Psychophysiology, 50 (2013), 1220–1225. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/psyp.12132

BRIEF REPORT

Reliability of the ERN across multiple tasks as a function of increasing errors

ALEXANDRIA MEYER,[a] ANJA RIESEL,[b] AND GREG HAJCAK PROUDFIT[a]

[a]Department of Psychology, Stony Brook University, Stony Brook, New York, USA
[b]Institut für Psychologie, Humbolt University, Berlin, Germany

Abstract

Event-related potential (ERP) studies of error-processing have characterized the error-related negativity (ERN) as a negative deflection occurring after the commission of an error at frontocentral sites. The ERN has frequently been examined in the context of individual differences and has been proposed as a neurobehavioral risk marker. Given this, it is important to characterize the psychometric properties of the ERN across multiple tasks as a function of increasing trial numbers in order to establish task-specific psychometric properties for efficient assessments in clinical or applied settings. The current study examines the internal reliability of the ERN across the flankers, Stroop, and go/no-go tasks as a function of error number. Results suggest that although the tasks all elicit the ERN reliably, important psychometric differences emerged indicating that the flankers task might be prioritized when assessing the ERN.

Descriptors: Anxiety, Individual differences, EEG/ERP, Psychometrics, Error-related negativity, Error processing

A flexible system for detecting errors is necessary for learning and for the mobilization of defensive responses in changing environments (Hajcak, 2012; Holroyd & Coles, 2002). Event-related potential (ERP) studies of action monitoring over the past 20 years have utilized the error-related negativity (ERN), a negative deflection appearing approximately 50 ms after the commission of an error that is maximal at frontocentral sites, to study error detection in humans (Falkenstein, Hohnsbein, Hoormann, & Blanke, 1991; Gehring, Goss, Coles, Meyer, & Donchin, 1993).

The ERN is thought to reflect the activity of a general error detection system that becomes active across a range of response and stimulus modalities (Gehring et al., 1993; van Veen & Carter, 2002), and a number of studies have examined error processing in the context of individual differences. For instance, the ERN appears increased in both obsessive-compulsive disorder (Endrass, Klawohn, Schuster, & Kathmann, 2008; Gehring, Himle, & Nisenson, 2000) and generalized anxiety disorder (Weinberg, Olvet, & Hajcak, 2010), as well as in relationship to personality traits that characterize anxiety, such as high negative affect (Hajcak, McDonald, & Simons, 2004), worry (Hajcak, McDonald, & Simons, 2003), and behavioral inhibition (Amodio, Master, Yee, & Taylor, 2008). In line with these findings, the ERN has been proposed as an endophenotype (Olvet & Hajcak, 2008) or neurobehavioral trait (Hajcak, 2012; Weinberg, Riesel, & Hajcak, 2012) that may be useful in identifying trajectories of risk for anxiety disorders (Meyer, Weinberg, Klein, & Hajcak, 2012).

The validity of an individual difference variable hinges on its reliability (Cronbach & Meehl, 1955). Reliability is defined as the tendency of a measure to reflect an individual's true score, and can be measured in three different ways: internal consistency, test-retest, and alternate forms. For an ERP component, internal consistency is indicated by the homogeneity of the ERP metric across trials within a single task. Although ERP components are derived by averaging many trials, if the trial-to-trial waveforms are unreliable, the average will also be unreliable (Simons & Miles, 1990). Existing research suggests that the ERN has good internal reliability (Larson, Baldwin, Good, & Fair, 2010; Olvet & Hajcak, 2009b), and high test-retest reliability over periods of weeks (Olvet & Hajcak, 2009b; Segalowitz et al., 2010), and even up to 2 years (Weinberg & Hajcak, 2011).

All of these studies assessed the psychometric properties of the ERN using variants of the flankers task; however, individual difference studies that employ the ERN have utilized a range of tasks—including probabilistic learning (Gründler, Cavanagh, Figueroa, Frank, & Allen, 2009; Nieuwenhuis, Nielen, Mol, Hajcak, & Veltman, 2005), go/no-go (Menon, Adleman, White, Glover, & Reiss, 2001; Torpey, Hajcak, & Klein, 2009), and Stroop (Hajcak et al., 2003; Hajcak & Simons, 2002). On the one hand, this may not be an issue because the ERN appears to have reasonable convergent validity (Segalowitz et al., 2010), and recent evidence suggests the ERN derived from different tasks tends to be at least moderately correlated, flankers and Stroop, $r = .37$, flankers and go/no-go, $r = .43$, and Stroop and go/no-go, $r = .33$ (Riesel, Weinberg, Endrass, Meyer, & Hajcak, 2013). However, the

Address correspondence to: Alexandria Meyer, Department of Psychology, Stony Brook University, Stony Brook, NY 11794-2500, USA. E-mail: ammeyer3@gmail.com

relationship between the ERN and individual differences in anxiety appears to be at least somewhat task-dependent (Gründler et al., 2009; Olvet & Hajcak, 2009a). Variability in the relationship between the ERN and individual difference measures could be attributable to differential reliability of the ERN across tasks. Indeed, a recent study found that an ERN task with worse psychometric properties was less related to individual differences in relation to psychosis (Foti, Kotov, & Hajcak, 2013). Additionally, recent evidence suggests that the internal reliability of the ERN is higher when recorded during a flankers task compared to both a go/no-go and Stroop. Specifically, the split-half reliabilities in the flankers, Stroop, and go/no-go tasks were $r = .81$, $r = .69$, and $r = .60$, respectively (Riesel et al., 2013).

An issue related to reliability of the ERN concerns the number of error trials necessary for a stable ERN—an issue particularly relevant because of low error rates in most speeded response tasks. Previous work from our group suggests that the ERN becomes stable after approximately 6 trials (Olvet & Hajcak, 2009b), and other studies have extended these findings to children and older adults (Pontifex et al., 2010). However, no studies to date have examined internal reliability of the ERN across multiple tasks, as a function of error number. Although tasks may have comparable psychometric properties when considering all error trials, some tasks may achieve adequate psychometric properties with fewer errors. In addition to helping characterize the psychometric properties of the ERN across multiple tasks, these data would have practical utility: assessments could be shorter if excellent psychometric properties can be attained with fewer error trials. This is particularly important when the ERN is assessed in more clinical or applied settings. Thus, the current study builds on and extends our previous study (Riesel et al., 2013) by directly comparing the ERN and its psychometric properties when derived from the flankers, Stroop, and go/no-go tasks as a function of increasing trial numbers.

Method

Participants

Forty-seven undergraduate students (20 female) from Stony Brook University participated in this study. Two participants committed fewer than six errors in at least one task and were therefore excluded from further analysis. Data from two subjects were excluded due to excessive electroencephalogram (EEG) artifacts. The final sample consisted of 43 participants (19 female). The mean age was 19.14 years ($SD = 1.42$); 38.6% of the sample was Caucasian/European, 45.5% was Asian-American, 6.8% was Hispanic, 2.3% was African-American, and 6.8% identified as "other."

Task and Procedure

The experiment consisted of three counterbalanced tasks: a modified flanker task, a go/no-go task, and a Stroop task. Tasks were administered using Presentation software (Neurobehavioral Systems, Inc., Albany, CA). Before each task, participants completed a practice block of 20 trials. All tasks consisted of 420 trials presented in seven blocks of 60 trials. Stimuli were presented for 200 ms with an intertrial interval that varied between 600 and 1,000 ms. Feedback was presented at the end of each block to encourage fast and accurate behavior. If performance was 75% correct or lower, the message, "Please try to be more accurate," was displayed; if performance was above 90% correct, the message, "Please try to respond faster," was displayed; otherwise the message, "You're doing a great job," was displayed.

Flankers task. On each trial, horizontally aligned arrowheads were presented: half of the trials were compatible (>>>>> or <<<<<) and half were incompatible (<<><< or >><>>); the order of trials was randomly determined. Participants were told to press the right mouse if the center arrow was facing to the right and to press the left mouse button if the center arrow was facing to the left. Viewing distance was approximately 65 cm, and the set of arrows filled 2° of visual angle vertically and 10° horizontally.

Stroop task. On each trial, one of three words was shown ("red," "green," or "blue"), and was presented in either red or green font. Participants were instructed to press the left mouse button if the word was presented in red and press the right button if the word was presented in green. Therefore, 1/3 of the trials were compatible (e.g., color word and font color require the same response), 1/3 were incompatible (e.g., color word and font color require different responses), and 1/3 were neutral (e.g., the color word "blue" in red or green font). Each word occupied between 2° and 3° of visual angle.

Go/no-go task. On each trial, a green triangle was presented. Participants were instructed to press the right mouse button in response to an upright triangle, occurring on 80% of trials. Participants were told to withhold responses to tilted triangles (10°), occurring on 20% of trials. Each triangle occupied 3° × 3° of visual angle.

pDetails regarding psychophysiological recording and ERP analyses are described in full elsewhere (Riesel et al., 2013).

Data Analysis

Statistical analyses were conducted using a significance level of $p = .05$. Analyses with three or more within-subjects levels used the Greenhouse-Geisser statistic with Bonferroni correction for multiple post hoc comparisons. We computed the ERN as a function of increasing trial numbers, deriving the ERN based on the first 2, 4, 6, 8, 10, 12, 14, 16, 18, and 20 error trials. A repeated measures analysis of variance (ANOVA) was performed with these averages across task as within-subject variables to determine if the ERN varied significantly within each task or as a function of increasing error number. We then correlated these averages with the grand-average ERN using Pearson's correlation coefficient. This metric suggests the degree to which the ERN based on a subset of error trials relates to the grand-average ERN. In addition, we calculated Cronbach's alpha (the average of all possible split-half reliabilities) as a function of the number of error trials. Because the number of error trials varied across participants, the full sample was only available when calculating alpha using the first 14 error trials. After this, as more trials were entered into the calculation (between 14 and 20), the number of participants included decreased differentially in each task. Overall, 41 participants committed at least 20 errors during the flankers and Stroop task, and 40 participants committed at least 20 errors during the go/no-go task.

Results

Figure 1 depicts ERN as a function of trial number for each task, and Figure 2 depicts the grand-average ERP waveform at FCz for differential trial numbers. The 3 (Task) × 10 (ERP average)

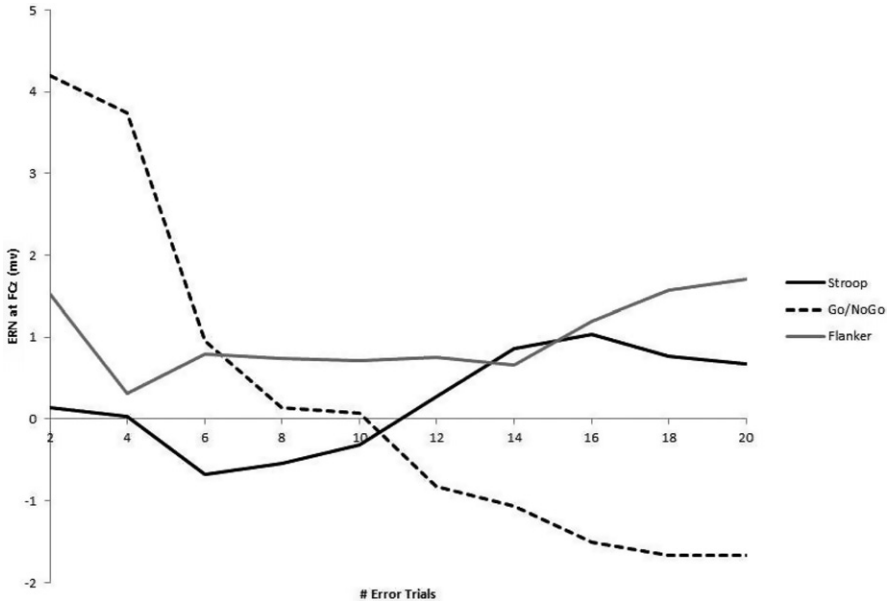

Figure 1. Magnitude of the ERN at FCz as a function of error number for each task.

repeated measures ANOVA revealed a significant interaction, $F(18,666) = 6.07$, $p < .001$, suggesting that the impact of increasing errors differentially impacted the ERN as a function of task. Follow-up ANOVAs were performed for each task to examine differences between smaller trial averages as two trials were increasingly added to ERN averages. For both the flankers and Stroop task, there were no significant differences when comparing the averages of increasing trial numbers (2, 4, 6, 8, 10, 12, 14, 16, 18, 20 error trials), $F(9,342) = .171$, $p = .84$, and $F(9,342) = .82$, $p = .46$, respectively. However, for the go/no-go task, the ERN averages of increasing trial number did differ, $F(1,342) = 12.26$, $p < .01$. This effect is evident visually in Figure 1, wherein the ERN for Stroop and flankers appear relatively stable, whereas the ERN for the go/no-go task decreases systematically as more error trials are added to the average. This effect is also illustrated in Figure 2, in which the ERN becomes more negative as more trials are added. Figure 3 presents correlation coefficients between the grand-average ERN and the ERN based on the average of fewer trials. All correlations are moderate to high, increasing as more trials were added (all correlations were significant at $p < .001$). For the flankers task, 8 error trials were required for the ERN to correlate .80 with the grand average, whereas 12 and 18 errors trials were required for the go/no-go and Stroop tasks, respectively.

Figure 4 presents Cronbach's alpha for the ERN for each task as more error trials were examined. A Cronbach's alpha exceeding .90 suggests excellent internal reliability, between .70 and .90 suggests high internal reliability, between .50 and .70 indicates moderate internal reliability, and below .50 low reliability. As can be seen in Figure 4, moderate internal reliability was achieved in both the flankers and go/no-go, after approximately 10 errors were committed. However, the reliability of the Stroop task remained low even after 20 errors per subject were examined.[1]

1. As previously reported in Riesel et al., 2013, participants committed significantly fewer errors during the go/no-go task ($M = 30.44$, $SD = 10.06$) than on the flankers ($M = 52.95$, $SD = 25.38$) or Stroop task ($M = 55.05$, $SD = 31.14$).

Discussion

The current study found that although the flankers, go/no-go, and Stroop task all elicit a reliable ERN, important psychometric differences between these tasks emerge when the number of errors are considered. For instance, the go/no-go task was reliable across participants: within the first 6 to 8 errors, the average ERN at FCz was highly correlated with the grand average ($r > .80$), the split-half correlations at frontocentral sites were moderate to high ($r = .70$–$.80$), and Cronbach's alpha achieved moderate reliability (.70) after 12 errors, suggesting good overall reliability. However, as can be seen in Figures 1 and 2, the magnitude of the ERN increased dramatically as more errors were committed. The high Pearson and split-half correlations and Cronbach's alpha suggest that this pattern of increasing ERN across the first 20 errors was consistent across participants—evident in the ERP averages in Figure 2. These data suggest that the magnitude of the ERN in the go/no-go task is highly dependent on the number of error trials included in averages (i.e., subjects making relatively few errors should have a smaller ERN compared to subjects making many errors in this task). Considering that participants committed significantly fewer errors during the go/no-go task than on the flankers or Stroop task, this may be particularly relevant. Performance-based differences should, therefore, be carefully evaluated in future studies utilizing go/no-go tasks to elicit the ERN in relation to individual differences, especially when fewer than 20 errors are committed.

In contrast to the go/no-go task, an examination of the psychometric properties of the ERN in the Stroop task suggested that although the average of the entire sample did not vary much within the first 20 errors, reliability was low: the Stroop task did not attain a Cronbach's alpha over .50, even after 20 errors were committed, suggesting the existence of a substantial amount of trial-to-trial variation in the ERN within this task.

Consistent with previous studies (Larson et al., 2010; Olvet & Hajcak, 2009b; Pontifex et al., 2010), the current analyses indicate that the ERN derived from the flankers task is highly reliable. The average ERN did not vary substantially across the task as more

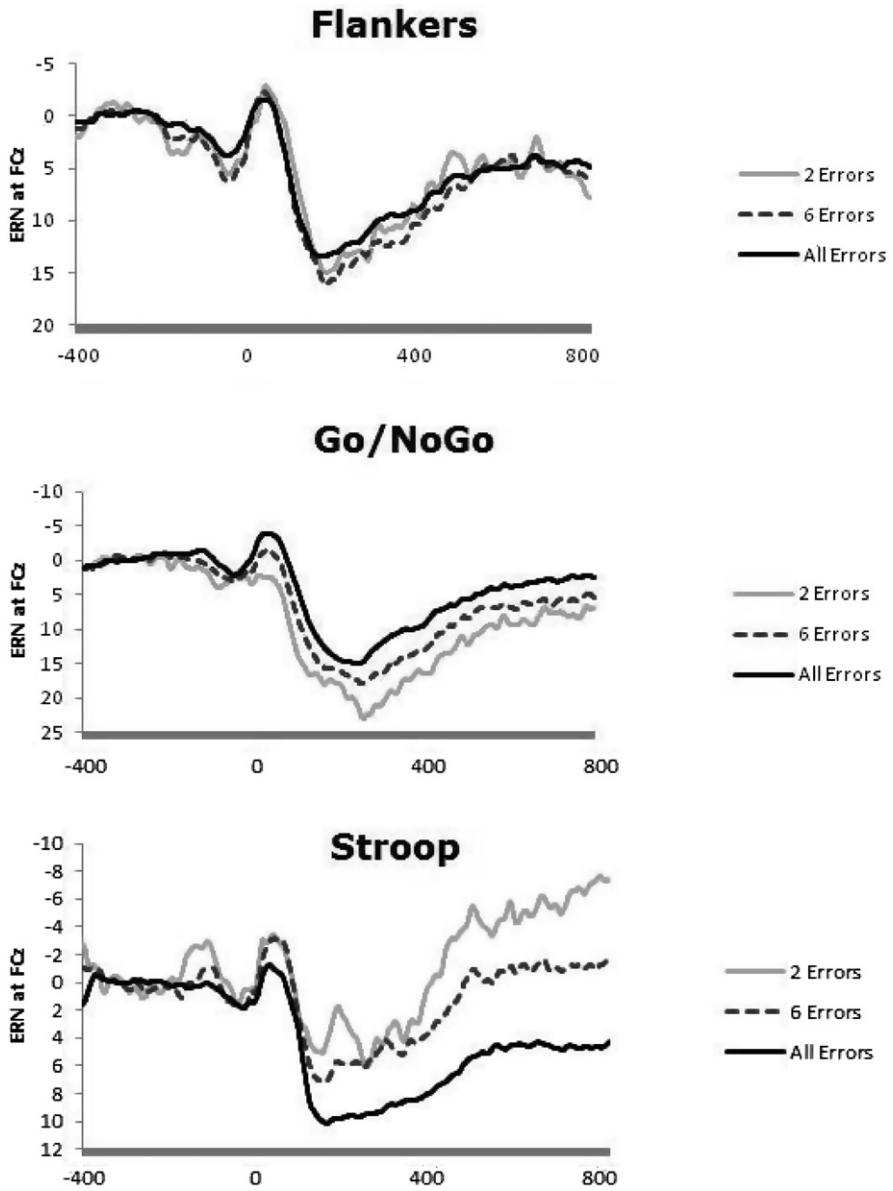

Figure 2. Grand-average ERP waveform at FCz for differential trial numbers for each task.

errors were added, and the magnitude of the ERN was consistent within individuals—high Cronbach's alpha was obtained after a relatively low number (e.g., seven) of errors. Taken together, the flankers task reached high reliability quickly and did not vary as more error trials were added—and, therefore, this task might be prioritized in studies seeking to relate the ERN to individual differences.

The differences in reliability observed may have been due to differences in cognitive control strategies. For example, the go/no-go task requires participants to inhibit any response on no-go trials, whereas in the Stroop and flankers task, participants make a response on every trial but must inhibit visual distractor elements some of the time. Consistent with this possibility, one meta-analysis suggests these three tasks engage differential patterns of neural activation based on the processing stage in which interference is being resolved (Nee, Wager, & Jonides, 2007).

It is also possible that the differences in reliability observed may be related to differences in structure between the three tasks. For example, during the go/no-go task there are only four practice trials during which participants could make an error of commission (no-go trials), compared to 20 practice trials for the flankers and Stroop tasks. Given that the ERN increases over the course of learning as errors become less expected (Holroyd & Coles, 2002), this relative lack of practice may account for the increase in ERN amplitude during the early portion of the go/no-go task. Furthermore, half of the trials were incompatible in the flankers task, whereas one in three trials were incompatible in the Stroop task. It is possible that variation in amount and type of interference across these tasks contributed to differences in internal reliability.

Considering that the validity of an individual difference variable hinges on its reliability, it will be important for future research to determine if the differential reliability of the ERN in these three tasks influences its relationship with anxiety. For instance, some evidence suggests that reliability of the ERN may differ between normative and clinical populations (Foti et al., 2013). Given the

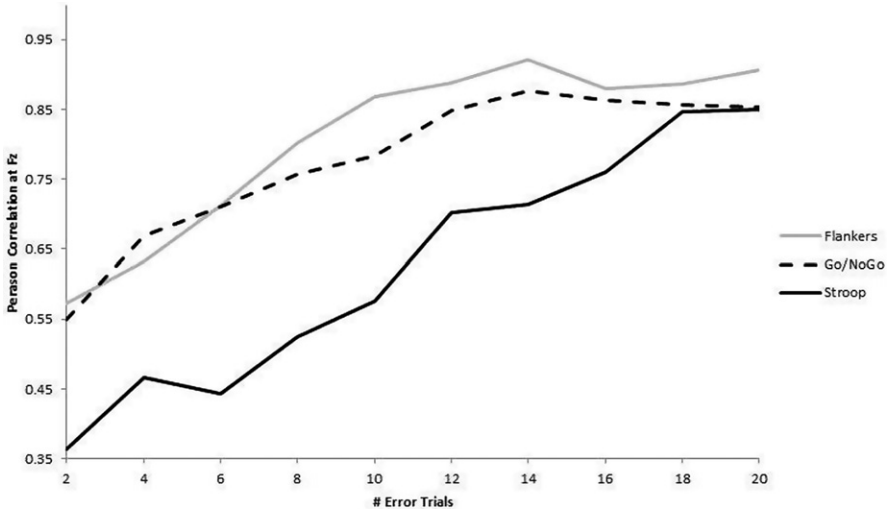

Figure 3. Correlation coefficients between the grand-average ERN and the ERN based on the average of fewer trials.

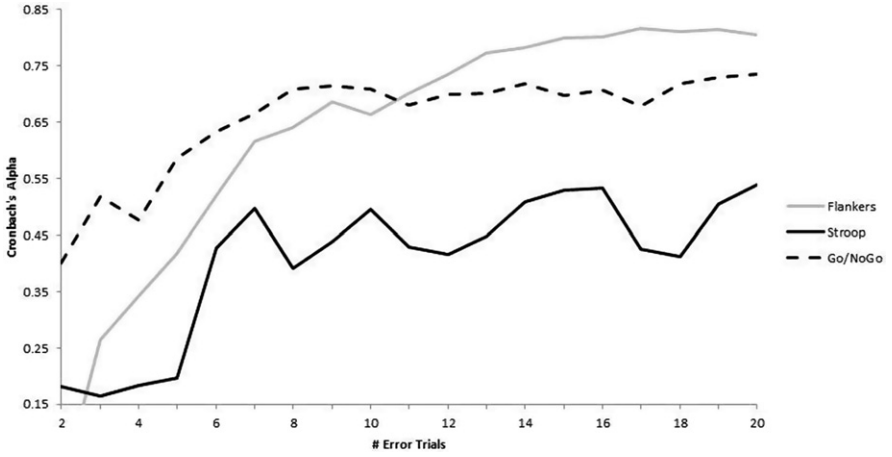

Figure 4. Cronbach's alpha of the ERN at FCz as a function of the number of error trials for each task.

current findings, we would expect the flankers task to be best suited for capturing individual differences in ERN related to anxiety and psychopathology. However, it is also possible that the different cognitive and neural processes associated with each task could impact the task-specific relationship between the ERN and individual difference measures.

It is important to note that the reliability of tasks in measuring the ERN may vary across the lifespan. Whereas this study examined reliability in a college-aged sample and previous studies have examined the reliability of the flankers task (Pontifex et al., 2010) and go/no-go task (Segalowitz et al., 2010) in older children, it will be important to compare the reliability of different tasks among younger children. This may be especially important when examining the ERN in relationship to development and trajectories of risk for anxiety disorders (Hajcak, 2012; Meyer et al., 2012).

References

Amodio, D. M., Master, S. L., Yee, C. M., & Taylor, S. E. (2008). Neurocognitive components of the behavioral inhibition and activation systems: Implications for theories of self-regulation. *Psychophysiology*, *45*, 11–19.doi: 10.1111/j.1469-8986.2007.00609.x

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. doi: 10.1037/h0040957

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. doi: 10.1016/j.neuropsychologia.2007.12.001

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*, 447–455. doi: 10.1016/0013-4694(91)90062-9

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*, 520. doi: 10.1037/a0032618

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

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

Gründler, T. O. J., Cavanagh, J. F., Figueroa, C. M., Frank, M. J., & Allen, J. B. (2009). Task-related dissociation in ERN amplitude as a function of obsessive–compulsive symptoms. *Neuropsychologia*, *47*, 1978–1987. doi: 10.1016/j.neuropsychologia.2009.03.010

Hajcak, G. (2012). What we've learned from mistakes. *Current Directions in Psychological Science*, *21*, 101–106. doi: 10.1177/0963721412436809

Hajcak, G., McDonald, N., & Simons, R. F. (2003). Anxiety and error-related brain activity. *Biological Psychology*, *64*, 77–90. doi: 10.1016/s0301-0511(03)00103-0

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

Hajcak, G., & Simons, R. F. (2002). Error-related brain activity in obsessive-compulsive undergraduates. *Psychiatry Research*, *110*, 63–72. doi: 10.1016/s0165-1781(02)00034-3

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679–709. doi: 10.1037/0033-295x.109.4.679

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*, 1167–1171. doi: 10.1111/j.1469-8986.2010.01022.x

Menon, V., Adleman, N. E., White, C. D., Glover, G. H., & Reiss, A. L. (2001). Error-related brain activation during a Go/NoGo response inhibition task. *Human Brain Mapping*, *12*, 131–143. doi: 10.1002/1097-0193(200103)12:3<131::aid-hbm1010>3.0.co;2-c

Meyer, A., Weinberg, A., Klein, D. N., & Hajcak, G. (2012). The development of the error-related negativity (ERN) and its relationship with anxiety: Evidence from 8 to 13 year-olds. *Developmental Cognitive Neuroscience*, *2*, 152–161. doi: 10.1016/j.dcn.2011.09.005

Nee, D. E., Wager, T. D., & Jonides, J. (2007). Interference resolution: Insights from a meta-analysis of neuroimaging tasks. *Cognitive, Affective, & Behavioral Neuroscience*, *7*, 1–17.

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

Olvet, D., & Hajcak, G. (2009a). The effect of trial-to-trial feedback on the error-related negativity and its relationship with anxiety. *Cognitive, Affective, & Behavioral Neuroscience*, *9*, 427–433. doi: 10.3758/cabn.9.4.427

Olvet, D. M., & Hajcak, G. (2008). The error-related negativity (ERN) and psychopathology: Toward an endophenotype. *Clinical Psychology Review*, *28*, 1343–1354. doi: 10.1016/j.cpr.2008.07.003

Olvet, D. M., & Hajcak, G. (2009b). The stability of error related brain activity with increasing trials. *Psychophysiology*, *46*, 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., 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*, 767–773. doi: 10.1111/j.1469-8986.2010.00974.x

Riesel, A., Weinberg, A., Endrass, T., Meyer, A., & Hajcak, G. (2013). The ERN is the ERN is the ERN? Convergent validity of error-related brain activity across different tasks. *Biological Psychology*, *93*, 377–385. doi: 10.1016/j.biopsycho.2013.04.007

Segalowitz, S. J., Santesso, D. L., Murphy, T. I., Homan, D., Chantziantoniou, D. K., & Khan, S. (2010). Retest reliability of medial frontal negativities during performance monitoring. *Psychophysiology*, *47*, 260–270. doi: 10.1111/j.1469-8986.2009.00942.x

Simons, R. F., & Miles, M. A. (1990). Nonfamilial strategies for the identification of subjects at risk for severe psychopathology: Issues of reliability in the assessment of event-related potential and other marker variables. In J. W. Rohrbaugh, R. Parasuraman, & R. Johnson, Jr. (Eds.), *Event-related brain potentials: Basic issues and applications* (pp. 343–363). New York, NY: Oxford University Press.

Torpey, D. C., Hajcak, G., & Klein, D. N. (2009). An examination of error-related brain activity and its modulation by error value in young children. *Developmental Neuropsychology*, *34*, 749–761. doi: 10.1080/87565640903265103

van Veen, V., & Carter, C. S. (2002). The anterior cingulate as a conflict monitor: fMRI and ERP studies. *Physiology & Behavior*, *77*, 477–482. doi: 10.1016/s0031-9384(02)00930-7

Weinberg, A., & Hajcak, G. (2011). Longer term test–retest reliability of error-related brain activity. *Psychophysiology*, *48*, 1420–1425. doi: 10.1111/j.1469-8986.2011.01206.x

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

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*, 84–100. doi: 10.1007/s11031-011-9269-y

(Received April 11, 2013; Accepted June 30, 2013)