge on how the choice of reference might affect both psychometric properties of ERPs and relationships between ERPs and other individual differences measures.

ERPs are often quantified in terms of the difference between two within-subject conditions (e.g., ERP activity on emotional vs. neutral trials). This scoring approach is used to isolate neural activity associated with one condition relative to another—this is also done to examine the specificity of differences that may exist between people (i.e., to control for potential differences in a baseline condition). Relative measures are often derived by computing subtraction-based difference scores (i.e., subtraction of condition-related mean amplitudes, or by scoring the ERP difference waveform—what we have previously referred to as δ measures; Meyer, Lerner, De Los Reyes, Laird, & Hajcak, 2017). Moreover, we have recently suggested using residualized scores as an alternative relative ERP difference approach (Meyer et al., 2017).

In addition, there is substantial variability in *where* and *how* ERPs are quantified. ERPs are typically measured at the site of the maximum (i.e., where an ERP is largest); however, many studies average across neighboring electrode sites and measure the ERP at a pooling of electrodes—often with the presumption that doing so increases signal and reduces noise. In terms of ERP quantification, one common approach is calculating the mean activity in a specific time period (i.e., mean amplitude). Alternatively, ERPs can be scored using peak-based methods, based on the determination of a local maximum or minimum; this approach includes simple peak (i.e., the single most extreme amplitude value) or peak-to-peak (i.e., the amplitude difference between the peak of the ERP of interest and another peak) quantification. In general, the advantage of peak scoring approaches is that they can account for individual variation in the timing of ERP peaks (i.e., mean amplitude scoring focuses on the same window across all individuals); however, peak scoring approaches have been criticized because they weight a single data point and can be biased measures (Luck, 2014). Scoring the area around a peak (i.e., mean activity of a component centered around the peak) is a hybrid method that may benefit from the relative advantages of both the peak and mean amplitude approaches.

In the present study, we systematically investigated the impact of these methodological choices on the error-related negativity (ERN)—in terms of both its internal consistency and relationship with anxiety. The ERN is a response-locked ERP that presents as a sharp negative deflection shortly after error commission over fronto-central electrodes and is a well-established electrophysiological marker of error processing. Increased ERN amplitudes have been reported in adult clinical groups with obsessive–compulsive disorder (OCD; Riesel, 2019), generalized anxiety disorder (GAD; Weinberg, Olvet, & Hajcak, 2010), and social anxiety disorder (Endrass, Riesel, Kathmann, & Buhlmann, 2014), as well as in pediatric OCD and anxiety (Meyer, 2017). Moreover, amplified error signaling can indicate risk for psychopathology, as shown by familial (Riesel et al., 2019) and prospective (Meyer, Hajcak, Torpey-Newman, Kujawa, & Klein, 2015) developmental studies. Thus, the ERN has emerged as a neural measure with substantial clinical utility (Hajcak et al., 2019).

ERPs always include baseline activity prior to an event of interest. In ERP studies examining stimulus-related processes, the 200 or 500 ms period prior to stimulus onset is often used as baseline. The choice of baseline is more complicated for the ERN insofar as differences between error and correct trials are evident *prior* to an incorrect button press; thus, using the 200 ms window prior to responses may include some error-related brain activity. Indeed, studies have used either the mean activity from −200 to 0 ms before response (Hajcak, McDonald, & Simons, 2004) or −500 to −300 ms (Weinberg et al., 2016) as baseline.

In terms of individual difference studies on the ERN, mastoid reference (Gehring, Himle, & Nisenson, 2000; Olvet & Hajcak, 2010), average reference (Endrass, Klawohn, Schuster, & Kathmann, 2008), and the CSD references schemes have all been used (Nelson, Jackson, Amir, & Hajcak, 2017). The ERN is scored most commonly at or around the FCz location (Kaczkurkin, 2013) or at pooled electrode sites surrounding FCz (Larson, Steffen, & Primosch, 2013). Some studies have scored the ERN as the average activity (i.e., mean amplitude) in a fixed window (Gehring et al., 2000), whereas others have quantified the ERN in terms of simple peak scoring (Nieuwenhuis, Nielen, Mol, Hajcak, & Veltman, 2005), peak-to-peak scoring (Klawohn, Endrass, Preuss, Riesel, & Kathmann, 2016), or area around the peak (Boksem, Tops, Kostermans, & De Cremer, 2008). Several studies have analyzed relative difference scores (i.e., error relative to correct trials [ΔERN]), using the subtraction of mean amplitude (Meyer et al., 2015), regression-based residual scores (Meyer et al., 2017), or area around the peak of the difference waveform (Chong & Meyer, 2019).

For the current investigation, two different samples were analyzed to systematically investigate the impact of common methodological choices—such as different references, baseline periods, and quantification methods—on psychometric properties of the ERN and its relationship to generalized anxiety. Because the ERN relates to both pathological and continuous variability of worry (Moser, Moran, Schroder, Donnellan, & Yeung, 2013) and in line

with the most common study approaches in clinical ERP research, the current study first reexamined previously reported data on the ERN in relation to clinical GAD; then, we conducted identical analyses in a new large adolescent data set in relation to continuous self-reported symptoms of worry.

Method

Samples and Measures

Sample 1 combines EEG data of adult participants from two previously published studies (Weinberg, Klein, & Hajcak, 2012; Weinberg et al., 2010), reanalyzed here using different methodological approaches. The sample includes 40 participants with a current diagnosis of generalized anxiety disorder (GAD), without comorbid depression, and 51 participants without current or past diagnosis of a psychiatric disorder (i.e., healthy controls [HC]). All participants were interviewed with the Structured Clinical Interview for *DSM–IV* Axis I Disorders (First, Spitzer, Gibbon, & Williams, 1995), to ensure they either met diagnostic criteria for current GAD or did not currently or previously meet criteria for any Axis I diagnosis, respectively. For additional information see Weinberg et al. (2010). Participants had a mean age of 25.6 years (GAD: 26.4, $SD = 9.9$; HC: 25.01, $SD = 8.3$) and were predominantly female (GAD: 95.0%, HC 84.3%). All participants gave informed consent prior to participation.

Sample 2 is from a large prospective study in female adolescents. Basic descriptions of Wave 1 of assessment can be found in Meyer, Carlton, Crisler, and Kallen (2018). The data for the current study stem from Wave 2 of assessment and have not been reported before. Self-report and flanker task EEG data were available from 195 adolescent girls. Data from 11 participants had to be excluded from analysis due to insufficient data quality ($n = 7$), less than six error trials ($n = 1$), or more than 45% errors ($n = 3$). The resulting sample included 184 girls with a mean age of 14.4 years (range = 10–17). All participants and their parents provided informed consent and assent, respectively. All adolescents completed a self-report with the Screen for Child Anxiety Related Emotional Disorders (SCARED; Birmaher et al., 1997), which measures anxiety symptoms in those 9−18 years of age with 38 items, each rated on a scale from 0 to 2. For the current study, we studied the association of the ERN with the GAD subscale.

EEG Data Collection and Experimental Task

Participants in both samples completed an identical arrowhead version of a flanker task, using Presentation software (Neurobehavioral Systems Inc., Berkeley, USA). In total, 330 trials were completed. Half of the stimuli were compatible and the other half incompatible, presented in random order in 11 blocks. Performance-sensitive feedback was provided during the breaks between blocks: *Please try to be more accurate* was shown when performance was 75% correct or lower and *Please try to respond faster* when performance was above 90% correct; otherwise, *You are doing a great job* was presented. Continuous EEG data were collected from 34 electrodes, positioned with an elastic cap in accordance with the 10/20 system, as well as on the right and left mastoids. Eye movements and blinks were recorded with four electrodes, two placed horizontally at the canthi of the eyes, two placed above and below the right eye. Data were preamplified with a BioSemi ActiveTwo System (Biosemi, Amsterdam, Netherlands) and digitized at 1024 Hz sampling rate. A common mode sense active electrode served as a recording reference.

EEG Analyses (i.e., Methodological Pathways)

EEG data were processed offline using Brain Vision Analyzer, Version 2.1 (Brain Products, Gilching, Germany). Initially, EEG data were rereferenced to the average of both mastoid electrodes and bandpass filtered from .1 to 30 Hz. Data were segmented into epochs from −500 to 1,000 ms around responses. Ocular artifacts were corrected using the algorithm by Gratton, Coles, and Donchin (1983), employing the horizontal and vertical eye channels. Further artifact rejection was performed automatically by rejecting epochs of data when voltage steps of more than 50 μV between sampling points, or a maximal absolute difference of more than 300 μV was present, or when low activity was detected, defined as a voltage difference less than .5 μV over 100 ms. Remaining artifacts were identified and removed based on visual inspection.

Subsequent analyses were then performed three times, each using a different reference scheme (see Figure S1 in the online supplemental material). Data either remained referenced to mastoid electrodes or were rereferenced to the average of all electrode sites, or else the current source density (CSD) transformation was applied (order of splines = 4, maximal degree of Legendre polynomials: 10; λ smoothing parameter = 10^{-5}). For all reference schemes, response-locked ERPs were averaged for correct and incorrect responses separately. Two baseline corrections were applied, using intervals from −500 to −300 ms or −200 to 0 ms before response. For all resulting ERP averages, several different quantification methods were used in relation to electrode FCz, and to a pool of electrodes (i.e., average of Cz, FC1, FC2, FCz, and Fz). Peak detection determined individual ERN peaks as the most negative deflection from −100 to 200 ms around response onset. Further, a preceding positive peak was identified within −150 to 50 ms relative to response onset. Peaks were visually inspected and corrected if necessary. ERP quantification included the following measures: mean amplitude between 0 and 100 ms after response, mean amplitude over 100 ms centered around the peak of the ERN (i.e., area around the peak of the ERN), and peak-to-peak amplitude (i.e., difference between the ERN peak and preceding positive peak). Further, several difference-score (i.e., ΔERN) measures were calculated: $\Delta ERN_{subtract}$ as the difference between error and correct mean amplitude scores in the 0- to 100-ms window relative to response onset; the mean amplitude over 100 ms centered around the peak of the difference waveform was quantified; ΔERN_{resid} scores were generated as the variance leftover in a regression (i.e., unstandardized residuals) wherein correct response mean amplitudes were entered predicting error mean amplitudes.

Statistical Analysis

In Sample 1, group differences in ERN amplitudes for the GAD and HC group were examined for all methodological choices using independent-samples *t* tests. Effect size estimates (Cohen's *d*) and respective 95% confidence intervals (CIs) were determined (Wuensch, 2012). In Sample 2, correlation between ERN scores

and the SCARED GAD subscale was examined with Pearson's *r*, and confidence intervals were determined via Fisher's *z* transformation. Further, internal consistency reliability of ERN scores was examined with a split-half approach, where the correlation between averages of odd- and even-numbered trials was determined and corrected using the Spearman-Brown prophecy formula (Nunnally, Bernstein, & Berge, 1967). All statistical analyses were conducted with SPSS Statistics, Version 23.0.

Results

Sample 1

Grand-average waveforms are presented in Figure 1 (Panel A), and results of internal consistency and between-subjects analyses are presented in Table 1 and Figure 2. For simple ERN measures, the choice of baseline did not impact internal consistency (mean $r = .80$, for both baseline periods; range = .72 to .88). However, the response-proximal baseline (i.e., −200 to 0 ms) appeared to be associated with larger between-subjects effect sizes than did the earlier baseline (i.e., −500 to −300 ms), and this pattern was especially prominent for ERN measures with average reference. Across all reference schemes, the baseline-independent peak-to-peak measures for both single and pooled electrodes had high internal consistency (mean $r = .85$) and rather large between-groups effect sizes (mean $d = .50$).

Regarding ΔERN measures, the response-proximal baseline showed higher internal consistency (mean $r = .79$) than did the earlier baseline (mean $r = .65$). Within the response-proximal baseline period, both the $\Delta ERN_{subtract}$ and the ΔERN_{resid} had high effect sizes across reference schemes (mean $d = .48$), whereas for the earlier baseline, only mastoid-referenced ΔERN measures resulted in significant group effects. For all ΔERN measures, across reference schemes and baselines, effect size estimates were somewhat higher for pooled (mean $d = .45$) than single (mean $d = .40$) electrode quantification, whereas no differences in internal consistency emerged. Despite these variations, all confidence intervals overlapped (see Figure 2), indicating that none of the analytic choices in Sample 1 produced significantly different results from the rest.

Sample 2

The SCARED GAD subscale yielded a sample mean of 5.9 ($SD = 4.6$). Results for the association of ERN amplitudes with the GAD subscale are presented in Table 1, and grand-average waveforms appear in Figure 1 (Panel B). Inspection of results for simple ERN measures indicated acceptable to good internal consistency ($r = .68$ to .88, mean $r = .83$) for both baseline periods. Significant correlations with GAD symptoms were found using only the response-proximal (i.e., −200 to 0 ms) baseline for mean-amplitude measures with mastoid or average reference. In contrast, ΔERN measures were characterized by significant or trend-level correlations (all $ps < .077$, r range = −.131 to −.225) with self-reported GAD symptoms for all quantification methods and across reference schemes. Further considering the various ΔERN measures, the earlier baseline (i.e., −500 to −300 ms) tended to be associated with larger effect sizes for mastoid and CSD references but not when using average reference; pooling was generally associated with lower effect sizes for both mastoid and average reference but higher effect sizes for the CSD reference. Internal consistency was good for the response-proximal baseline (r range = .72 to .92) and somewhat more variable but in the acceptable to good range for the earlier baseline period (r range = .60 to .94). Similar to findings in Sample 1, confidence intervals for associations between ERN and GAD symptoms all overlapped (see Figure 2).

Discussion

The current study examined the impact of common choices in ERP analyses on the internal consistency of the ERN and its relationship with individual differences in anxiety. In particular, we focused on different reference schemes, baseline periods, and quantification methods. Our approach was to first reanalyze previously published data on ERN in a relatively large sample of adults with clinical GAD versus healthy controls—to evaluate the impact of specific analytic choices in these data. The more response-proximal baseline (i.e., −200 to 0 ms) was generally associated with better internal consistency, although all measures—including difference measures—had acceptable to good internal consistency. Almost all measures using a mastoid reference yielded robust between-groups differences; the fact that both simple ERN and difference measures of the ERN produced comparable between-groups differences when the mastoid reference was utilized is consistent with meta-analytic findings in adults (Moser et al., 2013). Most simple ERN measures using the CSD reference also robustly differentiated GAD from healthy participants, and average referenced data were also associated with significant group differences, though only when using the more response-proximal (i.e., −200 to 0 ms) baseline. There was no clear impact of scoring method (i.e., mean amplitude vs. peak-based scoring, or using single vs. pooled electrodes) in these data.

We then examined the impact of these analytic decisions in a new data set, in relation to continuous GAD symptoms in a large nonclinical sample of adolescents. In this data set, all ERN measures had acceptable to excellent internal consistency that was on par with internal reliabilities found in adults in Study 1; pooling and scoring approach again had a negligible impact on internal consistency. Unlike in Study 1, difference ERN measures in Study 2 were more reliably related to GAD symptoms in the adolescent sample than were simple ERN measures. For simple ERN measures, only mean amplitudes with the response-proximal (i.e., −200 to 0 ms) baseline using mastoid or average reference were significantly correlated with GAD symptoms. In contrast, 34 out of 38 ΔERN measures were significantly correlated with GAD symptoms, thus indicating an overall more robust association with anxiety. It might be an interesting avenue for a meta-analysis to examine whether, in pediatric and adolescent populations, a stronger association of anxiety with ΔERN measures in contrast to simple ERN measures generalizes across studies and age ranges.

It is important to note that the two samples analyzed differed in several ways: Study 1 was composed of 91 adults who either had diagnosed GAD or no diagnosed psychopathology, whereas Study 2 was a community sample of female adolescents. Thus, it is somewhat difficult to interpret differences between studies, which could be due to multiple factors (e.g., age, diagnostic status, statistical power). Nonetheless, results of the present studies col-

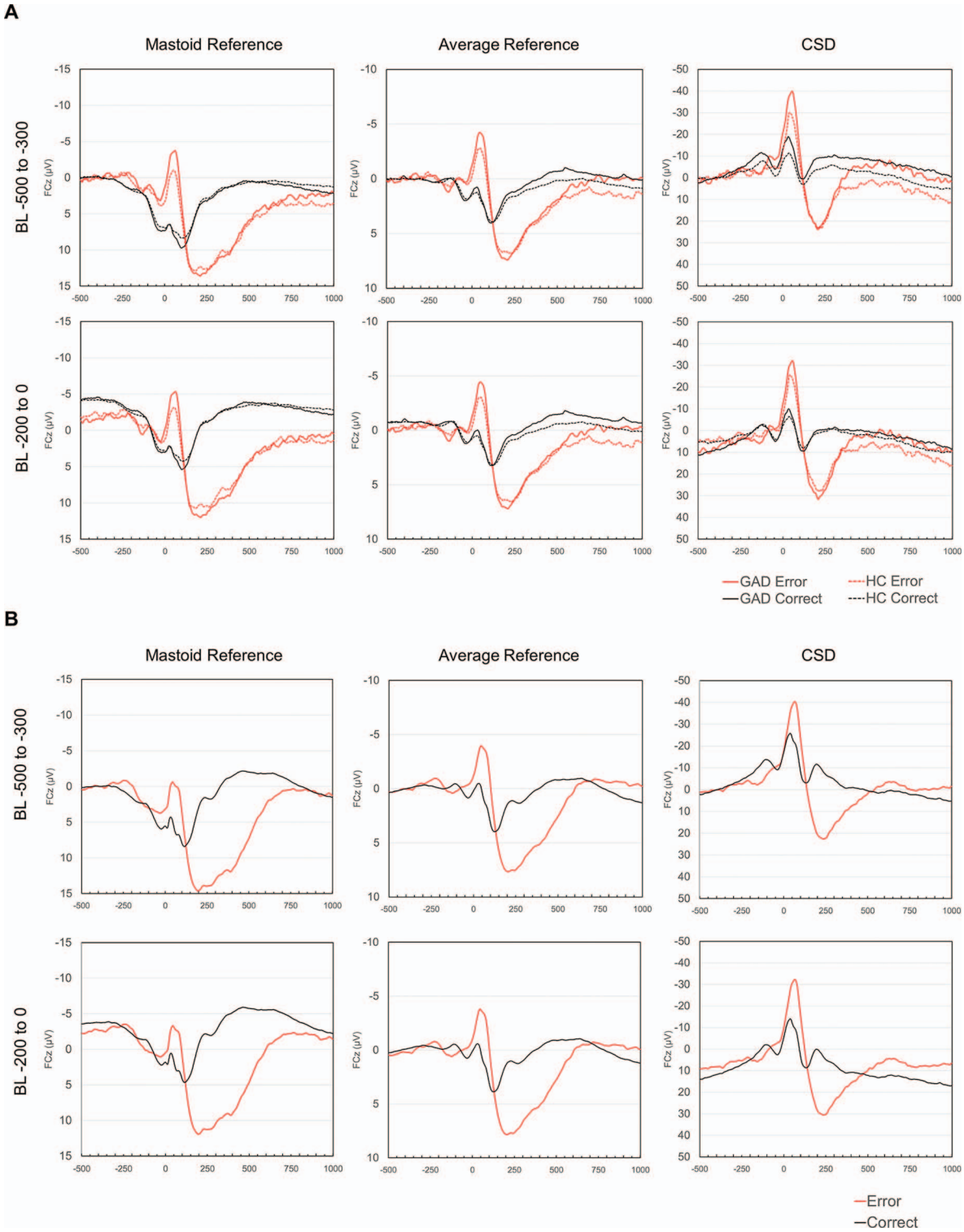

Figure 1. Grand-average waveforms for Sample 1 (Panel A) and Sample 2 (Panel B) with respect to the main analysis methods. CSD = current source density (transformation); BL = baseline; GAD = generalized anxiety disorder; HC = healthy controls. Please note different scales across columns. See the online article for the color version of this figure.

Table 1
Group Comparisons (Sample 1), Correlational Analysis (Sample 2), and Internal Consistency Analyses for ERPs Derived From Different Methodological Approaches

EEG analytic pathway (REF and quantification)	Sample 1						Sample 2					
	Baseline −200 to 0			Baseline −500 to −300			Baseline −200 to 0			Baseline −500 to −300		
	Group comparison			Group comparison			Correlation with GAD SCARED			Correlation with GAD SCARED		
	d	*p*	r_{int}	*d*	*p*	r_{int}	*r*	*p*	r_{int}	*r*	*p*	r_{int}
ERN measures												
MAST												
MA: FCz	**.515**	**.017**	.72	**.473**	**.028**	.71	**−.146**	**.048**	.76	−.114	.124	.83
MA: Pool	**.505**	**.019**	.72	**.439**	**.040**	.73	−.144	.051	.76	−.103	.165	.83
AAP: FCz	**.519**	**.016**	.74	**.444**	**.038**	.70	−.099	.180	.74	−.079	.288	.82
AAP: Pool	**.430**	**.045**	.78	.355	.096	.76	−.096	.196	.68	−.064	.387	.79
Peak to peak: FCz	**.567**	**.009**	.84	**.567**	**.009**	.84	−.112	.132	.77	−.112	.132	.77
Peak to peak: Pool	**.485**	**.024**	.86	**.485**	**.024**	.86	−.104	.160	.72	−.104	.160	.72
AVG												
MA: FCz	**.507**	**.019**	.76	.349	.102	.79	**−.148**	**.044**	.82	−.070	.346	.86
MA: Pool	**.549**	**.011**	.79	.349	.102	.84	**−.159**	**.031**	.81	−.059	.428	.84
AAP: FCz	**.432**	**.044**	.71	.270	.203	.77	−.129	.082	.84	−.046	.536	.89
AAP: Pool	**.484**	**.024**	.74	.275	.198	.84	−.129	.082	.83	−.021	.774	.86
Peak to peak: FCz	**.437**	**.041**	.82	**.437**	**.041**	.82	−.114	.123	.85	−.114	.123	.85
Peak to peak: Pool	**.463**	**.031**	.84	**.463**	**.031**	.84	−.100	.175	.82	−.100	.175	.82
CSD												
MA: FCz	**.500**	**.020**	.85	**.422**	**.048**	.73	−.087	.242	.86	−.047	.525	.90
MA: Pool	**.568**	**.009**	.85	**.463**	**.031**	.85	−.129	.081	.85	−.080	.278	.89
AAP: FCz	**.450**	**.036**	.85	.391	.068	.71	−.097	.190	.84	−.049	.507	.90
AAP: Pool	**.523**	**.015**	.84	**.422**	**.049**	.84	−.126	.088	.86	−.072	.330	.91
Peak to peak: FCz	**.514**	**.017**	.88	**.515**	**.017**	.88	−.080	.278	.88	−.080	.278	.88
Peak to peak: Pool	**.515**	**.017**	.87	**.515**	**.017**	.87	−.103	.163	.87	−.103	.163	.87
ΔERN measures												
MAST												
$\Delta ERN_{subtract}$: FCz	**.509**	**.018**	.79	**.617**	**.004**	.64	**−.207**	**.005**	.77	**−.225**	**.002**	.71
$\Delta ERN_{subtract}$: Pool	**.541**	**.012**	.79	**.632**	**.004**	.62	**−.195**	**.008**	.76	**−.205**	**.005**	.69
AAP diffwave: FCz	.410	.055	.79	**.511**	**.017**	.63	**−.205**	**.005**	.74	**−.215**	**.003**	.64
AAP diffwave: Pool	**.456**	**.034**	.77	**.545**	**.012**	.60	**−.182**	**.013**	.72	**−.183**	**.013**	.60
ΔERN_{resid}: FCz	**.547**	**.011**	.71	**.642**	**.003**	.60	**−.177**	**.016**	.75	**−.213**	**.004**	.71
ΔERN_{resid}: Pool	**.558**	**.010**	.71	**.642**	**.003**	.59	**−.170**	**.020**	.74	**−.195**	**.008**	.70
AVG												
$\Delta ERN_{subtract}$: FCz	.361	.091	.79	.304	.154	.66	**−.185**	**.012**	.80	**−.174**	**.018**	.65
$\Delta ERN_{subtract}$: Pool	**.439**	**.040**	.79	.380	.075	.67	**−.179**	**.015**	.77	**−.151**	**.041**	.66
AAP diffwave: FCz	.285	.179	.74	.224	.293	.59	**−.176**	**.017**	.82	**−.166**	**.024**	.71
AAP diffwave: Pool	.353	.098	.80	.292	.173	.69	**−.159**	**.031**	.82	−.133	.072	.68
ΔERN_{resid}: FCz	**.473**	**.027**	.75	.346	.105	.65	**−.178**	**.016**	.79	**−.166**	**.024**	.72
ΔERN_{resid}: Pool	**.539**	**.012**	.77	.418	.051	.68	**−.182**	**.013**	.77	−.144	.051	.67
CSD												
$\Delta ERN_{subtract}$: FCz	.376	.079	.86	.317	.138	.63	**−.157**	**.033**	.92	**−.191**	**.009**	.94
$\Delta ERN_{subtract}$: Pool	**.427**	**.047**	.86	.342	.109	.80	**−.193**	**.009**	.81	**−.217**	**.003**	.66
AAP diffwave: FCz	.275	.199	.82	.285	.180	.57	−.131	.077	.78	**−.168**	**.023**	.71
AAP diffwave: Pool	.334	.118	.83	.361	.091	.62	**−.147**	**.046**	.83	**−.173**	**.019**	.79
ΔERN_{resid}: FCz	**.448**	**.037**	.84	.203	.341	.64	−.137	.065	.83	**−.191**	**.009**	.77
ΔERN_{resid}: Pool	**.520**	**.016**	.84	.285	.182	.80	**−.179**	**.015**	.81	**−.212**	**.004**	.76

Note. Sample 1: $N = 91$ (generalized anxiety disorder: $n = 41$, healthy controls: $n = 50$); Sample 2: $N = 184$ adolescents. Boldface indicates statistically significant group comparisons or correlations. ERP = event-related potential; EEG = electroencephalogram; REF = reference; r_{int} = internal consistency; GAD = generalized anxiety disorder; SCARED = Screen for Child Anxiety Related Emotional Disorders; ERN = error-related negativity; MAST = linked mastoid reference; $\Delta ERN_{subtract}$ = error-correct mean amplitude scores (0–100 ms); Pool = fronto-central electrode pool (electrodes Fz, FCz, Cz, FC1, FC2); AAP = area around peak (100 ms); ΔERN_{resid} = regression-based residual ERN; AVG = average (reference); CSD = current source density; MA = mean amplitude (0–100 ms); Diffwave = difference waveform (error minus correct).

lectively suggest that the most common ERP analytic choices have only limited impact on both the psychometric properties of the ERN and its association with anxiety. Specifically, effect sizes were generally similar and overlapping. Along the same lines, internal consistency reliability across all analytic strategies was acceptable to good. Several additional consistent findings emerged across samples. Although studies often use a pooling of electrode sites based on the notion that this will increase internal consistency of the ERP score, we found no evidence that pooled electrode sites had superior internal consistency compared to single-site measures

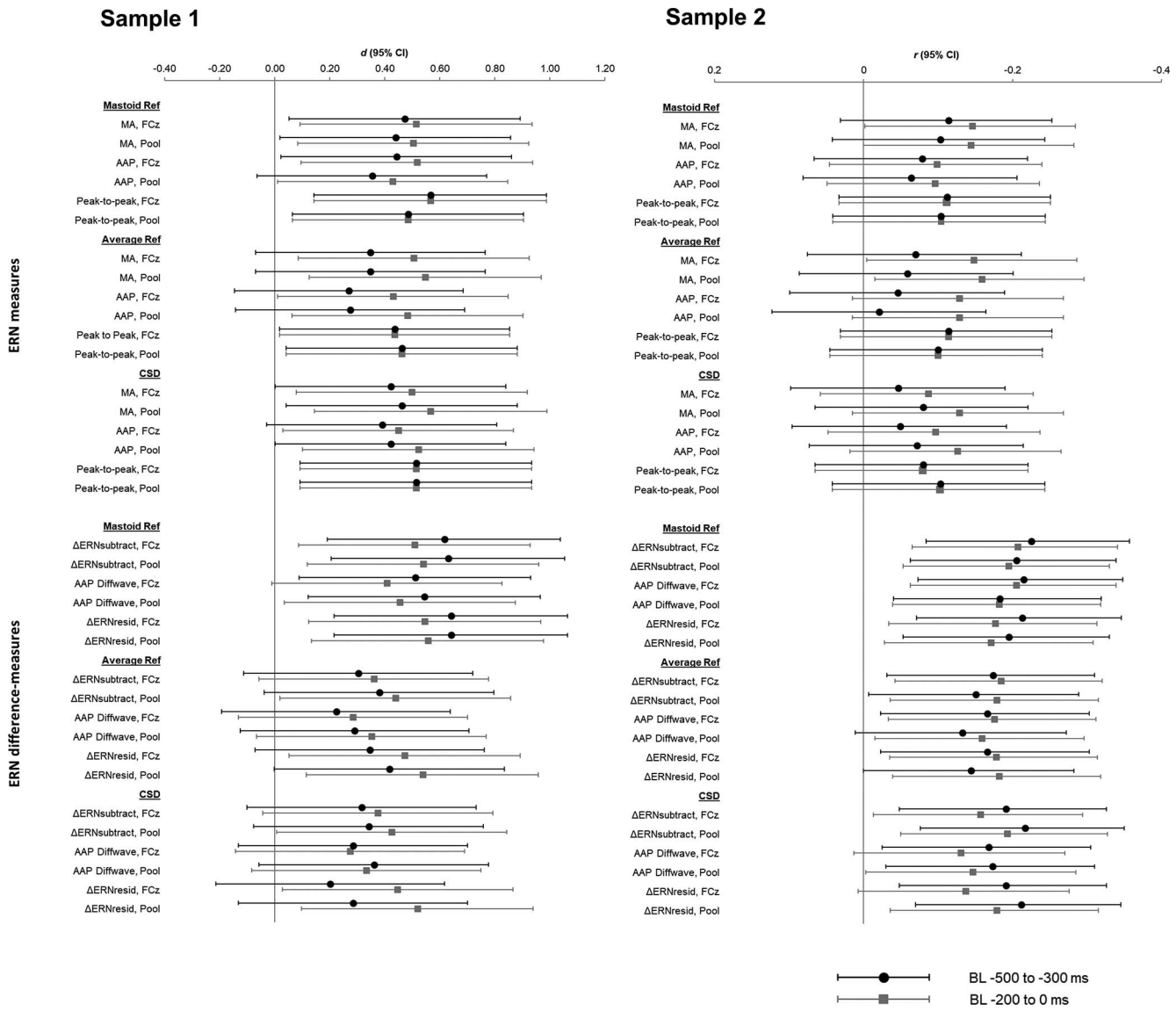

Figure 2. Forest plot of effect size estimates and 95% confidence intervals in Sample 1 and Sample 2. Sample 1: $N = 91$ (generalized anxiety disorder: $n = 41$, healthy controls: $n = 50$). Sample 2: $N = 184$ adolescents. BL = baseline; Ref = reference; ERN = error-related negativity; MAST = linked mastoid reference; Pool = fronto-central electrode pool (electrodes Fz, FCz, Cz, FC1, FC2); $\Delta ERN_{subtract}$ = error-correct mean amplitude scores (0–100 ms); AAP = area around peak (100 ms); ΔERN_{resid} = regression-based residual ERN; AVG = average (reference); CSD = current source density; MA = mean amplitude (0–100 ms); Diffwave = difference waveform (error minus correct).

of the ERN. Further, across both samples and study designs, almost all effects based on ΔERN measures and mastoid reference were significant—across all baseline periods and scoring approaches. This might be because ΔERN measures control for basic response processing as well as possibly overlapping stimulus-locked activity. In particular, the regression-based ERN measure might be best suited to control for suppression effects induced by stimulus-related activity (Meyer et al., 2017). Also, however, the mean amplitude $\Delta ERN_{subtract}$ and, in adults, the peak-to-peak ERN (in itself a form of difference score and baseline-independent) had overall good internal consistency and were robust regarding between-subjects effect sizes. Although the difference-based measures had somewhat lower internal consistency (though still within an acceptable range), it appears that a higher proportion of that reliable variance relates to other individual difference measures. Thus, despite concerns raised regarding the reliability of difference scores, ΔERN measures seem to better isolate the error-specific neural activity relevant to individual differences in anxiety.

There was also evidence that the baseline period impacts between-subjects differences, especially in considering simple versus difference ERN measures. For example, in both Studies 1 and 2, the response-proximal baseline was associated with

larger effect sizes for simple ERN measures when using mastoid reference, whereas the opposite was true for ΔERN measures using mastoid reference. These data suggest that a response-proximal baseline might be preferable for simple ERN measures when using a mastoid reference, whereas an earlier reference is preferable for difference measures. The choice of baseline might be specifically relevant to the ERN, because error and correct trials can differ well before response onset (i.e., see Figure 1). This is further complicated by the possibility that individual differences in stimulus-locked ERPs might overlap with and impact the ERN (Meyer et al., 2017; Riesel, Klawohn, Kathmann, & Endrass, 2017). The take-home suggestions for future research on ERN and anxiety would be to use mastoid reference, a difference-based ERN score, and an earlier baseline period (i.e., −500 to −300 ms).

We undertook these analyses to understand the concrete impact of common ERP analytic choices on both the psychometric properties of the ERN and its relationship with anxiety. In light of various reasonable choices available to ERP researchers, there is concern about the number of possible comparisons, p-hacking, and related concerns about replicability and false positive results (Baldwin, 2017; Luck & Gaspelin, 2017). These concerns could lead researchers to avoid analytic explorations of their data—an approach we took explicitly in the current study. Two things seem true: Effect sizes varied numerically, and some analytic paths produced results that fell below the threshold for statistical significance; however, internal reliability was uniformly high and between-subjects effect sizes were overlapping and generally consistent. No single analytic path produced statistically significant results—and there was not an outlying result based on any specific set of reasonable decisions. It is essential to ensure that methodological choices (e.g., quantification approaches, time frames) are not made based on data, because this inflates α error (Luck & Gaspelin, 2017); we would encourage researchers to specify specific analytic approaches a priori.

The current studies were limited in their focus on the ERN and anxiety. Therefore, similar analyses would be important in the context of other ERP measures and individual differences. As in the current study, we would only suggest conducting such analyses to understand the impact of methodological choices to guide future studies. Indeed, it might also be important for future studies to examine the impact of other methodological decisions, such as ocular correction methods.

References

Baldwin, S. A. (2017). Improving the rigor of psychophysiology research. *International Journal of Psychophysiology, 111,* 5–16. http://dx.doi.org/10.1016/j.ijpsycho.2016.04.006

Birmaher, B., Khetarpal, S., Brent, D., Cully, M., Balach, L., Kaufman, J., & Neer, S. M. (1997). The Screen for Child Anxiety Related Emotional Disorders (SCARED): Scale construction and psychometric characteristics. *Journal of the American Academy of Child & Adolescent Psychiatry, 36,* 545–553. http://dx.doi.org/10.1097/00004583-199704000-00018

Boksem, M. A., Tops, M., Kostermans, E., & De Cremer, D. (2008). Sensitivity to punishment and reward omission: Evidence from error-related ERP components. *Biological Psychology, 79,* 185–192. http://dx.doi.org/10.1016/j.biopsycho.2008.04.010

Chong, L. J., & Meyer, A. (2019). Understanding the link between anxiety and a neural marker of anxiety (the error-related negativity) in 5 to 7 year-old children. *Developmental Neuropsychology, 44,* 71–87. http://dx.doi.org/10.1080/87565641.2018.1528264

Endrass, T., Klawohn, J., Schuster, F., & Kathmann, N. (2008). Overactive performance monitoring in obsessive-compulsive disorder: ERP evidence from correct and erroneous reactions. *Neuropsychologia, 46,* 1877–1887. http://dx.doi.org/10.1016/j.neuropsychologia.2007.12.001

Endrass, T., Riesel, A., Kathmann, N., & Buhlmann, U. (2014). Performance monitoring in obsessive-compulsive disorder and social anxiety disorder. *Journal of Abnormal Psychology, 123,* 705–714. http://dx.doi.org/10.1037/abn0000012

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. (1995). *Structured Clinical Interview for DSM–IV Axis I Disorders*. New York, NY: New York State Psychiatric Institute.

Gehring, W. J., Himle, J., & Nisenson, L. G. (2000). Action-monitoring dysfunction in obsessive-compulsive disorder. *Psychological Science, 11,* 1–6. http://dx.doi.org/10.1111/1467-9280.00206

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55,* 468–484. http://dx.doi.org/10.1016/0013-4694(83)90135-9

Hajcak, G., Klawohn, J., & Meyer, A. (2019). The utility of event-related potentials in clinical psychology. *Annual Review of Clinical Psychology, 15,* 71–95. http://dx.doi.org/10.1146/annurev-clinpsy-050718-095457

Hajcak, G., McDonald, N., & Simons, R. F. (2004). Error-related psychophysiology and negative affect. *Brain and Cognition, 56,* 189–197. http://dx.doi.org/10.1016/j.bandc.2003.11.001

Hajcak, G., Weinberg, A., MacNamara, A., & Foti, D. (2012). ERPs and the study of emotion. In E. S. Kappenman and S. J. Luck (Eds.), *The Oxford handbook of event-related potential components* (pp. 441–474). http://dx.doi.org/10.1093/oxfordhb/9780195374148.013.0222

Kaczkurkin, A. N. (2013). The effect of manipulating task difficulty on error-related negativity in individuals with obsessive-compulsive symptoms. *Biological Psychology, 93,* 122–131. http://dx.doi.org/10.1016/j.biopsycho.2013.01.001

Kayser, J., & Tenke, C. E. (2015). On the benefits of using surface Laplacian (current source density) methodology in electrophysiology. *International Journal of Psychophysiology, 97,* 171–173. http://dx.doi.org/10.1016/j.ijpsycho.2015.06.001

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., . . . Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology, 51,* 1–21. http://dx.doi.org/10.1111/psyp.12147

Klawohn, J., Endrass, T., Preuss, J., Riesel, A., & Kathmann, N. (2016). Modulation of hyperactive error signals in obsessive-compulsive disorder by dual-task demands. *Journal of Abnormal Psychology, 125,* 292–298. http://dx.doi.org/10.1037/abn0000134

Larson, M. J., Steffen, P. R., & Primosch, M. (2013). The impact of a brief mindfulness meditation intervention on cognitive control and error-related performance monitoring. *Frontiers in Human Neuroscience, 7,* 308. http://dx.doi.org/10.3389/fnhum.2013.00308

Luck, S. J. (2014). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology, 54,* 146–157. http://dx.doi.org/10.1111/psyp.12639

Meyer, A. (2017). A biomarker of anxiety in children and adolescents: A review focusing on the error-related negativity (ERN) and anxiety across development. *Developmental Cognitive Neuroscience, 27,* 58–68. http://dx.doi.org/10.1016/j.dcn.2017.08.001

Meyer, A., Carlton, C., Crisler, S., & Kallen, A. (2018). The development of the error-related negativity in large sample of adolescent females:

Associations with anxiety symptoms. *Biological Psychology, 138,* 96–103. http://dx.doi.org/10.1016/j.biopsycho.2018.09.003

Meyer, A., Hajcak, G., Torpey-Newman, D. C., Kujawa, A., & Klein, D. N. (2015). Enhanced error-related brain activity in children predicts the onset of anxiety disorders between the ages of 6 and 9. *Journal of Abnormal Psychology, 124,* 266–274. http://dx.doi.org/10.1037/abn0000044

Meyer, A., Lerner, M. D., De Los Reyes, A., Laird, R. D., & Hajcak, G. (2017). Considering ERP difference scores as individual difference measures: Issues with subtraction and alternative approaches. *Psychophysiology, 54,* 114–122. http://dx.doi.org/10.1111/psyp.12664

Moser, J. S., Moran, T. P., Schroder, H. S., Donnellan, M. B., & Yeung, N. (2013). On the relationship between anxiety and error monitoring: A meta-analysis and conceptual framework. *Frontiers in Human Neuroscience, 7,* 466. http://dx.doi.org/10.3389/fnhum.2013.00466

Nelson, B. D., Jackson, F., Amir, N., & Hajcak, G. (2017). Attention bias modification reduces neural correlates of response monitoring. *Biological Psychology, 129,* 103–110. http://dx.doi.org/10.1016/j.biopsycho.2017.08.059

Nieuwenhuis, S., Nielen, M. M., Mol, N., Hajcak, G., & Veltman, D. J. (2005). Performance monitoring in obsessive-compulsive disorder. *Psychiatry Research, 134,* 111–122. http://dx.doi.org/10.1016/j.psychres.2005.02.005

Nunnally, J. C., Bernstein, I. H., & Berge, J. M. F. (1967). *Psychometric theory* (McGraw-Hill Series in Psychology, Vol. 226). Available at https://www.jstor.org/

Olvet, D. M., & Hajcak, G. (2010). The effect of trial-to-trial feedback on the error-related negativity and its relationship with anxiety. *Cognitive, Affective & Behavioral Neuroscience, 10,* 157.

Riesel, A. (2019). The erring brain: Error-related negativity as an endophenotype for OCD-A review and meta-analysis. *Psychophysiology, 56*(4): e13348. http://dx.doi.org/10.1111/psyp.13348

Riesel, A., Klawohn, J., Grützmann, R., Kaufmann, C., Heinzel, S., Bey, K., . . . Kathmann, N. (2019). Error-related brain activity as a transdiagnostic endophenotype for obsessive-compulsive disorder, anxiety and substance use disorder. *Psychological Medicine, 49,* 1207–1217. http://dx.doi.org/10.1017/S0033291719000199

Riesel, A., Klawohn, J., Kathmann, N., & Endrass, T. (2017). Conflict monitoring and adaptation as reflected by N2 amplitude in obsessive-compulsive disorder. *Psychological Medicine, 47,* 1379–1388. http://dx.doi.org/10.1017/S0033291716003597

Weinberg, A., Dieterich, R., & Riesel, A. (2015). Error-related brain activity in the age of RDoC: A review of the literature. *International Journal of Psychophysiology, 98,* 276–299. http://dx.doi.org/10.1016/j.ijpsycho.2015.02.029

Weinberg, A., Klein, D. N., & Hajcak, G. (2012). Increased error-related brain activity distinguishes generalized anxiety disorder with and without comorbid major depressive disorder. *Journal of Abnormal Psychology, 121,* 885–896. http://dx.doi.org/10.1037/a0028270

Weinberg, A., Meyer, A., Hale-Rude, E., Perlman, G., Kotov, R., Klein, D. N., & Hajcak, G. (2016). Error-related negativity (ERN) and sustained threat: Conceptual framework and empirical evaluation in an adolescent sample. *Psychophysiology, 53,* 372–385. http://dx.doi.org/10.1111/psyp.12538

Weinberg, A., Olvet, D. M., & Hajcak, G. (2010). Increased error-related brain activity in generalized anxiety disorder. *Biological Psychology, 85,* 472–480. http://dx.doi.org/10.1016/j.biopsycho.2010.09.011

Wuensch, K. L. (2012). *Using SPSS to obtain a confidence interval for Cohen's d*. Retrieved from http://core.ecu.edu/psyc/wuenschk/SPSS/CI-d-SPSS.pdf

Received December 6, 2018
Revision received May 6, 2019
Accepted June 13, 2019 ■