Psychometric Reliability of ERN and Pe Across Flanker, Stroop, and Go/No-Go Tasks: A Direct and Conceptual Replication

Amanda Holbrook[1], Bohyun Park[1], Scott A. Baldwin[2], Anja Riesel[3], Michael J. Larson[2,4], and Peter Clayson*[1]

[1] Department of Psychology, University of South Florida, Tampa, FL, USA

[2] Department of Psychology, Brigham Young University, Provo, Utah

[3] Department of Psychology, Universität Hamburg, Hamburg, Germany

[4] Neuroscience Center, Brigham Young University, Provo, Utah

**Corresponding author at*: Department of Psychology, University of South Florida, 4202 East Fowler Avenue, Tampa, FL, US, 33620-7200. *Email:* clayson@usf.edu

Abstract

The effectiveness of error-related negativity (ERN) in assessing individual differences hinges on its psychometric reliability. Despite evidence that task used to record ERN moderates reliability, this moderation is rarely examined within the same sample, risking inaccurate generalizations of psychometrics. A direct and conceptual replication of Meyer et al. (2013, *Psychophysiology*) was conducted in 182 participants to assess internal consistency of ERN from flanker, go/no-go, and Stroop tasks as a function of increasing trials. Analyses were extended to include error positivity (Pe) and difference scores (ΔERN, ΔPe), and generalizability theory and multilevel models were used to statistically compare internal consistency across tasks. Overall, data supported the internal consistency of results across three tasks in a healthy undergraduate sample, with values ranging from .70 to .97 when examining all data. However, estimates were in part outside the confidence intervals of the original study, and ERN scores showed lower internal consistency than previously reported for a flanker task and higher internal consistency than previously reported for a Stroop task. Pe score internal consistency was similar across tasks when examining the average number of error trials. These findings underscore the importance of examining reliability in each study rather than relying on universal trial cutoffs. Overall, a flanker task may be better suited for studies of ERN due to higher internal consistency of ERN scores when including data from all error trials, However, exclusively using a single task is discouraged because understanding the functional significance of ERN and Pe requires considering task-specific nuances and the varying contributions of cognitive processes, such as cognitive control or response inhibition.

Keywords: psychometrics; internal consistency; event-related potential (ERP); ERP psychometric reliability; error-related negativity (ERN); error positivity (Pe)

1. Introduction

Event-related brain potentials (ERPs), particularly error-related negativity (ERN), provide insights into performance monitoring—the ability to monitor ongoing performance and adapt behavior in line with goals (Larson et al., 2014). ERN is a negative-going deflection in the scalp-recorded ERP that peaks within 100 ms of an erroneous response and is larger for error responses than for correct responses (for review, see Larson et al., 2014). Whether ERN is suitable as a measure of individual differences depends on psychometric internal consistency (Clayson, 2024b; Clayson, Brush, & Hajcak, 2021; Hajcak et al., 2017). Although internal consistency is a characteristic of observed data, the internal consistency of ERN scores is often inferred based on published psychometric information. However, this inference erroneously assumes that the internal consistency of ERN scores remains stable across various contexts, including sample characteristics, recording/analysis procedures, and the task used to record ERN (Clayson & Miller, 2017b). Therefore, the present study aimed to determine whether ERN scores yield similar internal consistency within the same sample of participants across flanker, go/no-go, and Stroop tasks, which are commonly used to examine individual differences in performance monitoring.

The functional role of ERN is explained by several theoretical frameworks (Falkenstein et al., 2000; Holroyd & Coles, 2002). A prominent theory proposes that ERN amplitude reflects the detection of competing response options on a single trial (Botvinick et al., 2001). Greater conflict detection may be reflective of greater cognitive control, supporting adaptive changes aimed at preventing future errors (Larson et al., 2012). ERN is sensitive to individual differences in clinical characteristics (Martin et al., 2018; Michael et al., 2021; Pasion & Barbosa, 2019; Riesel, 2019) and may aid in predicting first onset of psychopathology (Meyer et al., 2015, 2018) and prospective clinical symptom severity (Banica et al., 2020), and understanding familial risk for psychopathology (Carrasco et al., 2013; Morand-Beaulieu et al., 2024). However, the clinical utility of ERN is threatened by lower internal consistency of ERN scores found in psychopathology groups than healthy controls (Clayson, 2020). Identifying moderators that optimize the internal consistency of ERN scores is crucial for enhancing the clinical utility of ERN.

Meta-analytic work examining 189 internal consistency estimates from 43 studies revealed that task moderated ERN's internal consistency (Clayson, 2020). When analyzing adjusted coefficient alphas using eight trials, task significantly moderated internal consistency: Go/no-go tasks yielded higher estimates ($\hat{\alpha}$ = .74) than flanker tasks ($\hat{\alpha}$ =.67), Stroop tasks ($\hat{\alpha}$ =.56), and picture/word tasks ($\hat{\alpha}$ = .27), although go/no-go tasks yielded estimates similar to those from Simon tasks ($\hat{\alpha}$ = .73). These findings indicate that ERN scores from a go/no-go task may yield the highest internal consistency at low trial counts. This meta-analysis demonstrated how the task used for recording ERN impacts internal consistency. However, within-person variations must also be investigated, since the meta-analysis was based on between-study comparisons, leaving open the possibility that non-task related factors contributed to differences in ERN score internal consistency.

Few studies have examined the possibility of ERN yielding different estimates of internal consistency for different tasks in the same sample of participants. Meyer et al. (2013) examined the internal consistency of ERN scores across flanker, go/no-go, and Stroop tasks. They showed that ERN from a flanker task obtained adequate internal consistency (i.e., 0.7 coefficient alpha) with as few as 7 trials and that a go/no-go task required at least 12 trials. Notably, ERN scores from a Stroop task never obtained internal consistency above .5 even with as many as 20 trials included. Meyer et al. (2013) concluded that a flanker task may be the best suited task of the three paradigms for studying individual differences in ERN due to its superior psychometric properties in this study. This conclusion has likely influenced the widespread adoption of flanker tasks for recording ERN. Indeed, recent reviews and meta-analyses of the relationship between ERN and mental health symptoms suggest flanker tasks are used in approximately 33-77% of ERN studies (Martin et al., 2018; Michael et al., 2021; Pasion & Barbosa, 2019; Riesel, 2019). Therefore, the present study sought to replicate Meyer et al. (2013) by testing within-participant task differences in the internal consistency of ERN scores in a large sample.

A limitation of Meyer et al. (2013) was relying solely on visual inspection of internal consistency estimates, rather than statistical comparison, to justify conclusions that ERN internal consistency differs between tasks. Another limitation was the use of coefficient alpha for internal consistency of ERN scores as a function of increasing trial numbers, which can be problematic, as the use of coefficient alpha requires participants to have an equal number of observations (i.e., error trials). This requirement excludes trials and participants unequally (e.g., some participants are included in analysis at low trial counts and excluded at higher trial counts), reducing generalizability and potentially altering between-person variance. The present study addresses these limitations by statistically comparing estimates of internal consistency. Specifically, Bayesian multilevel location-scale models and generalizability theory estimate between- and within-subject variance components from all available trials, tasks, and participants (Baldwin et al., 2015; Clayson et al., 2022; Clayson & Miller, 2017b). By simultaneously estimating the internal consistency of ERN from all tasks in the same model, post-estimation contrasts can then be used to statistically evaluate whether tasks differ in internal consistency estimates (Clayson et al., 2024).

Finally, a limitation of previous psychometric studies is the exclusion of other performance monitoring ERPs, particularly error positivity (Pe). Pe is a positive-going ERP that peaks 200-400 ms following an erroneous response and is associated more frequently than ERN with the conscious processing of errors (Orr & Carrasco, 2011; Overbeek et al., 2005). Like ERN, Pe shows promise as an individual difference measure of psychopathology but has been studied less frequently (Hanna et al., 2024; Lutz et al., 2021; Macedo et al., 2021; Vallet et al., 2021). Studies of the internal consistency of Pe have primarily focused on flanker tasks, finding that adequate internal consistency, as measured by coefficient alpha, can be achieved with as few as 6-12 trials[1] (Olvet & Hajcak, 2009; Pontifex et al., 2010; Rietdijk et al., 2014). However, it remains unclear how internal consistency differs for Pe scores measured during Stroop and go/no-go tasks. Including ERN and Pe in the same study is critical for allowing direct comparisons of internal consistency and examining how moderators, such as task used for recording, affect different stages of performance monitoring, which is key for understanding the functional significance and clinical utility of these indices.

1.1. Present Study

The present study directly compares whether the internal consistency of ERN recorded during a flanker task is superior to the internal consistency of ERN recorded during go/no-go and Stroop tasks. This comparison is important because flanker tasks remain among the most widely used tasks for studying individual differences in ERN, due in part to the influential research by Meyer et al. (2013). Therefore, the present study sought to confirm the findings of Meyer et al. (2013) and support ongoing work using ERN as a measure of individual differences in performance monitoring. In addition to a direct and conceptual replication of the original study,

[1] WebPlotDigitizer (Rohatgi, 2020) was used to retrieve data points from Figure 4c in Rietdijk et al. (2019). An α of .7 was obtained for Pe with 12 error trials.

the present study extends the scope of analysis by examining Pe and difference score activity (error minus correct).

Aim 1. The present study directly replicated Meyer et al. (2013) to determine the number of trials needed for adequate internal consistency of ERN scores recorded during flanker, Stroop, and go/no-go tasks using a classical test theory estimate (i.e., coefficient alpha [α]). We expected to replicate the findings from Meyer et al. (2013) such that ERN scores from flanker and go/no-go tasks would reach an α of at least .7 with 7 and 12 trials included in the average, respectively[2]. We expected ERN scores from the Stroop task not to reach $\alpha \geq .7$ even with 20 trials included in estimation. α was estimated including up to 20 trials to be consistent with Meyer et al. (2013).

Aim 2. The second aim was to extend the findings of Meyer et al. (2013) by conducting a conceptual replication using generalizability theory estimates of internal consistency (i.e., dependability [ϕ]) to determine the number of trials needed for ERN scores to reach adequate internal consistency ($\geq .7$). We expected internal consistency (i.e., dependability) of ERN scores to reach .7 when including at least 7 trials for the flanker task and 12 trials for the go/no-go task, but not for the Stroop task even when including 20 trials.

Additional exploratory analyses were conducted on ERN and Pe. ERN scores, Pe scores, their correct-trial counterparts (correct-response negativity [CRN] and post-correct positivity [Pc]), and their difference scores (ΔERN, ΔPe; error minus correct) were analyzed to determine the number of trials required to reach three thresholds of internal consistency (.7, .8, .9) using

[2] WebPlotDigitizer (Rohatgi, 2020) was used to retrieve data points from Figure 4 in Meyer et al. (2013). An α of .7 was obtained with 11 and 8 trials for the flanker and go/no-go task, respectively. We noted this discrepancy in the preregistration and a priori decided to base our hypotheses off the description provided in the Discussion section of Meyer et al. (2013).

generalizability theory. To include a classical test-theory alternative for estimation of internal consistency of difference scores, exploratory analyses of residualized difference scores were conducted for all three tasks to determine how many trials are required to reach three thresholds of internal consistency (.7, .8, .9; Clayson, Baldwin, & Larson, 2021).

Aim 3. The third aim was to statistically compare task differences in internal consistency (i.e., dependability and intraclass correlation coefficient [ICC]) of ERP scores. Post-estimation contrasts from the Bayesian multilevel models were used to compare internal consistency of ERN scores for all three tasks. We expected that dependability estimates would differ between tasks, such that ERN and ΔERN scores from the flanker and go/no-go tasks would demonstrate higher internal consistencies than the Stroop task when including 20 error trials. We also expected to find task differences in ICC estimates such that ERN from the flanker and go/no-go tasks would demonstrate higher internal consistency (i.e., ICC) than the Stroop task when using a trial-independent estimate of reliability. Additional task comparisons using ICCs and dependability estimates for CRN, Pe, Pc, ΔPe, and ΔERN were conducted as exploratory analyses.

2. Method

The present manuscript is a secondary data analysis of data collected as part of a registered report (https://osf.io/8cbua/; Clayson et al., 2023). The Clayson et al. (2023) study evaluated the convergent and divergent validity of ERP components recorded during each task. Internal consistency (i.e., dependability) was reported to describe ERN, CRN, Pe, and Pc scores for the data including all trials. The current study differs from the published report in that the current study aims to examine internal consistency estimates as a function of trial number to determine whether there are task differences in psychometric internal consistency. These specific analyses were not performed on these data prior to the current project. All statistical analyses and hypotheses were preregistered on Open Science Framework (OSF; https://osf.io/p2hr3) prior to analyzing data. Raw EEG data are posted on OpenNeuro (Clayson & Larson, 2023), and summary data and statistical analysis code can be found on OSF (https://osf.io/8cbua/).

2.1. Participants

A sample of 205 participants was collected from two study sites: Brigham Young University (BYU) in Provo, UT, USA (*n*=118) and the University of South Florida (USF) in Tampa, FL, USA (*n*=64). Each site used the same EEG amplifier system and followed identical study protocols. Participants were recruited from undergraduate courses at each university, compensated with course credit, and provided written informed consent prior to study participation. Participants were eligible to participate if they were between the ages of 18 and 65 years. Exclusion criteria included uncorrected visual impairment and self-reported history of head trauma (e.g., loss of consciousness for more than 10 minutes) or neurological disease (e.g., craniocerebral trauma, meningitis, stroke, epilepsy, multiple sclerosis, brain tumor, or migraines more often than four times per month). Consistent with the to-be-replicated study (Meyer et al., 2013), any participants with fewer than six error trials for any tasks after artifact rejection (*n*=23) were excluded from further analyses, resulting in a final sample of 182 participants. A summary of demographic characteristics is presented in Table 1.

2.2. Experimental Tasks

Participants completed an arrowhead flanker task, Stroop task, and a go/no-go task that were identical to the versions used in Meyer et al. (2013). The order of task presentation was counterbalanced (using a fully crossed Latin-square) across participants and sites. Tasks were presented in E-Prime (Psychology Software Tools, Pittsburg, PA). These tasks are described in Clayson, McDonald, et al. (in press), and they are also described in the online supplement.

2.3. Electrophysiological Data Recording and Reduction

EEG data were collected from 128 passive Ag/AgCl scalp sites using a 129-channel hydrocel geodesic sensor net and an amplifier system from Magstim Electrical Geodesics, Inc. (Magstim EGI; Eugene, OR). Data were recorded with a 20K nominal gain and a delta-sigma sinc lowpass filter with a half-power cutoff of 2,000 Hz. The hydrocel net's sensor layout is shown in Clayson et al. (2011). EEG was online referenced to the vertex electrode and the ground sensor was located at PCz. EEG was recorded continuously at a 500 Hz sampling rate with a 24-bit analog-to-digital converter. Impedances were maintained below 50 kΩ for all tasks.

Data processing steps were intended to closely replicate those in Meyer et al. (2013), and a comparison of data recording and processing pipelines is provided in the supplementary material for the parent project on OSF (https://osf.io/qabgh). Continuous EEG was rereferenced offline to an algebraic link of two channels: TP9 and TP10 (i.e., linked mastoids). Continuous EEG was filtered offline using a fourth-order (24 dB/oct) infinite impulse response (IIR) Butterworth filter with half-amplitude cutoffs at .10 and 30 Hz implemented in ERPLab v8.02 (Lopez-Calderon & Luck, 2014). Response-locked epochs were individually extracted for each participant from 400 ms prior to the participant's button press to 800 ms following a button press.

Independent components analysis (ICA) implemented in the ERP PCA Toolkit v2.95 (Dien, 2010) was used to remove eye blinks and horizontal and vertical saccadic eye movement from the segmented waveforms. Epoched EEG data from all channels were processed through a binary version of EEGLab's *runica* function called *binica* (Delorme & Makeig, 2004). Any ICA components that correlated at .9 or above with the scalp topography of a blink template and at .8 or above with the scalp topography of vertical and horizontal saccade templates were excluded. Templates for artifact correction included those automatically generated by the ERP PCA Toolkit and those created by the present authors (Dien, 2010). Movement artifacts were removed using temporal principal component analysis (PCA) with a promax rotation, identifying factors accounting for amplitude differences exceeding 200 µV within a trial (Dien, 2010). Channels were marked as globally bad if they had an absolute correlation with the nearest six neighboring channels that fell below .4. On a trial-wise basis, channels with a voltage difference of 100 µV through the duration of the epoch, channels that were flat, and channels with more than a 30 µV difference with the nearest six neighbors were marked as bad for the epoch. Channels that were marked as bad for more than 20% of epochs were considered globally bad. Bad channels were interpolated using spherical splines (Perrin et al., 1989), but if more than 10% of channels were marked bad for an epoch, the entire epoch was rejected. Epochs associated with a response time (RT) below 100 ms or above 700 ms were excluded (see Meyer et al., 2013).

Response-locked epochs were separately averaged for each participant, task (flanker, go/no-go, Stroop), and event type (correct, error). Epochs were baseline adjusted from 400 ms to 200 ms before the participant’s response. Consistent with Meyer et al. (2013), ERN was quantified using a time-window mean amplitude of the average activity from 0 to 100 ms following participant response at single electrode site: 6 [FCz]. Pe was quantified using a time-window mean amplitude as the average activity from 200 to 400 ms at a single electrode site: 62 [Pz]. ΔERN was calculated by subtracting CRN amplitude from ERN amplitude for flanker and Stroop tasks. For the go/no-go task, ΔERN was calculated by subtracting ERN on no-go trials from CRN on go trials. ΔPe was calculated by subtracting Pc amplitude from Pe amplitude.

Residualized difference scores were calculated by regressing error-trial ERPs on correct-trial ERPs.

2.4. Analytical Plan

Aim 1. We directly replicated Meyer et al. (2013) by examining α of ERN as a function of increasing trial numbers (first 2, 4, 6, 8, 10, 12, 14, 16, 18, and 20 error trials) separately for each task. A plot of α as a function of the number of error trials similar to Figure 4 from Meyer et al. (2013) was created, and interpretation of α across trials and tasks was based on visual inspection of the plot.

Exploratory analyses were performed on data from participants with at least 20 error trials from each task to examine internal consistency using a fully within-person comparison. This analysis rules out between-participant differences in internal consistency when participants are dropped from analyses for having too few trials.

Aim 2. For the remainder of analyses, Bayesian location-scale multilevel models were used to estimate between-person, between-trial, and error variances for each ERP and task. These variance components were then used to estimate psychometric reliability. The models, their implementation, and the formulas for estimating psychometric reliability are described in the supplementary materials.

Aim 3. To examine task differences in internal consistency (i.e., dependability, ICC) of each ERP, post-estimation contrasts from the Bayesian location-scale multilevel models were used to estimate variance components. Eq. (1) and (2) were used to calculate dependability estimates using 20 error trials and the average number of correct trials. These models are the psychometric equations are described in the supplementary material.

2.5. Deviations from Preregistration

First, 15 participants were recorded using a 250 Hz sampling rate instead of the preregistered 500 Hz sampling rate due to experimenter error. These participants were included in the final analyses because the use of time-window mean amplitudes mitigates the impact of different sampling rates and background EEG noise on ERP measurements (Clayson et al., 2013, 2023) Second, population intercepts were removed from the Bayesian location-scale multilevel models, and 0-intercept models were used instead, pushing the intercept to each level of TaskEvent (the unique combinations of task and event). While functionally equivalent to the model with a single population intercept, (co)variance estimates of each level of TaskEvent are directly estimated during model fitting facilitating the estimation of reliability coefficients.

2.6. Support for Hypotheses

Aim 1. Consistent with the preregistered plan, the hypothesis for Aim 1 would be supported if the following three criteria are met. 1) The estimated internal consistency using coefficient alpha (α) of ERN scores for the flanker task is within the 95% confidence interval (CI) of the original estimate (0.54, 0.82)[3] from Meyer et al. (2013) at seven trials, 2) the estimated internal consistency (α) of ERN scores for the go/no-go task is within the 95% CI of the original estimate (0.55, 0.82) at 12 trials, and 3) the 95% CI for α for ERN scores from the Stroop task from the present dataset does not include .7 with up to 20 trials included. Based on this pre-registration, if the hypothesis for Aim 1 is supported, the current study would be considered a successful direct replication of Meyer et al. (2013).

Aim 2. The hypothesis for Aim 2 would be supported if the following three criteria are met. 1) The estimated internal consistency using ϕ of ERN scores using seven trials for the flanker task is within the 95% CI of the original study estimate (0.54, 0.82), 2) the estimated

[3] 95% CIs were calculated based on α of .7 obtained from 43 participants (Meyer et al., 2013).

internal consistency (ϕ) of ERN scores using 12 trials for the go/no-go task is within the 95% CI of the original study estimate (0.55, 0.82), and 3) the 95% credible interval for ϕ for ERN scores from the Stroop task from the present dataset does not include .7 with up to 20 trials included. If the hypothesis for Aim 2 is supported, the current study would be considered a successful conceptual replication of Meyer et al. (2013).

Aim 3. The hypothesis for Aim 3 would be supported if: 1) the 95% credible interval of the dependability post-estimation contrasts for flanker vs. Stroop and go/no-go vs. Stroop do not contain zero, such that ERN and ΔERN scores for the flanker task and go/no-go tasks yield greater dependability than ERN and ΔERN scores for the Stroop task, and 2) the 95% credible interval of the ICC post-estimation contrast for flanker vs. Stroop and go/no-go vs. Stroop does not contain zero, such that ERN scores for the flanker task and go/no-go tasks yield greater ICCs than ERN scores for the Stroop task.

3. Results

Summary information for ERP scores and behavioral data for each task is presented in Table 2. Grand average waveforms for ERN, ΔERN, Pe, and ΔPe are shown in Figures 1 and 2.

3.1. Aim 1

Direct Replication. Figure 3A shows α for ERN scores from flanker, go/no-go, and Stroop tasks as a function of increasing the number of error trials[4]. Similar to Meyer et al. (2013), ERN scores from flanker and go/no-go tasks consistently demonstrated higher internal consistency than ERN scores from the Stroop task across trials. In Meyer et al. (2013), ERN scores from a flanker task demonstrated lower reliability than ERN scores from a go/no-go task

[4] Plots for α for CRN, Pe, and Pc as a function of increasing trials are presented in the supplementary materials on OSF.

when including 2-10 error trials and then higher reliability than ERN scores from a go/no-go task when examining more than 10 error trials. However, the present findings show that ERN scores from a go/no-go task consistently yielded higher reliability than ERN scores from a flanker task across all trial counts (2-20 trials).

Next, the trial cutoffs identified by Meyer et al. (2013) were examined for each task. When examining data from seven error trials, α for flanker ERN scores was 0.45 (95% CI: 0.32, 0.57)[5]. Therefore, the point estimate for α of flanker ERN scores did not fall within the 95% CI of the estimate (0.54, 0.82) from Meyer et al. (2013) and, based on the preregistered criterion, hypothesis #1 was not supported for the flanker task. When examining data from 12 error trials, α for go/no-go ERN scores reached 0.61 (95% CI: 0.51, 0.69) which is within the 95% CI of the original estimate (0.55, 0.82) from Meyer et al. (2013). Thus, hypothesis #1 was supported for the go/no-go task. When examining data from 20 error trials, α for Stroop ERN scores was 0.61 (95% CI: 0.52, 0.70). The 95% CI for the Stroop task includes .7, thus the reliability of ERN scores from the Stroop task at 20 trials was higher than expected and hypothesis #1 was not supported for the Stroop task.

Extension. Next, α was examined as a function of increasing error trials for only participants with at least 20 error trials for each task (*n* = 124) to rule out between-participant differences in internal consistency when participants are dropped from analyses for having too few trials (see Figure 3b). When compared to analysis of the full sample (see Figure 3a), ERN scores from both flanker and go/no-go tasks consistently demonstrated numerically lower reliability across 2-20 trials, while ERN scores from the Stroop task demonstrated numerically

[5] The number of participants available to calculate α for the flanker task at 7 trials was 181. The number of participants available to calculate α for the go/no-go task at 12 trials was 172. The number of participants available to calculate α for the Stroop task at 20 trials was 151.

higher reliability at low trial counts. The 95% CIs for α for ERN scores from the flanker and go/no-go tasks, but not the Stroop task, included .7 with 20 error trials included in estimation.

Interim Summary. When examining the reliability of ERN scores using α, ERN from a go/no-go task demonstrated the numerically highest reliability across 2-20 trials. Go/no-go ERN scores exhibited α levels similar to those found in Meyer et al. (2013), while ERN scores from a flanker task demonstrated lower reliability than suggested by Meyer et al. (2013), and ERN scores from a Stroop task demonstrated higher reliability than expected. Thus, hypothesis #1 was partially supported.

3.2. Aim 2

Conceptual Replication. When examining data from seven error trials, ϕ of ERN scores from the flanker task was 0.41 (95% credible interval [CrI]: 0.35, 0.47); see Figure 4[6]), which is not within the 95% CI of the original estimate (0.54, 0.82) from Meyer at al. (2013). Thus, hypothesis #2 was not supported for the flanker task. At 12 trials, ϕ for go/no-go ERN scores reached 0.56 (95% CrI: 0.48, 0.63) which is within the 95% CI of the original estimate (0.55, 0.82) from Meyer at al. (2013), supporting hypothesis #2 for the go/no-go task. At 20 trials, ϕ for Stroop ERN scores was 0.62 (95% CrI: 0.55, 0.68). The 95% CI for the Stroop task excludes .7, supporting hypothesis #2 for the Stroop task.

Extension. Table 2 shows the number of trials required for ERN, CRN, Pe, Pc, ΔERN, and ΔPe scores to reach acceptable levels of ϕ (.7, .8, .9) for each task. The go/no-go task required the fewest trials to reach each reliability threshold for ERN scores, followed by the flanker then the Stroop task. Notably, all three tasks required at least 20 error trials for ERN

[6] Plots for dependability for CRN, Pe, and Pc as a function of increasing trial number are presented in the supplementary materials on OSF.

scores to reach ϕ of.7. The go/no-go task required the fewest number of trials to reach each reliability threshold for Pe scores. For all tasks, fewer trials were needed for Pe and ∆Pe scores to reach each reliability threshold compared to ERN and ∆ERN scores.

Classical test theory estimates of internal consistency of residualized difference scores are shown in Figure 5. Similar to estimates of coefficient alphas above, participants with too few trials retained for estimating reliability were dropped from reliability analysis as the number of trials examined increased. ∆ERN scores for flanker, go/no-go, and Stroop reached reliability of .7 at 35, 25, and 31 trials, respectively. These estimates are based on 132 participants who had at least 35 flanker error trials, 106 participants who had at least 25 go/no-go error trials, and 111 participants who had at least 31 Stroop trials. ∆ERN scores reached reliability of .8 for go/no-go at 35 trials, however, this estimate is based on only 42 participants who had at least 35 go/no-go error trials. ∆ERN scores from flanker and Stroop failed to reach reliability of .8 with 40 error trials included. The decision not to further increase the number of error trials available for estimation was due to the loss in sample size at higher numbers of trials.

Interim Summary. When examining the reliability of ERN scores using ϕ, the go/no-go and Stroop tasks exhibited numerically similar reliability estimates to those expected based on findings from Meyer et al. (2013). ERN scores from a flanker task demonstrated lower ϕ than expected. Thus, hypothesis #2 was partially supported. ERN scores from a go/no-go task consistently exhibited the numerically highest ϕ levels across 2-20 error trials and required the fewest number of error trials to reach ϕ thresholds of .7, .8, and .9.

3.3. Aim 3

Task Comparisons. Plots of point estimates of ϕ when including 20 error trials and ICC (mean of the posterior distributions) and associated 95% interval estimates for each ERP and task are presented in Figure 6. Results from post-estimation contrasts from the multilevel models for ERN, CRN, Pe, Pc, ΔERN, and ΔPe are presented in Table 3. While ϕ of ERN and ΔERN scores were expected to be higher for the flanker and go/no-go tasks than for the Stroop task, no task differences were observed. Thus, hypothesis #3 was not supported.

Exploratory analyses of task differences in ϕ for CRN, Pe, Pc, and ΔPe scores were conducted (see Figure 6 and Table 3). Pe scores from the go/no-go task yielded higher ϕ than scores from the flanker and Stroop tasks, whereas ΔPe scores from the go/no-go task yielded higher ϕ than scores from the Stroop task but not flanker task.

ICC estimates provided a trial-independent characterization of reliability across all ERPs and tasks (Clayson, Brush, et al., 2021). Although ICC for ERN was expected to be higher for the flanker and go/no-go tasks than the Stroop task, task differences were not observed. Thus, hypothesis #3 was not supported. Exploratory analyses of task differences for ICC of ΔERN, CRN, Pe, Pc, and ΔPe scores were conducted. Pe scores from the go/no-go task yielded higher ICC than Pe scores from the Stroop and flanker tasks, and ΔPe scores from the go/no-go task yielded higher ICC than ΔPe scores from the Stroop task but not flanker task. Task differences in ICC were not observed for ΔERN, Pc, and CRN scores.

3.4. Exploratory Analyses

Based on the above analyses, the number of trials needed to achieve adequate internal consistency of ERP components varied across tasks. However, ERP studies typically average all trials from a participant prior to scoring, and the above approaches consequently do not characterize the typical ERP component scores used in analysis. Neither the original studies (Meyer et al., 2013; Riesel et al., 2013) nor a recent replication study (Clayson et al., 2023) statistically compared internal consistency estimates across tasks. Therefore, exploratory analyses evaluated internal consistency estimates of each ERP component score using the average number of error trials (and average number of correct trials for difference scores) from each task (see Table 2)[7]. Plots of point estimates of ϕ when including the average number of error trials and associated 95% interval estimates for each ERP and task are presented in Figure 7. Results from post-estimation contrasts from the multilevel models are presented in Table 3.

For ERN, when including the average number of error trials, scores from the flanker task showed higher ϕ than those from the go/no-go and Stroop tasks, while ϕ of ERN scores from the Stroop and go/no-go tasks did not differ. For ΔERN, scores from the flanker task showed higher ϕ than those from the go/no-go task, but not from the Stroop task. ϕ of ΔERN scores from the Stroop and go/no-go tasks did not differ. For ΔPe, scores from the flanker task showed higher ϕ than those from the go/no-go and Stroop tasks, while dependability of ΔPe scores from the go/no-go and Stroop tasks did not differ. No task differences emerged for Pe scores.

Interim Summary. ϕ and ICC of ERN and ΔERN did not differ across tasks when examining 20 error trials. Thus, hypothesis #3 was not supported. In contrast, Pe scores from a go/no-go task demonstrated statistically higher ϕ and ICC than Pe scores from flanker and Stroop tasks. ΔPe scores from a go/no-go task demonstrated higher ϕ than ΔPe scores from a Stroop task but not flanker task.

When examining task differences using the average number of recorded trials for each task, ERN scores from a flanker task showed higher ϕ than scores from go/no-go and Stroop

[7] A chi-square test showed a significant difference in the average number of error trials among the three tasks (χ^2 (2) = 10.36, *p* = .01). Results from pairwise chi-square tests showed the average number of error trials from the flanker task (*n*=60) significantly differed from the average number of error trials from the go/no-go task (*n*=30) (χ^2 (1) = 10, *p* = .002). The average number of error trials from the flanker task did not differ from the Stroop task (*n*=42) (χ^2 (1) = 3.18, *p* = .07). The average number of error trials from the go/no-go task did not differ from the Stroop task (χ^2 (1) = 2, *p* = .16).

tasks. ∆ERN scores from a flanker task showed higher ϕ than scores from a go/no-go task but not Stroop task. For ∆Pe scores, a flanker task showed higher dependability than a Stroop and go/no-go task. No task differences were present for Pe scores.

4. Discussion

The current study examined the internal consistency of ERP measures of performance monitoring during flanker, go/no-go, and Stroop tasks. Direct and conceptual replications of Meyer et al. (2013) assessed internal consistency of ERN scores as a function of increasing error trials and task used for recording ERN. Findings for the internal consistency of ERN scores from a go/no-go task were successfully replicated. Findings for a Stroop task were partially replicated; Internal consistency of ERN scores were as-expected for the conceptual replication (i.e., dependability) and higher than expected for the direct replication (i.e., coefficient alpha). Findings for a flanker task were not replicated as ERN scores demonstrated lower internal consistency than expected. The failure to replicate the internal consistency of ERN scores from Meyer et al. (2013) suggests that the widespread adoption of flanker tasks based on assumptions of superior internal consistency at low-trial counts may be inappropriate and highlights the need for psychometric reliability to be examined and reported on a study-by-study basis (see Clayson & Miller, 2017a, 2017b).

4.1. Direct Replication

Following the procedure of the replicated study, a visual examination of coefficient alpha across trials indicated that ERN scores from a go/no-go task consistently demonstrated the highest internal consistency across each trial count from 2 to 20 trials. This finding differs from Meyer et al. (2013), wherein ERN scores from a flanker task showed the highest internal consistency when including at least 12 error trials. However, relying solely on visual inspection of internal consistency estimates can lead to erroneous conclusions, because uncertainty around the observed differences is not readily interpretable. Therefore, post-estimation contrasts from Bayesian multilevel models were used to statistically compare the internal consistency of ERN scores across tasks and interpret task differences.

4.2. Conceptual Replication

Analyses for Aim 2 indicated a greater number of trials were required for ERN scores to reach an internal consistency threshold of .7 than those reported by Meyer et al. (2013). Meyer et al. (2013) found that 7 and 12 trials were sufficient for a flanker and go/no-go task with coefficient alpha, respectively, while 24 and 22 trials were required to reach an internal consistency threshold of .7 (ϕ) in the current study. Regarding the Stroop task, like Meyer et al. (2013), more than 20 trials were required for ERN scores to reach a reliability threshold of .7. The discrepancy with Meyer et al. (2013) for flanker and go/no-go tasks illustrates a common issue in ERN research, where participants with too few error trials based on published psychometric information are excluded. If Meyer et al. (2013)'s recommendations had been used in the current study, it would have resulted in the inclusion of participants with unreliable data, thereby inflating Type I and Type II errors (Hedge et al., 2018; Loken & Gelman, 2017; Rouder & Haaf, 2019). These results emphasize the critical need to evaluate and report internal consistency on a study-by-study basis and caution against using universal trial cutoffs for data inclusion. Instead, subject-level data quality and internal consistency estimates are recommended to identify individual participants that show ERN with poor internal consistency. This approach circumvents limitations of relying on group-level reliability estimates and group-level numbers of trials, which can obscure poor data quality of individual participants (see Clayson, Brush, et al., 2021). The analysis of subject-level reliability is readily available in the open-source ERP Reliability Analysis (ERA) Toolbox (Clayson & Miller, 2017a).

Analyses for Aim 3 showed that, contrary to expectations, internal consistency of ERN and ΔERN scores did not differ across tasks when including 20 error trials. Importantly, all three tasks required more than 20 trials to obtain dependability of .7. A well-regarded psychometric text recommends a minimum threshold of .7 for novel or preliminary work, .8 for developed research areas focused on group differences, and .9 or .95 for clinical decision-making (Nunally & Bernstein, 1994). Similar thresholds are advised for ERP research (Clayson & Miller, 2017b). Given that studies of ERN are not typically considered novel or preliminary in most contexts, higher thresholds of internal consistency should be used to support the examination of individual differences. As such, examining task differences in ERN score internal consistency when including more error trials may be more informative.

Exploratory analyses were used to examine task differences in the dependability of ERN scores using the average number of error trials for each task. Results showed that ERN scores from a flanker task demonstrated higher dependability than scores from go/no-go and Stroop tasks, and ΔERN scores from a flanker task demonstrated higher dependability than scores from a go/no-go task but not a Stroop task. Given that each task was the same length (420 trials) and that common practice in ERN research is to use all available trials for analyses, a flanker task may be best suited for studying individual differences in ERN based on higher estimates of internal consistency due to the inclusion of more error trials than the other two tasks.

Unlike ERN, task differences in internal consistency emerged for Pe scores at 20 error trials. Pe scores from a go/no-go task exhibited higher dependability than those from flanker and Stroop tasks, and ΔPe scores from a go/no-go task exhibited higher dependability than those from a Stroop task but not a flanker task. No task differences emerged for Pe using mean error trials. These findings suggest any of the three tasks may be appropriate for measuring Pe; however, a go/no-go task may be preferable when few error trials are expected.

4.3. Moving Forward

Overreliance on a single paradigm for recording ERN can limit understanding of how error-related performance monitoring relates to individual differences. ERN represents a manifestation of multiple neural generators, involves various neurotransmitters, and is influenced by several cognitive, affective, motivational, and motor processes (Gehring et al., 2012). Consequently, task-related differences in ERN amplitude could reflect engagement of different component processes contributing to ERN. Indeed, ERN and Pe indices measured during go/no-go, flanker, and Stroop tasks demonstrate low convergence even after adjusting for measurement error (Clayson et al., 2023; Riesel et al., 2013). This conclusion is supported by the strong convergence of ERN when examining three different versions of flanker tasks (Clayson et al., 2024), and the stronger concordance of ERN across versions of flanker tasks than across different types of tasks.

Ideally, task choice should be guided by interest in the component processes contributing to ERN during that task. However, the functional significance of ERN and Pe remains not well understood (Clayson, Kappenman, et al., 2021; Overbeek et al., 2005), especially in specific contexts (e.g., task), complicating task selection and discouraging the adoption of a single paradigm. A notable distinction among flanker, go/no-go, and Stroop tasks is the way errors are committed. Errors across all three tasks reflect a failure to override a prepotent response; However, in a go/no-go task, ERN and Pe are measured from commission errors on no-go trials, in which participants fail to inhibit a motor response. In contrast, flanker and Stroop tasks require a motor response on every trial, with errors reflecting the selection of an incorrect option between two conflicting responses. Therefore, if response inhibition is of interest, it would be sensible to choose a go/no-go task for measuring ERN and Pe.

To understand the functional significance of ERN and Pe recorded during different tasks and to inform task decisions for specific contexts, future research might consider “multiverse” analyses, where multiple task features or data-processing pipelines could be analyzed within the same study (see Clayson, 2024a). For example, using both a go/no-go and flanker task could highlight whether different forms of error commission differentially relate to clinically meaningful variables, such as behavioral inhibition. Even within the same task, different data-processing decisions (e.g., ocular artifact correction procedures, approaches to scoring), could be examined to identify approaches optimized for examining individual differences. However, a multiverse approach must carefully balance the research aims against potential drawbacks, such as participant burden and fatigue. Taken together, narrowing performance-monitoring research to a single task would likely limit a comprehensive understanding of changes in error processing related to contextual influences, individual differences, or psychopathology.

The current study’s findings can be used as a guide during planning stages of ERN studies, such as determining the required task length to retain an adequate number of error trials for the study of individual differences. A task with 420 trials resulted in an adequate number of error trials for ERN scores from a flanker task to reach an overall dependability of .86, whereas dependability of ERN scores did not reach .8 for the go/no-go and Stroop tasks due to higher participant accuracy on these tasks. Therefore, the current task length may be sufficient for flanker tasks but the inclusion of more trials may be necessary for go/no-go and Stroop tasks.

The present study closely followed the procedures described in Meyer et al. (2013) and incorporated input from an original co-author. However, potential deviations from the original study could impact the present findings. First, tasks were adapted to use with E-Prime Software because Presentation was incompatible with the Magstim EGI hardware setup. These E-Prime tasks, modeled on the original paradigms, were functionally identical and are available on OSF for future replications (https://osf.io/tzq8n/). Second, the original study used active electrodes, while the current study used passive electrodes. Although passive and active electrodes can provide comparable data quality (Mathewson et al., 2017), differences in data quality between the two studies could influence internal consistency estimates. However, this possible explanation is unlikely because the present study replicated the internal consistency estimates of ERN for the go/no-go task. If data quality were worse in the present study, worse internal consistency across all three tasks would be expected. Third, minor adjustments were made to the data-processing pipeline due to the different software used for recording and analyzing EEG. For example, the original study used a regression-based approach for ocular artifact correction, but the present study used ICA based on findings it leads to improve data quality (Clayson, Baldwin, Rocha, et al., 2021). Fourth, the sample size in the present study (*n* = 182) was over four times larger than the original study (*n* = 43), and sampling error could contribute to differences in reliability estimates. Additionally, the demographic characteristics of the samples differed: the present sample was predominantly White (85%), with Asians being the next largest group (8%), while the original study's sample was predominantly Asian-American (46%) with Caucasians/Europeans as the next largest group (39%). Despite both studies sampling college undergraduates using identical inclusion/exclusion criteria, these demographic differences are notable. Hopefully, recent efforts to mobilize the field to fully characterize the demographics of samples in psychophysiology studies will provide a foundation for understanding how these characteristics can influence study interpretations (Gatzke-Kopp et al., 2023; Kissel & Friedman, 2023).

Given these methodological and demographic differences, the extent to which the present study constitutes a direct (i.e., close) replication of Meyer et al. (2013) is debatable. While we closely followed their procedures and collaborated with an original co-author, variations between studies could influence outcomes. This raises questions about what defines a direct replication in studies of ERPs (Clayson et al., 2019). Although variations are sometimes unavoidable (e.g., EEG hardware differences between labs, different sample characteristics), a key aspect of a direct replication is the replication of the core procedures and conditions critical for eliciting the phenomenon under investigation (Simons, 2014), but ERP studies involve many researcher degrees of freedom (e.g., Clayson et al., 2022). We acknowledge that some may interpret our study as a conceptual rather than direct replication due to methodological and demographic differences. Although there are various definitions of replication (Vachon et al., 2021), a recent relevant definition describes it as a study designed such that its outcome serves as diagnostic evidence of the original study’s claim (Nosek & Errington, 2020). The present findings yield outcomes that might decrease confidence in the claims of the original research, despite study differences. Replications by independent labs—even when minor differences exist—are vital for testing the repeatability of a study’s findings.

The present study used the same tasks as Meyer et al. (2013), but these represent only a subset of possible flanker, Stroop, and go/no-go tasks. Present findings might not generalize to other versions of these tasks due to important methodological differences, particularly those related to stimulus-response ensembles described by the dimensional overlap model (Kornblum et al., 1990). For example, the traditional semantic Stroop task involves interference between relevant (i.e., font color) and irrelevant (i.e., word) stimulus dimensions, both of which overlap with response options: words “red”, “green”, or “blue” appear in either red, green or blue font, and each font color (and word) is associated with a unique response (Stroop, 1935). In contrast, the current version of the Stroop task eliminated the one-to-one mapping between the irrelevant dimension (i.e., word) and participant response by only presenting words (red, green, blue) in red or green font. As a result, the relevant and irrelevant stimulus dimensions did not overlap as they do in the traditional Stroop. Similarly, the modified version of the flanker task used in the present study has overlap between stimulus and response dimensions, unlike the original letter version of the task (Eriksen & Eriksen, 1974). It is possible that there is an assumption that tasks with the same name measure the same construct, simply because they share the same name—fallacious reasoning is referred to as the jingle fallacy (Kelley, 1927). Taken together, different stimulus-response dimensions and causes of conflict in each task might contribute to differences in internal consistency, and systematic examinations of conflict type would help clarify how they contribute to changes in ERN and Pe.

The study had limitations. The current study is limited by use of a single-data processing pipeline and a single version of each task (Clayson, 2024a), and a single data-processing pipeline was chosen to be consistent with Meyer et al. (2013). Internal consistency of ERP scores can vary as a function of data-processing decisions (Clayson, Baldwin, Rocha, et al., 2021) and task features, even for the same paradigm (Clayson et al., 2024). Thus, results may differ when using other processing decisions or task variants (e.g., different versions of a Stroop task). Although the present study focused on internal consistency, other approaches could be used to identify paradigms optimal for recording ERN in certain contexts. For example, data from a paradigm might relate to certain individual differences more robustly than another paradigm. Relationships with external correlates can help guide which task to use for recording ERN, but this should not limit investigations into the particular features that lead to changes in ERN. Similary, ERN scores might show higher test-retest reliability when recorded using certain paradigms, and future research might consider examining which pardigms yield both the highest stability over time and within sessions.

Taken together, these findings highlight the importance of task selection in performance-monitoring ERP research as the choice of task can significantly impact the internal consistency of these indices, which consequently impacts the ability to observe relationships with external correlates. The failure to replicate the internal consistency of ERN scores from Meyer et al. (2013) for flanker and Stroop tasks, despite the use of identical tasks in a similarly aged sample, underscores the need for internal consistency to be examined and reported in any study of individual differences in ERN amplitude due to the numerous contextual factors that could contribute to internal consistency (Clayson & Miller, 2017b). Findings of task differences for Pe and ERN score internal consistency emphasize the inappropriateness of generalizing a trial cutoff for data inclusion from a single task to other tasks (see also Clayson, 2020). Once all data from each task are included, a flanker task appears best suited for recording ERN, ∆ERN, and ∆Pe, as the data from the flanker task show the highest internal consistency due to inclusion of more error trials. Despite differences in the number of trials and observed internal consistency to Meyer et al. (2013), all tasks in the present study achieved reasonable internal consistency (see Table 2) when all data were included in analysis and are therefore suitable for investigating associations with individual differences. Consequently, the choice of tasks should primarily be guided by hypotheses and the specific research question. Present findings should not be construed as unequivocal support for any single task, given the ongoing uncertainty surrounding the functional significance of ERN and Pe and the causes of variability in internal consistency across flanker, Stroop, and go/no-go tasks.

Data Availability

The present study used publicly available data from OpenNeuro (ds004883). More information can be found on the original projects OSF page: https://osf.io/8cbua/.

Conflict of Interest

Authors have no known conflicts of interest related to this work.

References

Baldwin, S. A., Larson, M. J., & Clayson, P. E. (2015). The dependability of electrophysiological measurements of performance monitoring in a clinical sample: A generalizability and decision analysis of the ERN and Pe. *Psychophysiology*, *52*(6), 790–800. https://doi.org/10.1111/psyp.12401

Banica, I., Sandre, A., Shields, G. S., Slavich, G. M., & Weinberg, A. (2020). The error-related negativity (ERN) moderates the association between interpersonal stress and anxiety symptoms six months later. *International Journal of Psychophysiology*, *153*, 27–36. https://doi.org/10.1016/j.ijpsycho.2020.03.006

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295x.108.3.624

Carrasco, M., Harbin, S. M., Nienhuis, J. K., Fitzgerald, K. D., Gehring, W. J., & Hanna, G. L. (2013). Increased Error-Related Brain Activity in Youth with Obsessive-Compulsive Disorder and Unaffected Siblings. *Depression and Anxiety*, *30*(1), 39–46. https://doi.org/10.1002/da.22035

Clayson, P. E. (2020). Moderators of the internal consistency of error-related negativity scores: A meta-analysis of internal consistency estimates. *Psychophysiology*, *57*(8), e13583. https://doi.org/10.1111/psyp.13583

Clayson, P. E. (2024a). Beyond single paradigms, pipelines, and outcomes: Embracing multiverse analyses in psychophysiology. *International Journal of Psychophysiology*, *197*, 112311. https://doi.org/10.1016/j.ijpsycho.2024.112311

Clayson, P. E. (2024b). The psychometric upgrade psychophysiology needs. *Psychophysiology*, *61*(3), e14522. https://doi.org/10.1111/psyp.14522

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology*, *50*(2), 174–186. https://doi.org/10.1111/psyp.12001

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2021). Evaluating the internal consistency of subtraction-based and residualized difference scores: Considerations for psychometric reliability analyses of event-related potentials. *Psychophysiology*, *58*(4), e13762. https://doi.org/10.1111/psyp.13762

Clayson, P. E., Baldwin, S. A., Rocha, H. A., & Larson, M. J. (2021). The data-processing multiverse of event-related potentials (ERPs): A roadmap for the optimization and standardization of ERP processing and reduction pipelines. *NeuroImage*, *245*, 118712. https://doi.org/10.1016/j.neuroimage.2021.118712

Clayson, P. E., Brush, C. J., & Hajcak, G. (2021). Data quality and reliability metrics for event-related potentials (ERPs): The utility of subject-level reliability. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *165*, 121–136. https://doi.org/10.1016/j.ijpsycho.2021.04.004

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*(11), e13437. https://doi.org/10.1111/psyp.13437

Clayson, P. E., Carbine, K. A., Baldwin, S. A., Olsen, J. A., & Larson, M. J. (2021). Using generalizability theory and the ERP Reliability Analysis (ERA) Toolbox for assessing test-retest reliability of ERP scores part 1: Algorithms, framework, and implementation. *International Journal of Psychophysiology*, *166*, 174–187. https://doi.org/10.1016/j.ijpsycho.2021.01.006

Clayson, P. E., Clawson, A., & Larson, M. J. (2011). Sex differences in electrophysiological indices of conflict monitoring. *Biological Psychology*, *87*(2), 282–289. https://doi.org/10.1016/j.biopsycho.2011.03.011

Clayson, P. E., Kappenman, E. S., Gehring, W. J., Miller, G. A., & Larson, M. J. (2021). A commentary on establishing norms for error-related brain activity during the arrow flanker task among young adults. *NeuroImage*, *234*, 117932. https://doi.org/10.1016/j.neuroimage.2021.117932

Clayson, P. E., Keil, A., & Larson, M. J. (2022). Open science in human electrophysiology. *International Journal of Psychophysiology*, *174*, 43–46. https://doi.org/10.1016/j.ijpsycho.2022.02.002

Clayson, P. E., & Larson, M. J. (2023). *Registerd Replication Report of ERN/Pe Psychometrics* [Dataset]. Openneuro. https://doi.org/10.18112/OPENNEURO.DS004602.V1.0.1

Clayson, P. E., Mcdonald, J. B., Park, B., Holbrook, A., Baldwin, S. A., Riesel, A., & Larson, M. J. (2023). Registered replication report of the construct validity of the error-related negativity (ERN): A multi-site study of task-specific ERN correlations with internalizing and externalizing symptoms. *Psychophysiology*, e14496. https://doi.org/10.1111/psyp.14496

Clayson, P. E., & Miller, G. A. (2017a). ERP Reliability Analysis (ERA) Toolbox: An open-source toolbox for analyzing the reliability of event-related brain potentials. *International*

Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology, *111*, 68–79. https://doi.org/10.1016/j.ijpsycho.2016.10.012

Clayson, P. E., & Miller, G. A. (2017b). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. *International Journal of Psychophysiology*, *111*, 57–67. https://doi.org/10.1016/j.ijpsycho.2016.09.005

Clayson, P. E., Rocha, H. A., Baldwin, S. A., Rast, P., & Larson, M. J. (2022). Understanding the Error in Psychopathology: Notable Intraindividual Differences in Neural Variability of Performance Monitoring. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *7*(6), 555–565. https://doi.org/10.1016/j.bpsc.2021.10.016

Clayson, P. E., Rocha, H. A., McDonald, J. B., Baldwin, S. A., & Larson, M. J. (2024). A registered report of a two-site study of variations of the flanker task: ERN experimental effects and data quality. *Psychophysiology*, e14607. https://doi.org/10.1111/psyp.14607

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dien, J. (2010). The ERP PCA Toolkit: An open source program for advanced statistical analysis of event-related potential data. *Journal of Neuroscience Methods*, *187*(1), 138–145. https://doi.org/10.1016/j.jneumeth.2009.12.009

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: A tutorial. *Biological Psychology*, *51*(2), 87–107. https://doi.org/10.1016/S0301-0511(99)00031-9

Gatzke-Kopp, L., Keil, A., & Fabiani, M. (2023). Diversity and representation. *Psychophysiology*, *60*(11), e14431. https://doi.org/10.1111/psyp.14431

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2012). The error-related negativity (ERN/Ne). In *The Oxford handbook of event-related potential components* (pp. 231–291). Oxford University Press.

Hajcak, G., Meyer, A., & Kotov, R. (2017). Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects. *Journal of Abnormal Psychology*, *126*(6), 823–834. https://doi.org/10.1037/abn0000274

Hanna, G. L., Liu, Y., Rentschler, L. G., Hanna, B. S., Arnold, P. D., & Gehring, W. J. (2024). Altered Error Monitoring and Decreased Flanker Task Accuracy in Pediatric Obsessive–Compulsive Disorder. *Child Psychiatry & Human Development*. https://doi.org/10.1007/s10578-024-01711-4

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

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*(4), 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Kelley, T. L. (1927). *Interpretation of educational measurements* (p. 353). World Book Co.

Kissel, H. A., & Friedman, B. H. (2023). *Participant diversity in Psychophysiology*. *60*(11), e14369. https://doi.org/10.1111/psyp.14369

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility--A model and taxonomy. *Psychological Review*, *97*(2), 253–270. https://doi.org/10.1037/0033-295X.97.2.253

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). *The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? 9*(2). https://doi.org/10.1177/1094428105284919

Larson, M. J., Clayson, P. E., & Baldwin, S. A. (2012). Performance monitoring following conflict: Internal adjustments in cognitive control? *Neuropsychologia*, *50*(3), 426–433. https://doi.org/10.1016/j.neuropsychologia.2011.12.021

Larson, M. J., Clayson, P. E., & Clawson, A. (2014). Making sense of all the conflict: A theoretical review and critique of conflict-related ERPs. *International Journal of Psychophysiology*, *93*(3), 283–297. https://doi.org/10.1016/j.ijpsycho.2014.06.007

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fnhum.2014.00213

Lutz, M. C., Kok, R., & Franken, I. H. A. (2021). Event-related potential (ERP) measures of error processing as biomarkers of externalizing disorders: A narrative review. *International Journal of Psychophysiology*, *166*, 151–159. https://doi.org/10.1016/j.ijpsycho.2021.06.002

Macedo, I., Pasion, R., Barbosa, F., & Ferreira-Santos, F. (2021). A dimensional approach to the neuronal correlates of anxiety, depression, and perfectionism: A transdiagnostic dissociation of error-related brain activity. *Behavioural Brain Research*, *408*, 113271. https://doi.org/10.1016/j.bbr.2021.113271

Martin, E. A., McCleery, A., Moore, M. M., Wynn, J. K., Green, M. F., & Horan, W. P. (2018). ERP indices of performance monitoring and feedback processing in psychosis: A meta-analysis. *International Journal of Psychophysiology*, *132*, 365–378. https://doi.org/10.1016/j.ijpsycho.2018.08.004

Mathewson, K. E., Harrison, T. J. L., & Kizuk, S. A. D. (2017). High and dry? Comparing active dry EEG electrodes to active and passive wet electrodes. *Psychophysiology*, *54*(1), 74–82. https://doi.org/10.1111/psyp.12536

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*(2), 266–274. https://doi.org/10.1037/abn0000044

Meyer, A., Nelson, B., Perlman, G., Klein, D. N., & Kotov, R. (2018). A neural biomarker, the error-related negativity, predicts the first onset of generalized anxiety disorder in a large sample of adolescent females. *Journal of Child Psychology and Psychiatry*, *59*(11), 1162–1170. https://doi.org/10.1111/jcpp.12922

Meyer, A., Riesel, A., & Hajcak Proudfit, G. (2013). Reliability of the ERN across multiple tasks as a function of increasing errors. *Psychophysiology*, *50*(12), 1220–1225. https://doi.org/10.1111/psyp.12132

Michael, J. A., Wang, M., Kaur, M., Fitzgerald, P. B., Fitzgibbon, B. M., & Hoy, K. E. (2021). EEG correlates of attentional control in anxiety disorders: A systematic review of error-related negativity and correct-response negativity findings. *Journal of Affective Disorders*, *291*, 140–153. https://doi.org/10.1016/j.jad.2021.04.049

Morand-Beaulieu, S., Banica, I., Freeman, C., Ethridge, P., Sandre, A., & Weinberg, A. (2024). Neural response to errors among mothers with a history of recurrent depression and their adolescent daughters. *Development and Psychopathology*, 1–15. https://doi.org/10.1017/S0954579424001780

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nunally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill.

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

Orr, J. M., & Carrasco, M. (2011). The Role of the Error Positivity in the Conscious Perception of Errors. *Journal of Neuroscience, 31*(16), 5891–5892. https://doi.org/10.1523/JNEUROSCI.0279-11.2011

Overbeek, T. J. M., Nieuwenhuis, S., & Ridderinkhof, K. R. (2005). Dissociable components of error processing: On the functional significance of the Pe vis-à-vis the ERN/Ne. *Journal of Psychophysiology*, *19*(4), 319–329. https://doi.org/10.1027/0269-8803.19.4.319

Pasion, R., & Barbosa, F. (2019). ERN as a transdiagnostic marker of the internalizing-externalizing spectrum: A dissociable meta-analytic effect. *Neuroscience & Biobehavioral Reviews*, *103*, 133–149. https://doi.org/10.1016/j.neubiorev.2019.06.013

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, *72*(2), 184–187. https://doi.org/10.1016/0013-4694(89)90180-6

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. https://doi.org/10.1111/j.1469-8986.2010.00974.x

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

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*(3), 377–385. https://doi.org/10.1016/j.biopsycho.2013.04.007

Rietdijk, W. J. R., Franken, I. H. A., & Thurik, A. R. (2014). Internal consistency of event-related potentials associated with cognitive control: N2/P3 and ERN/Pe. *PloS One*, *9*(7), e102672. https://doi.org/10.1371/journal.pone.0102672

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Vachon, B., Curran, J. A., Karunananthan, S., Brehaut, J., Graham, I. D., Moher, D., Sales, A. E., Straus, S. E., Fiander, M., Paprica, P. A., & Grimshaw, J. M. (2021). Replication Research Series-Paper 1: A concept analysis and meta-narrative review established a comprehensive theoretical definition of replication research to improve its use. *Journal of Clinical Epidemiology*, *129*, 176–187. https://doi.org/10.1016/j.jclinepi.2020.07.006

Vallet, W., Neige, C., Mouchet-Mages, S., Brunelin, J., & Grondin, S. (2021). Response-locked component of error monitoring in psychopathy: A systematic review and meta-analysis of error-related negativity/positivity. *Neuroscience & Biobehavioral Reviews*, *123*, 104–119. https://doi.org/10.1016/j.neubiorev.2021.01.004

Table 1

Summary Demographic Information for Participants

	M	SD
Age	20	2
	n	%
Sex		
Female	117	64.3
Male	65	35.7
Gender		
Women	115	63.2
Men	65	35.7
Bigender	1	0.55
Nonbinary	1	0.55
Race		
White	155	85.2
Asian	14	7.7
American Indian/Alaskan	2	1.1
Black/African American	2	1.1
Multiracial	9	5
Hispanic/Latino/Spanish Origin	44	24
Sexual Orientation		
Heterosexual	164	90.1
Bisexual/Pansexual	12	6.6
Gay/Lesbian	1	0.55
Queer	1	0.55
Asexual	1	0.55
Unsure	3	1.6
Income		
Below $15,000	17	9.3
$15,000 to $30,000	12	6.6
$30,001 to $45,000	12	6.6
$45,001 to $60,000	18	9.9
$60,001 to $75,000	21	11.5
Over $75,000	101	55.5
Unknown	1	0.55

Note: Participants provided their parent/guardian income if they were a full-time student or dependent.

Table 2

Summary Information for Event-Related Potential Number of Trials, Amplitude (µV), Accuracy, Dependability, and Number of Trials Required to Reach Dependability Thresholds

		Dependability Thresholds			Overall
	M (SD)	0.7	0.8	0.9	Dependability
Flanker					
Accuracy	79% (0.15)				
Error trials	60 (41)				
Correct trials	319 (76)				
ERN	-.5 (3.0)	24	40	87	.86 (.82, .89)
CRN	2.5 (2.7)	23	39	86	.97 (.96, .98)
Pe	-0.4 (1.8)	17	30	65	.89 (.86, .91)
Pc	3.6 (3.5)	45	76	167	.94 (.93, .95)
ΔERN	-3.0 (2.7)	31	57	159	.80 (.75, .85)
ΔPe	4.0 (3.8)	14	25	60	.90 (.87, .92)
Go/no-go					
Accuracy	88% (0.09)				
Error trials	30 (24)				
Correct trials	283 (53)				
ERN	-1.5 (3.5)	22	37	81	.76 (.70, .81)
CRN	2.0 (2.7)	23	39	84	.97 (.96, .97)
Pe	6.0 (4.3)	11	19	41	.86 (.83, .89)
Pc	0.3 (1.6)	60	102	223	.92 (.90, .93)
ΔERN	-3.5 (3.2)	30	54	150	.70 (.62, .76)
ΔPe	5.6 (4.1)	12	21	50	.84 (.80, .88)
Stroop					
Accuracy	83% (0.13)				
Error trials	42 (42)				
Correct trials	323 (70)				
ERN	-1.1 (2.8)	28	49	106	.77 (.72, .82)
CRN	2.4 (2.9)	20	34	77	.97 (.97, .98)
Pe	3.2 (3.6)	17	29	64	.85 (.81, .88)
Pc	-0.1 (1.8)	43	73	158	.95 (.93, .96)
ΔERN	-3.5 (2.9)	30	54	148	.76 (.70, .81)
ΔPe	3.3 (3.5)	19	33	82	.83 (.78, .87)

Note: Accuracy is presented as the percentage of correct trials. One participant was missing behavioral data for all tasks, and one additional participant was missing behavioral data for the flanker task. Error and correct trials are presented as the number of ERP trials retained following artifact and trial rejection procedures. The number of trials required to reach dependability thresholds for each ERP are presented in ‘Dependability Thresholds’ column. Dependability obtained for each ERP using actual data (i.e., total available trials) is presented in 'Overall

Dependability' column. ERN = error-related negativity; CRN = correct-related negativity; Pe = post-error positivity; Pc =post-correct positivity; ΔERN = ERN minus CRN; ΔPe = Pe minus Pc

Table 3

Post-Estimation Contrasts for Task Comparisons

	Dependability (20 trials)	Dependability (all trials)	ICC
ERN			
Flanker v. Go/no-go	-0.02 [-0.09, 0.06]	**0.09 [0.04, 0.16]**	-0.01 [-0.04, 0.02]
Stroop v. Go/no-go	-0.06 [-0.14, 0.03]	0.01 [-0.05, 0.08]	-0.02 [-0.05, 0.01]
Flanker v. Stroop	0.04 [-0.03, 0.13]	**0.08 [0.03, 0.14]**	0.01 [-0.01, 0.04]
CRN			
Flanker v. Go/no-go	0.003 [-0.004, 0.01]	0.003 [-0.004, 0.01]	-0.001 [-0.02, 0.02]
Stroop v. Go/no-go	**0.01 [0.001, 0.02]**	**0.01 [0.001, 0.02]**	0.02 [-0.01, 0.04]
Flanker v. Stroop	-0.01 [-0.01, 0.001]	-0.01 [-0.004, 0.001]	-0.01 [-0.03, 0.005]
Pe			
Flanker v. Go/no-go	**-0.08 [-0.14, -0.02]**	0.02 [-0.01, 0.06]	**-0.06 [-0.10, -0.01]**
Stroop v. Go/no-go	**-0.08 [-0.14, -0.02]**	-0.06 [-0.01, 0.03]	**-0.05 [-0.10, -0.01]**
Flanker v. Stroop	-0.002 [-0.08, 0.07]	0.04 [-0.004, 0.08]	-0.001 [-0.04, 0.04]
Pc			
Flanker v. Go/no-go	**0.03 [0.01, 0.05]**	**0.03 [0.01, 0.05]**	**0.01 [0.001, 0.02]**
Stroop v. Go/no-go	**0.03 [0.01, 0.05]**	**0.03 [0.01, 0.05]**	**0.01 [0.003, 0.03]**
Flanker v. Stroop	-0.003 [-0.02, 0.01]	-0.003 [-0.02, 0.01]	-0.002 [-0.01, 0.01]
ΔERN			
Flanker v. Go/no-go	-0.01 [-0.11, 0.09]	**0.10 [0.03, 0.18]**	-0.004 [-0.02, 0.01]
Stroop v. Go/no-go	-0.003 [-0.10, 0.10]	0.06 [-0.02, 0.14]	-0.002 [-0.02, 0.01]
Flanker v. Stroop	-0.01 [-0.10, 0.08]	0.04 [-0.02, 0.11]	-0.002 [-0.02, 0.01]
ΔPe			
Flanker v. Go/no-go	-0.03 [-0.08, 0.03]	**0.05 [0.01, 0.09]**	-0.02 [-0.04, 0.01]
Stroop v. Go/no-go	**-0.08 [-0.15, -0.01]**	-0.02 [-0.07, 0.03]	**-0.03 [-0.06, -0.01]**
Flanker v. Stroop	0.05 [-0.02, 0.13]	**0.07 [0.03, 0.12]**	0.02 [-0.01, 0.04]

Note: If the 95% credible interval of the contrast excludes zero, this is interpreted as evidence of a difference. In such instances, the contrast is shown in bold font. The Dependability (20 trials) column refers to the dependability estimates using only 20 error trials, and the Dependability (all trials) column refers to the dependability estimates using the average number of error trials recorded during each task. Both dependability columns include the average number of correct trials recorded during each task. ERN = error-related negativity; CRN = correct-related negativity; Pe = post-error positivity; Pc =post-correct positivity; ΔERN = ERN minus CRN; ΔPe = Pe minus Pc; ICC = intraclass correlation coefficient

Figure Captions

Figure 1. Grand average response-locked error-related negativity (ERN) waveforms as a function of task. ERN waveforms reflect activity from 0 to 100 ms (shaded) at electrode FCz. Difference waveforms represent error minus correct activity for each task. Waveforms are only shown from participants recorded using the 500 Hz sampling rate.

Figure 2. Grand response-locked average post-error positivity (Pe) waveforms as a function of task. Pe waveforms reflect activity from 200 to 400 ms (shaded) at electrode Pz. Difference waveforms represent error minus correct activity for each task. Waveforms are only shown from participants recorded using the 500 Hz sampling rate.

Figure 3. Coefficient alpha (α) for error-related negativity scores as a function of increasing error trials and task for full sample (Figure A) and including only individuals with 20 trials for all tasks for a fully within-person comparison (Figure B).

Figure 4. Dependability (ϕ) estimates of error-related negativity scores as a function of increasing error trials and task.

Figure 5. Internal consistency of residualized difference scores for error-related negativity as a function of increasing error trials and task.

Figure 6. Intraclass correlation coefficient (ICC) and dependability (ϕ) estimates for each event-related potential and task. ϕ and associated 95% credible intervals are based on 20 error trials and the average number of correct trials for each task. ERN = error-related negativity; CRN = correct-related negativity; Pe = post-error positivity; Pc = post-correct positivity; ΔERN = ERN minus CRN; ΔPe = Pe minus Pc.

Figure 7. Dependability (ϕ) estimates for each event-related potential and task. ϕ and associated 95% credible intervals are based on the average number of error and correct trials for each task. ERN = error-related negativity; CRN = correct-related negativity; Pe = post-error positivity; Pc = post-correct positivity; ΔERN = ERN minus CRN; ΔPe = Pe minus Pc.

Figure 1

ERN

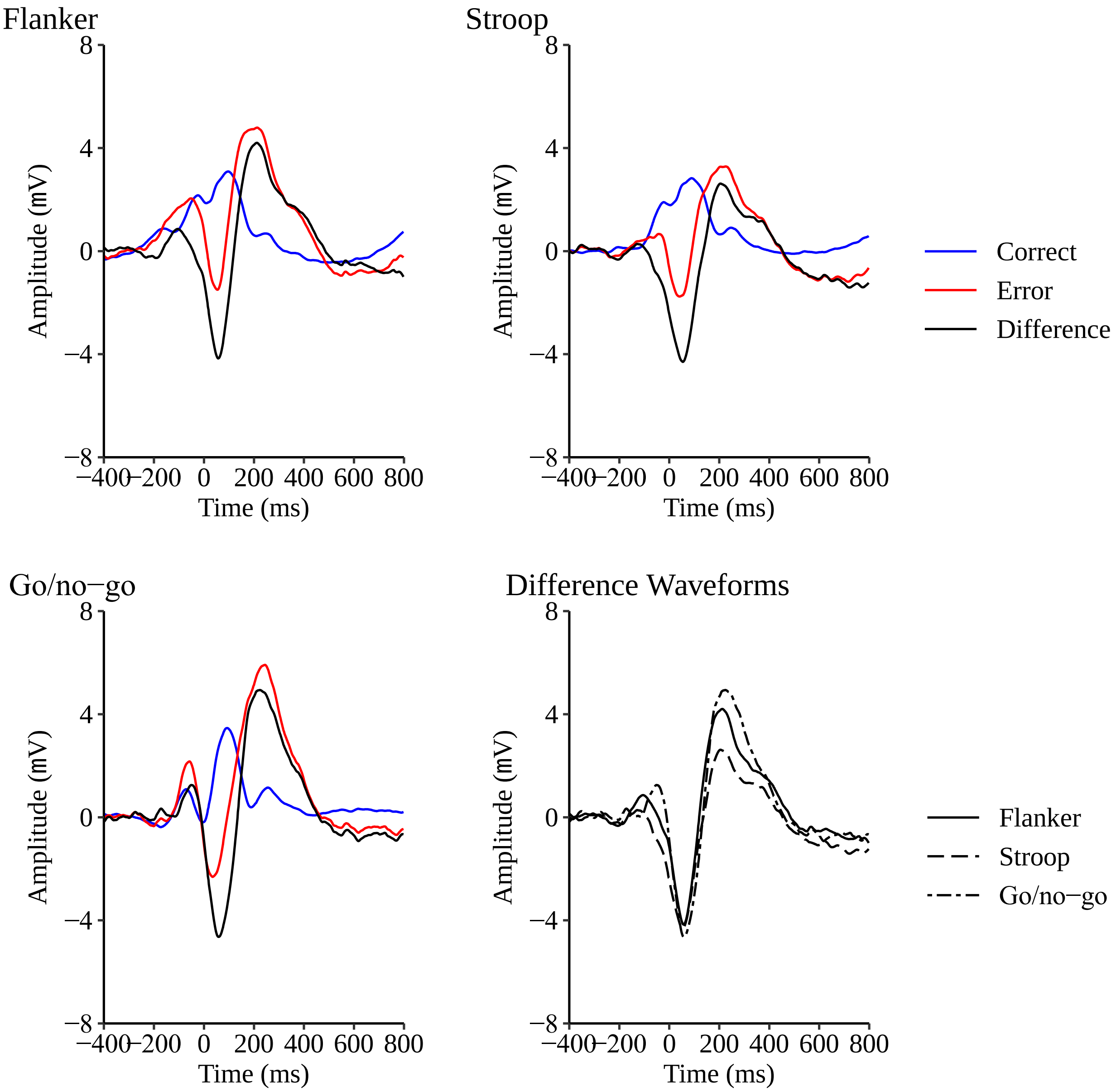

Flanker
Stroop
Go/no−go
Difference Waveforms
Amplitude (mV)
Time (ms)
8
4
0
−4
−8
−400
−200
0
200
400
600
800
Correct
Error
Difference
Flanker
Stroop
Go/no−go

Figure 2

Pe

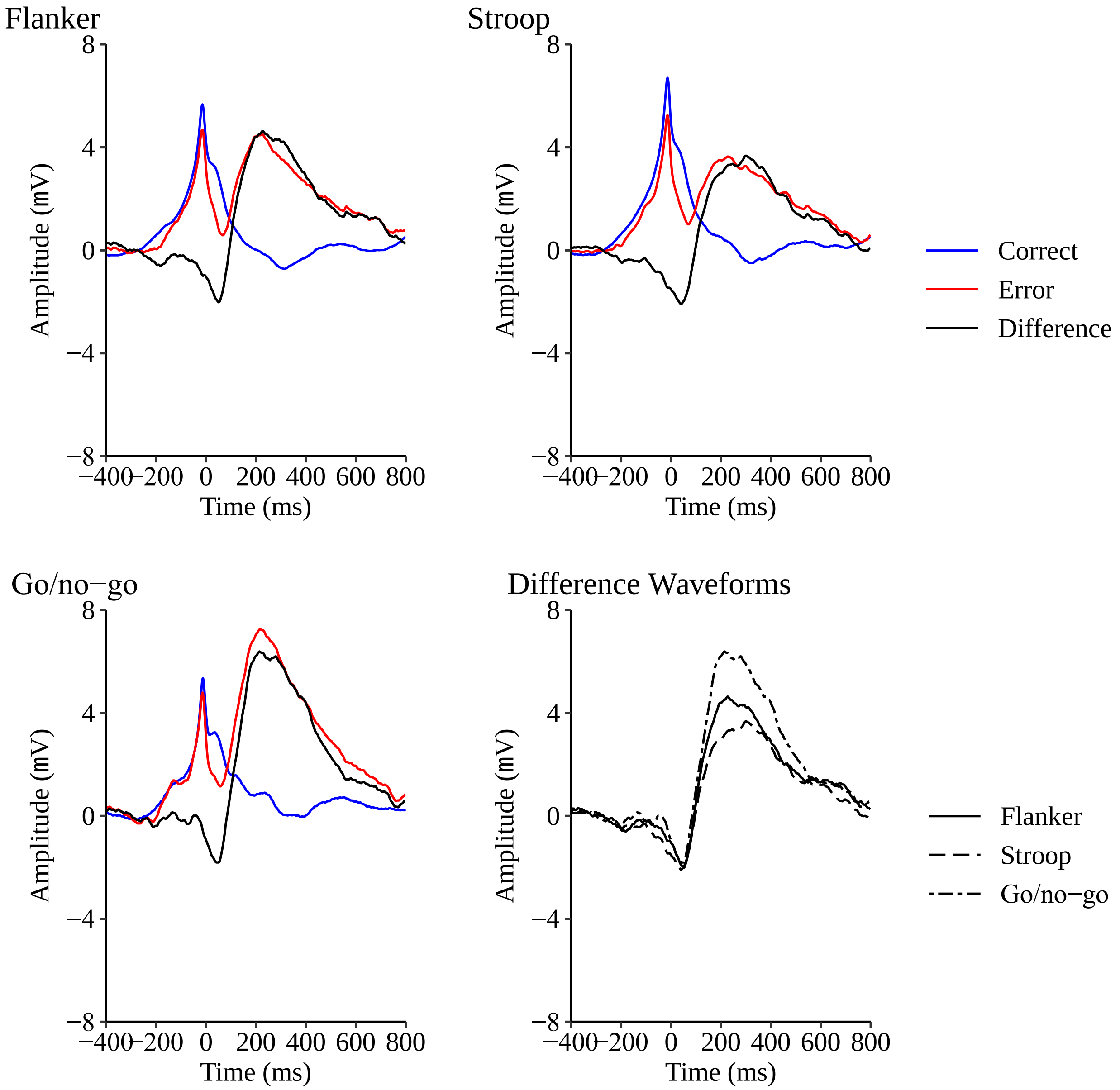

Flanker
Stroop
Go/no−go
Difference Waveforms
Amplitude (mV)
8
4
0
−4
−8
−400
−200
0
200
400
600
800
Time (ms)
Correct
Error
Difference
Flanker
Stroop
Go/no−go

Figure 3

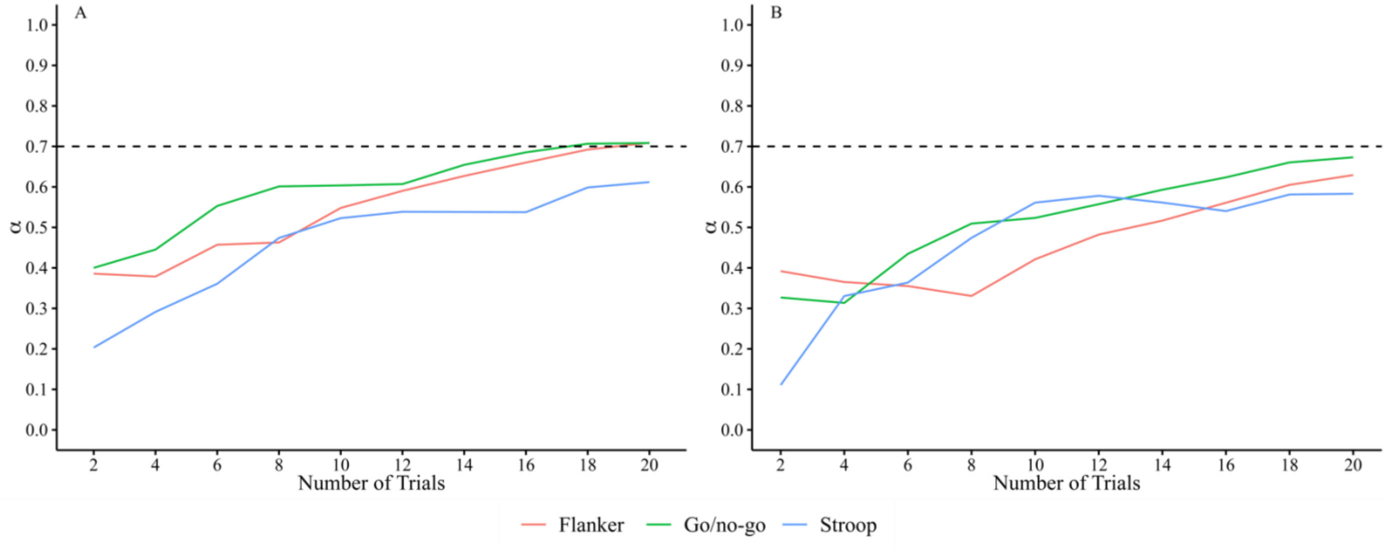

A
B
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
α
2
4
6
8
10
12
14
16
18
20
Number of Trials
Flanker
Go/no-go
Stroop

Figure 4

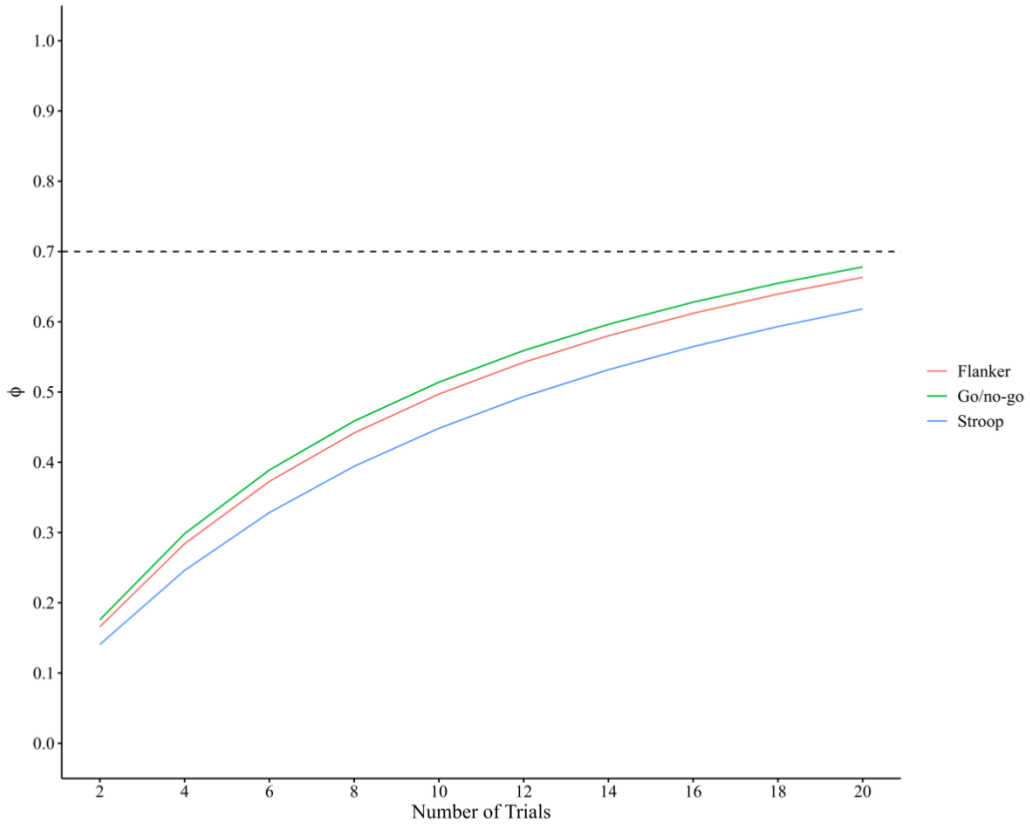

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
ϕ
2
4
6
8
10
12
14
16
18
20
Number of Trials
Flanker
Go/no-go
Stroop

Figure 5

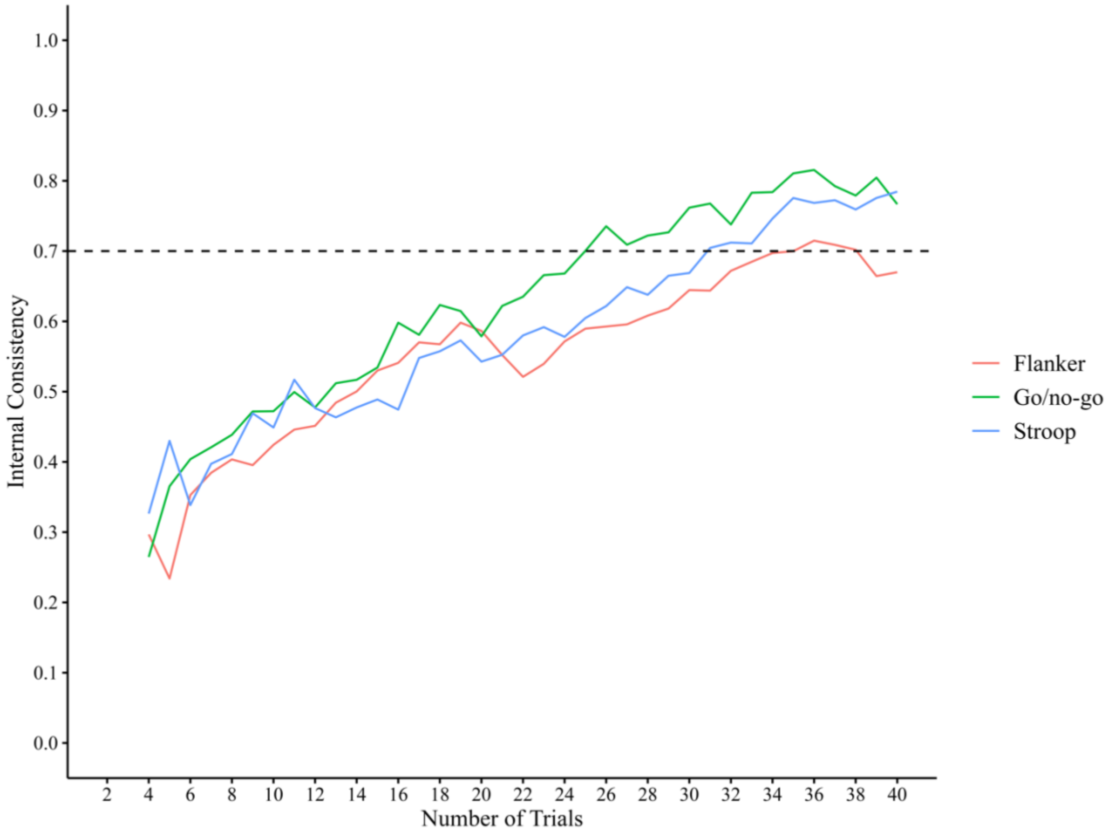

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Internal Consistency
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
Number of Trials
Flanker
Go/no-go
Stroop

Figure 6

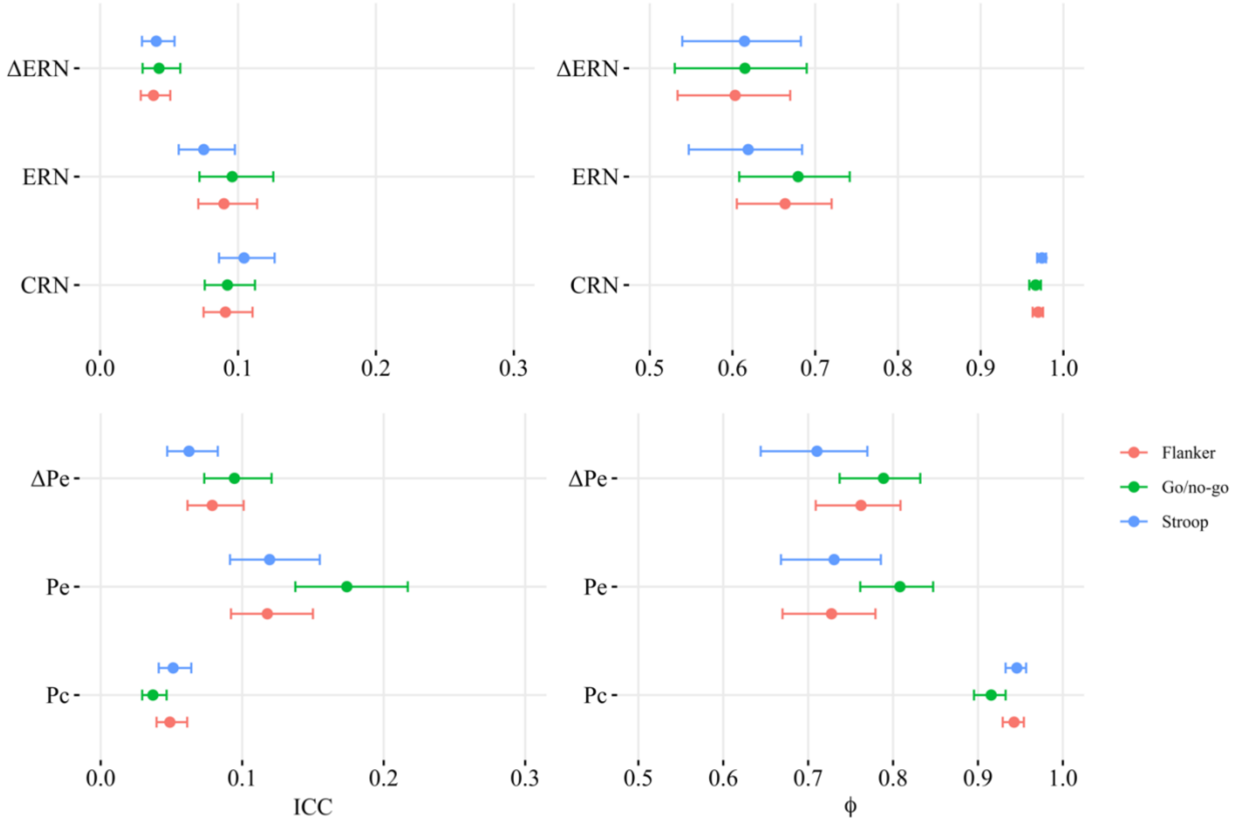
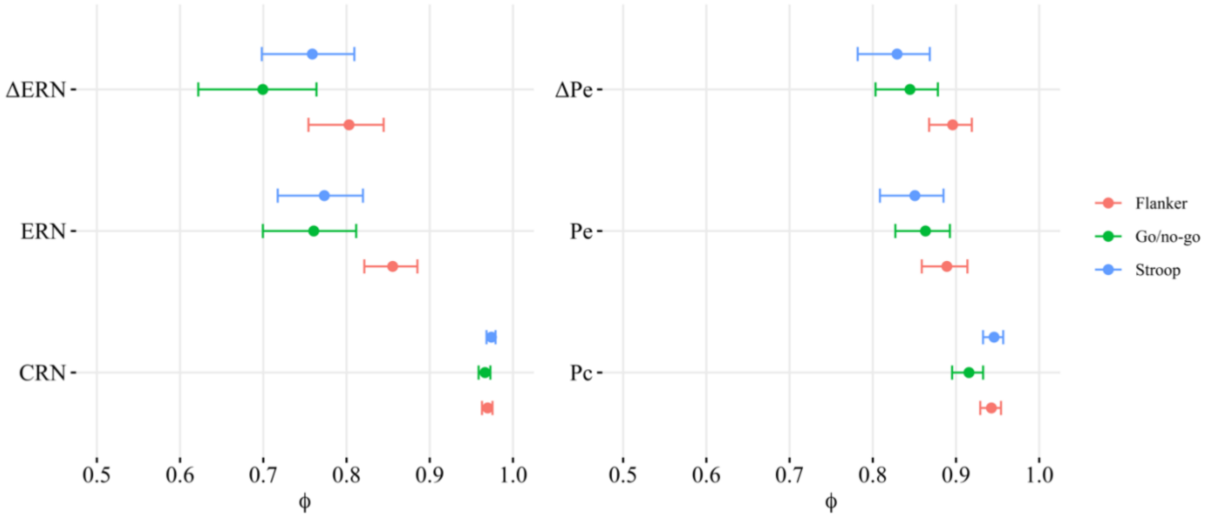

ΔERN
ERN
CRN
0.0
0.1
0.2
0.3
0.5
0.6
0.7
0.8
0.9
1.0
ΔPe
Pe
Pc
ICC
ϕ
Flanker
Go/no-go
Stroop

Figure 7



Supplementary Method

Experimental Tasks

Participants completed a flanker task, Stroop task, and a go/no-go task that were identical to the versions used in Meyer et al. (2013). The order of task presentation was counterbalanced (using a fully crossed Latin-square) across participants and sites. Tasks were presented in E-Prime (Psychology Software Tools, Pittsburg, PA).

All tasks included 7 blocks of 60 trials, resulting in 420 total trials. Each task began with 20 practice trials. The total duration of all three tasks combined was approximately 60 minutes. The order of trials within each task was randomly determined within each block. Stimuli were presented for 200 ms, on a black background, followed by a jittered inter-trial interval (ITI) ranging from 600 to 1,000 ms. Performance-based feedback was presented at the end of each block to encourage fast and accurate behavior. If performance accuracy was below or equal to 75%, the feedback "Please try to be more accurate" was shown. If accuracy was above 90%, the feedback "Please try to respond faster" was shown. If accuracy fell between 75% and 90%, the feedback "You're doing a great job" was shown. For the go/no-go task, feedback was based on accuracy calculated for no-go trials only (i.e., errors of commission).

Flanker task. On each trial, 5 horizontally aligned white arrowheads were presented, with arrows in the array either all pointed the same direction (i.e., congruent trial: >>>>> or <<<<<) or with the central arrow pointed in the opposite direction of other arrows (i.e., incongruent trial: <<><< or >><>>). Participants were instructed to either click the left or right mouse button that corresponded with the direction of the center arrow. Each of the 7 trial blocks consisted of 50% congruent and 50% incongruent trials. At a viewing distance of 65 cm, the arrow array occupied 2° of visual angle vertically and 10° horizontally.

Stroop task. On each trial, one of three color words was shown (‘red’, ‘green’, or ‘blue’), presented in either red or green font. Participants were instructed to click the left mouse button if a word was presented in red font and click the right mouse button if a word was presented in green font. Trial blocks consisted of 1/3 congruent trials (e.g., the word “red” in red font), 1/3 incongruent trials (e.g., the word “red” in green font), and 1/3 neutral trials (e.g., the word ‘blue’ in red or green font). At a viewing distance of 65 cm, 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 click the right mouse button if the triangle was facing upright, but to withhold responding if the triangle was slightly tilted. Trial blocks consisted of 80% go (i.e., upright triangles) and 20% no-go (i.e., 10° titled triangles) trials. At a viewing distance of 65 cm, the triangle occupied 3°×3° of the visual angle.

Analytical Plan

Aim 2. Bayesian location-scale multilevel models were used to estimate between-person, between-trial, and error variances for each ERP and task. These variance components were then used to estimate psychometric reliability. Location-scale multilevel models are well suited to ERP data and accommodate the nested nature of the data (trials nested within tasks nested within participants). The following models were implemented and are described here using Wilkinson notation:

ERP ~ 0 + TaskEvent + (0 + TaskEvent|p|Participant) + (0 + TaskEvent|q|Trial)

Sigma ~ 0 + TaskEvent + (0 + TaskEvent|p|Participant) + (0 + TaskEvent|q|Trial)

Each model was fit within *R* (R Development Core Team, 2021) using *brms*, a front end for *Stan* (Stan Development Team, 2021). Sigma refers to the scale portion of the model, and separate models were used to predict ERN and Pe amplitudes. The same fixed and random effects were modeled on the location and scale portions of the model. The population intercepts were suppressed in the models, resulting in the estimation of separate intercepts for each level of “TaskEvent”. The independent variable “TaskEvent” included six levels representing all unique combinations of Task (flanker, go/no-go, Stroop) and Event (error, correct). The models included participants and trials as random intercepts and “TaskEvent” as random slopes, and covariances were modeled between random intercepts and slopes, which is reflected by the “|p|” and “|q|” notations. The priors for fixed effect estimates used a normal distribution with a mean of 0 and a standard deviation of 3. All variance components were estimated using a Student’s *t* prior distribution with 10 degrees of freedom, a 0 location parameter, and a 2 scale parameter. An LKJ prior distribution with a shape parameter of 2 was used, which can be considered mildly informative.

Using the variance components estimated from the location-scale models and Eq. 1, ϕ of ERN scores were estimated as a function of increasing trial numbers (2, 4, 6, 8, 10, 12, 14, 16, 18, and 20 error trials) separately for each task.

Eq. (1)

$$\phi = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_i^2}{n_{trials}} + \frac{\sigma_e^2}{n_{trials}}}$$

ϕ was estimated as a function of between-person variance (σ_p^2), between-trial variance (σ_i^2), error variance (σ_e^2), and a given number of trials (n_{trials}) (see Baldwin et al., 2015; Clayson, Brush, & Hajcak, 2021; Clayson, Carbine, Baldwin, et al., 2021; Clayson & Miller, 2017b)

Eq. (1) was used to determine how many trials were required for ERN, CRN, Pe, and Pc scores to reach three thresholds of internal consistency (ϕ ≥ .7, .8, .9). Eq. (2), shown below and in (Clayson, Baldwin, et al., 2021), was used to determine how many trials were required for ΔERN and ΔPe scores to reach different thresholds of internal consistency. The number of error trials is limited compared to the number of correct trials because some high-performing participants commit few errors. Therefore, the number of error trials typically place a ceiling on the internal consistency of ΔERN and ΔPe scores. As a result, analyses of the internal consistency of difference scores used the average number of correct trials for each task (see Table 2)[1] and focused on the impact of increasing the number of error trials on internal consistency.

Eq. (2)

$$\phi_D = \frac{\sigma_X^2(p) + \sigma_Y^2(p) - 2\sigma_{XY}(p)}{\sigma_X^2(p) + \sigma_Y^2(p) - 2\sigma_{XY}(p) + \frac{\sigma_X^2(i)}{n'_{iX}} + \frac{\sigma_Y^2(i)}{n'_{iY}} - \frac{2\sigma_{XY}(i)}{\ddot{n}'_i} + \frac{\sigma_X^2(e)}{n'_{iX}} + \frac{\sigma_Y^2(e)}{n'_{iY}} - \frac{2\sigma_{XY}(e)}{\ddot{n}'_i}}$$

The dependability coefficient of a subtraction-based difference score (ϕ_D) can be estimated using the between-person variances of X ($\sigma_X^2(p)$), Y ($\sigma_Y^2(p)$), and their covariance ($2\sigma_{XY}(p)$); the within-person (residual) variance for X ($\sigma_X^2(e)$), Y ($\sigma_Y^2(e)$), and their covariance ($2\sigma_{XY}(e)$); and between-trial variances for X ($\sigma_X^2(i)$), Y ($\sigma_Y^2(i)$), and their covariance ($2\sigma_{XY}(i)$). The contribution of within-person and between-trial variances are impacted by the average number of X trials (n'_{iX}) and of Y trials (n'_{iY}) and by the harmonic mean of X and Y trials ($\ddot{n}'_i$) (see Clayson, Baldwin, & Larson, 2021). Although Eq (2) includes the covariance between residuals for sake of completeness, correct and error trials are not naturally cooccurring events but are two separate conditions. Therefore, error covariances were set to zero (Clayson, Baldwin, & Larson, 2021) .

[1] While the average number of correct trials for difference score estimation are used throughout analyses, we also preregistered that these analyses would be performed including 200 correct trials for each task in order to limit differences associated with differing correct trials available from each task. Analyses using 200 correct trials are reported in the supplementary results below. Intraclass correlation coefficients (ICCs) are reported as a trial-independent estimate of reliability across tasks.

Next, Eq. (3), shown below, was used to calculate the internal consistency of residualized difference scores and determine how many trials are required to reach three thresholds of internal consistency (i.e., ≥ .7, .8, .9). The average number of correct trials from each task was used for the correct-trial count and, like the internal consistency of subtraction-based difference scores, the focus was on examining the impact of increasing the number of error trials on the internal consistency of the residualized scores.

Eq. (3)

$$\rho_{RR\prime} = \frac{\rho_{YY} + \rho^2{}_{XY}\rho_{XX} - 2\rho^2{}_{XY}}{1 - \rho^2{}_{XY}}$$

The estimation of the internal consistency of residualized scores used internal consistencies of X (ρ_{XX}) and Y (ρ_{YY}) and their squared correlation ($2\rho^2{}_{XY}$) (Clayson, Baldwin, & Larson, 2021).

Aim 3. To examine task differences in internal consistency (i.e., dependability, ICC) of each ERP, post-estimation contrasts from the Bayesian location-scale multilevel models were used to estimate variance components. Eq. (1) and (2) were used to calculate dependability estimates using 20 error trials and the average number of correct trials. Eq. (4), shown below, was used to calculate intraclass correlation coefficients (ICCs) for each ERP from each task, which can be interpreted as the proportion between-person variance to total variance. Eq. (5), shown below, was used to calculate ICCs for subtraction-based difference scores (Clayson, Baldwin, & Larson, 2021).

Eq. (4)

$$\mathrm{ICC} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_i^2 + \sigma_e^2}$$

Eq. (5)

$$\mathrm{ICC}_D = \frac{\sigma_X^2(p) + \sigma_Y^2(p) - 2\sigma_{XY}(p)}{\sigma_X^2(p) + \sigma_Y^2(p) - 2\sigma_{XY}(p) + \sigma_X^2(i) + \sigma_Y^2(i) - 2\sigma_{XY}(i) + \sigma_X^2(e) + \sigma_Y^2(e) - 2\sigma_{XY}(e)}$$

Task differences in dependability and ICCs were supported when the 95% credible interval of the post-estimation contrasts included zero. Eq (5) includes the covariance between residuals for sake of completeness, but error covariances were set to zero (Clayson, Baldwin, & Larson, 2021).

Supplementary Results

Aim 1

Supplementary Figure 1 shows coefficient alpha (α) for event-related brain potential (ERP) component scores from flanker, go/no-go, and Stroop tasks as a function of increasing error trials for post-error positivity (Pe) and correct trials for correct-response negativity (CRN) and post-correct positivity (Pc). CRN scores from the go/no-go task reached α of .7 at 16 trials, while CRN scores from the Stroop task reached α of .7 at 18 trials. CRN scores from the flanker task did not reach α of .7 with 20 trials included. Pe scores reached α of .7 at 16, 14, and 18 trials for flanker, go/no-go, and Stroop tasks, respectively. Pc scores did not reach acceptable levels of α with up to 20 trials included for any task.

Aim 2

Dependability. Supplementary Figure 2 shows dependability (ϕ) estimates for ERP scores from flanker, go/no-go, and Stroop tasks as a function of increasing error trials (Pe) or correct trials (CRN, Pc). CRN scores from the Stroop task reached ϕ of .7 at 20 trials. CRN scores from the go/no-go and flanker tasks reached ϕ of .7 with 23 trials included. Pe scores reached ϕ of .7 at 18, 11, and 17 trials for the flanker, go/no-go, and Stroop tasks, respectively. Pc scores did not reach acceptable levels of ϕ with up to 20 trials included for any task.

Dependability Thresholds. Supplementary Table 1 shows the number of trials required for ΔERN and ΔPe scores to reach acceptable levels of ϕ (.70, .80, .90) for the flanker, go/no-go, and Stroop tasks when including 200 correct trials in estimation. Similar to analyses using the average number of correct trials (see manuscript), the go/no-go task required the fewest number of trials to reach each reliability threshold for ΔPe, followed by the flanker then Stroop task. ΔERN scores from all tasks required a similar number of trials to reach dependability thresholds of .7 (approximately 30 trials) and .8 (approximately 60 trials). ΔERN scores from the go/no-go task required the fewest trials to reach dependability threshold of .9, followed by the Stroop then flanker task. Overall, ΔPe scores required fewer trials than ΔERN scores to reach each reliability threshold regardless of the task. Compared to analyses using the average number of correct trials, more trials were needed for ΔERN and ΔPe scores from each task to reach each reliability threshold when including 200 correct trials.

Residualized Difference Score. Results for residualized ΔPe are shown below in Supplementary Figure 3. ΔPe scores reached an internal consistency of .7 at 16 trials, 12 trials, and 25 trials for flanker, go/no-go, and Stroop, respectively. The number of participants available for each calculation was 171 participants with at least 16 flanker error trials, 172 participants with at least 12 go/no-go error trials, and 141 participants with at least 25 Stroop error trials. ΔPe from go/no-go reached an internal consistency of .8 at 26 trials, based on 98 participants with at least 26 go/no-go error trials. ΔPe from flanker and Stroop did not reach an internal consistency of .8 even with 40 error trials included. The decision not to further increase the number of error trials available for estimation was due to the loss in sample size at a higher number of trials.

Aim 3

Analyses of task comparisons of ϕ were repeated including 20 error trials and 200 correct trials from each task to control for the different number of available correct trials across tasks. Results are shown below in Supplementary Figure 4 and Table 2. Results were similar to analyses including the average number of correct trials (see manuscript), such that ϕ of ERN and ΔERN scores did not differ across tasks, failing to support hypothesis #3.

Consistent with findings using the average number of correct trials, Pe scores from the go/no-go task yielded higher ϕ than Pe scores from the Stroop and flanker tasks, whereas ΔPe scores from the go/no-go task yielded higher ϕ than scores from the Stroop task, but not flanker task.

Exploratory Site Analysis

Supplementary Table 3 provides demographic characteristics by site. Numerical comparisons show that participants from USF showed greater racial, ethnic, and income diversity, whereas those from BYU showed more diversity in sex and gender.

To determine the influence of study site (USF, BYU) on ERP internal consistency, site was added as a covariate to the Bayesian location-scale multilevel models used for Aim 2 (see *Analytical Plan* in manuscript). Models including Site were estimated using the same model parameters and priors as the models implemented for the manuscript. The following models were implemented and are described here using Wilkinson notation:

ERP ~ 0 + TaskEvent + TaskEvent:Site + (0 + TaskEvent|p|Participant) + (0 + TaskEvent|q|Trial)

Sigma ~ 0 + TaskEvent + TaskEvent:Site + (0 + TaskEvent|p|Participant) + (0 + TaskEvent|q|Trial)

To justify interpretation of models including site, leave-one-out cross-validation via Pareto smoothed importance sampling (PSIS-LOO) was used to compare model fits (Vehtari et al., 2016) using the *R* package *loo* (Vehtari et al., 2024). PSIS-LOO compares the predictive accuracy of models. A given model is expected to show better fit over another when 1) the difference in expected log predictive density ($\Delta\widehat{elpd}_{loo}$) is greater than 4 and 2) $\Delta\widehat{elpd}_{loo}$ is greater than 2 times the standard error of $\Delta\widehat{elpd}_{loo}$. Two models are considered to have comparable fits when 1) $\Delta\widehat{elpd}_{loo}$ is less than 4 or 2) $\Delta\widehat{elpd}_{loo}$ is less than 2 times the standard error of $\Delta\widehat{elpd}_{loo}$.

The ERN location-scale model that included Site showed improved fit compared to the ERN model that excluded Site based on PSIS-LOO criteria (see Supplementary Table 4), supporting the interpretation of ERN reliability when controlling for site. As such, analyses for Aim 2 were repeated using variance components derived from the model including Site to determine if patterns of dependability results remained consistent.

When accounting for site, results continued to show that ERN scores from the go/no-go task demonstrate the highest numerical dependability when including 2-20 errors trials. Task comparisons of ERN scores when including the average number of error trials were also repeated. ERN and ΔERN scores from the flanker task continued to demonstrate the highest dependability due to the availability of more error trials. Thus, patterns of results for the model including Site were largely comparable to results reported in the manuscript.

The Pe location-scale model including Site did not show improved fit over the corresponding Pe model without Site based on PSIS-LOO criteria; therefore, interpreting Pe models that account for site was not justified based on model fit (see Supplementary Table 4).

References

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., Gelman, A., Goodrich, B., Piironen, J., Nicenboim, B., & Lindgren, L. (2024). *Efficient LOO-CV and WAIC for Bayesian models—Loo-package* [Computer software]. http://mc-stan.org/loo/reference/loo-package.html#references-1

Vehtari, A., Gelman, A., & Gabry, J. (2016). *Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC | Statistics and Computing*. *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Supplementary Table 1

Number of Error Trials Required for Difference Scores to Reach Three Thresholds of Reliability (Using 200 Correct Trials)

	Dependability Thresholds		
	0.7	0.8	0.9
Flanker			
ΔERN	33	64	229
ΔPe	15	26	68
Go/no-go			
ΔERN	31	59	189
ΔPe	13	22	54
Stroop			
ΔERN	32	60	206
ΔPe	20	36	97

Note: ΔERN = error-related negativity minus correct-related negativity; ΔPe = post-error positivity minus post-correct positivity

Supplementary Table 2

Task Comparisons Using Post-Estimation Contrasts from Bayesian Multilevel Models (Using 20 Error Trials and 200 Correct Trials)

	Dependability
ERN	
Flanker v. Go/no-go	-0.02 [-0.09, 0.06]
Stroop v. Go/no-go	-0.06 [-0.14, 0.02]
Flanker v. Stroop	0.04 [-0.03, 0.13]
CRN	
Flanker v. Go/no-go	-0.001 [-0.01, 0.01]
Stroop v. Go/no-go	0.01 [-0.004, 0.02]
Flanker v. Stroop	-0.006 [-0.02, 0.002]
Pe	
Flanker v. Go/no-go	**-0.08 [-0.14, -0.02]**
Stroop v. Go/no-go	**-0.08 [-0.14, -0.02]**
Flanker v. Stroop	-0.002 [-0.08, 0.07]
Pc	
Flanker v. Go/no-go	**0.03 [0.001, 0.05]**
Stroop v. Go/no-go	**0.03 [0.01, 0.06]**
Flanker v. Stroop	-0.004 [-0.02, 0.02]
ΔERN	
Flanker v. Go/no-go	-0.01 [-0.11, 0.08]
Stroop v. Go/no-go	-0.003 [-0.10, 0.10]
Flanker v. Stroop	-0.01 [-0.10, 0.08]
ΔPe	
Flanker v. Go/no-go	-0.03 [-0.09, 0.03]
Stroop v. Go/no-go	**-0.08 [-0.15, -0.01]**
Flanker v. Stroop	0.05 [-0.02, 0.13]

Note: If the 95% credible interval of the contrast excludes zero, this is interpreted as evidence of a difference. In such instances, the contrast is shown in bold font. ERN = error-related negativity; CRN = correct-related negativity; Pe = post-error positivity; Pc =post-correct positivity; ΔERN = ERN minus CRN; ΔPe = Pe minus Pc

Supplementary Table 3

Summary Demographic Information for Participants by Site

Site	USF (*n* = 64)		BYU (*n* = 118)	
	M	*SD*	*M*	*SD*
Age	19.25	2.24	20.14	-1.89
	n	%	*n*	%
Sex				
Female	48	75.00%	69	58.47%
Male	16	25.00%	49	41.53%
Gender				
Women	47	73.44%	68	57.63%
Men	16	25.00%	49	41.53%
Bigender	0	0.00%	1	0.85%
Nonbinary	1	1.56%	0	0.00%
Race				
White	43	67.19%	112	94.92%
Asian	14	21.88%	0	0.00%
American Indian/Alaskan	2	3.13%	0	0.00%
Black/African American	2	3.13%	0	0.00%
Multiracial	3	4.69%	6	5.08%
Hispanic/Latino/Spanish Origin	26	40.63%	18	15.25%
Sexual Orientation				
Heterosexual	49	76.56%	115	97.46%
Bisexual/Pansexual	10	15.63%	2	1.69%
Gay/Lesbian	1	1.56%	0	0.00%
Queer	1	1.56%	0	0.00%
Asexual	1	1.56%	0	0.00%
Unsure	2	3.13%	1	0.85%
Income				
Below $15k	9	14.06%	8	6.78%
$15,000 to $30,000	4	6.25%	8	6.78%
$30,001 to $45,000	6	9.38%	6	5.08%
$45,001 to $60,000	11	17.19%	7	5.93%
$60,001 to $75,000	9	14.06%	12	10.17%
Over $75,000	25	39.06%	76	64.41%
Unknown	0	0.00%	1	0.85%

Note: Participants provided their parent/guardian income if they were a full-time student or dependent.

Supplementary Table 4

Comparison of Model Fit Summary when Excluding vs. Including Site Variable

ERP	$\widehat{\text{elpd}}_{\text{loo}}$	$\text{SE}(\widehat{\text{elpd}}_{\text{loo}})$	$\Delta\widehat{\text{elpd}}_{\text{loo}}$	$\Delta\text{SE}(\widehat{\text{elpd}}_{\text{loo}})$
ERN				
Excluding Site	-670,536.8	389.39		
Including Site	-670,541.3	389.44	-4.53	1.87
Pe				
Excluding Site	-660,442.0	486.12		
Including Site	-660,446.2	486.16	-4.20	3.27

Note: Log predictive density = $(\widehat{\text{elpd}}_{\text{loo}})$; Standard error of log predictive density = $\text{SE}(\widehat{\text{elpd}}_{\text{loo}})$. $\Delta\widehat{\text{elpd}}_{\text{loo}}$ and $\Delta\text{SE}(\widehat{\text{elpd}}_{\text{loo}})$ represent the difference in $\widehat{\text{elpd}}_{\text{loo}}$ and $\text{SE}(\widehat{\text{elpd}}_{\text{loo}})$, respectively, between models excluding and including Site variable.

Supplementary Figure Captions

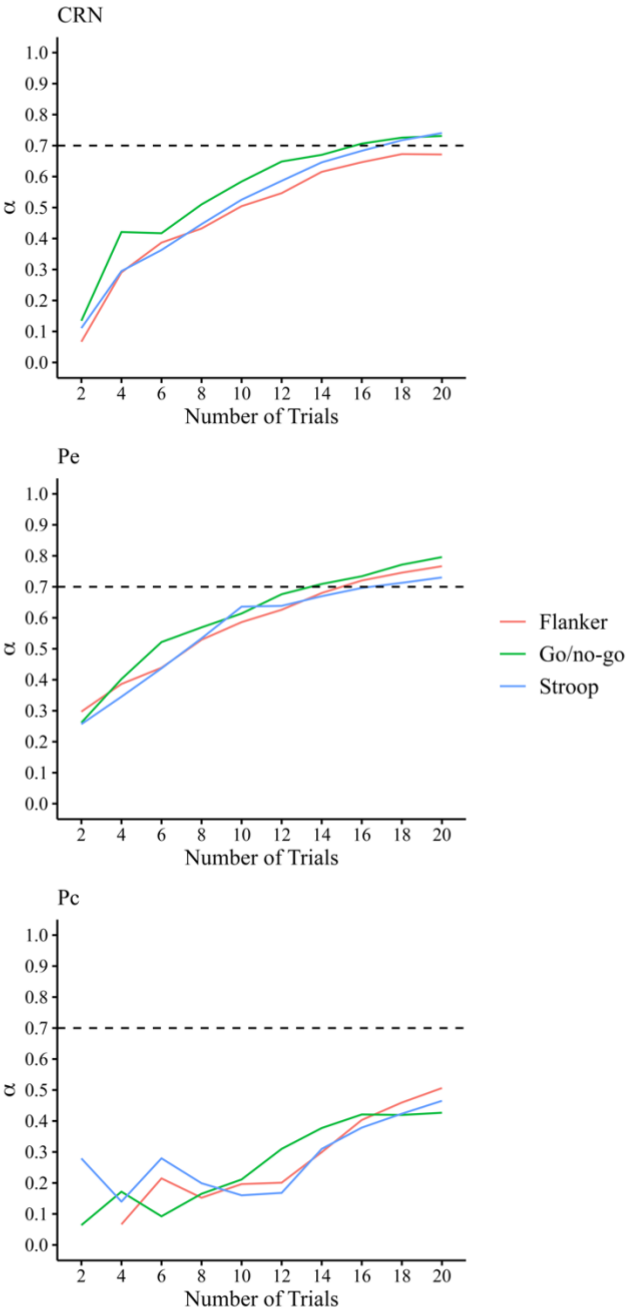
Supplementary Figure 1. Coefficient alpha (α) for correct-related negativity (CRN), post-error positivity (Pe), and post-correct positivity (Pc) as a function of increasing trials.

Supplementary Figure 2. Dependability (ϕ) for correct-related negativity (CRN), post-error positivity (Pe), and post-correct positivity (Pc) as a function of increasing trials.

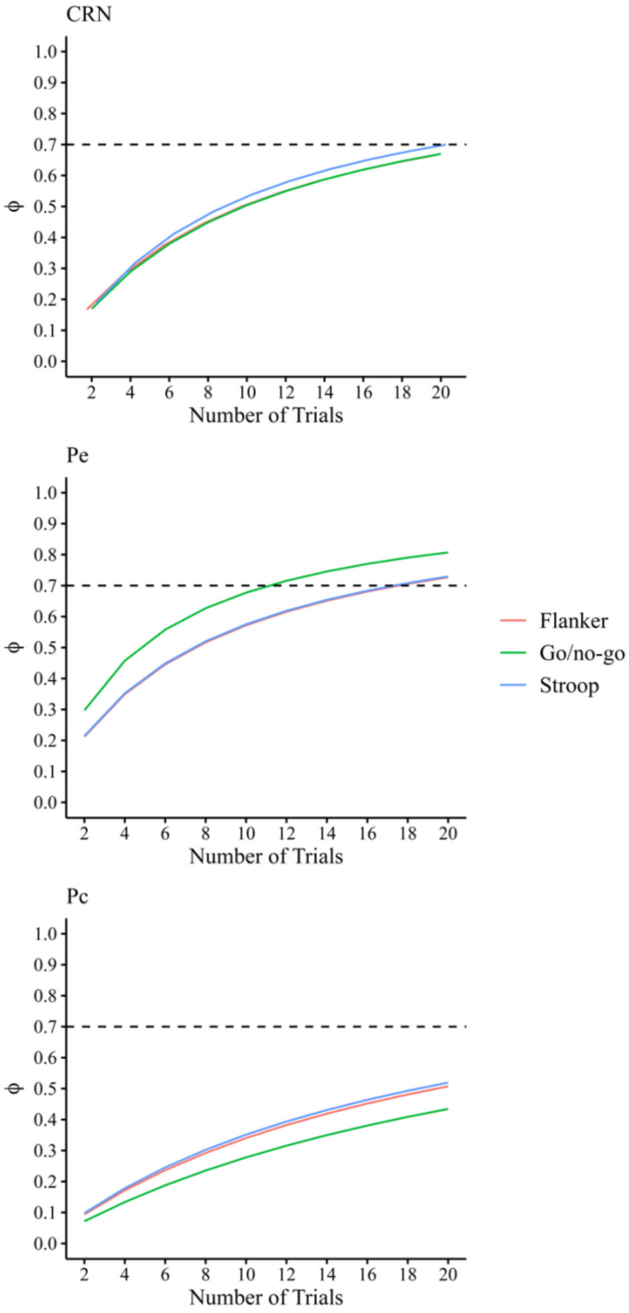
Supplementary Figure 3. Residualized difference score reliability for post-error positivity as a function of increasing error trials.

Supplementary Figure 4. Dependability (ϕ) for each event-related potential and task based on 200 correct trials and 20 error trials. ERN = error-related negativity; CRN = correct-related negativity; Pe = post-error positivity; Pc =post-correct positivity; ΔERN = ERN minus CRN; ΔPe = Pe minus Pc.

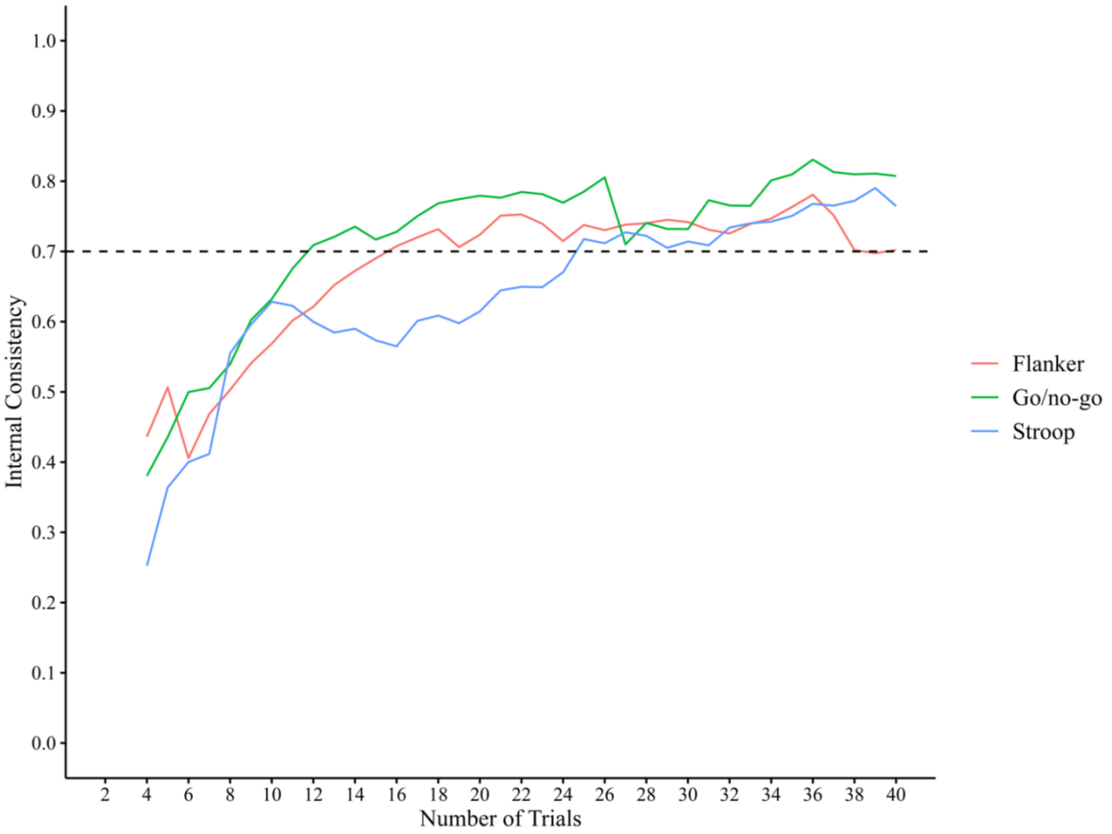
Supplementary Figure 1



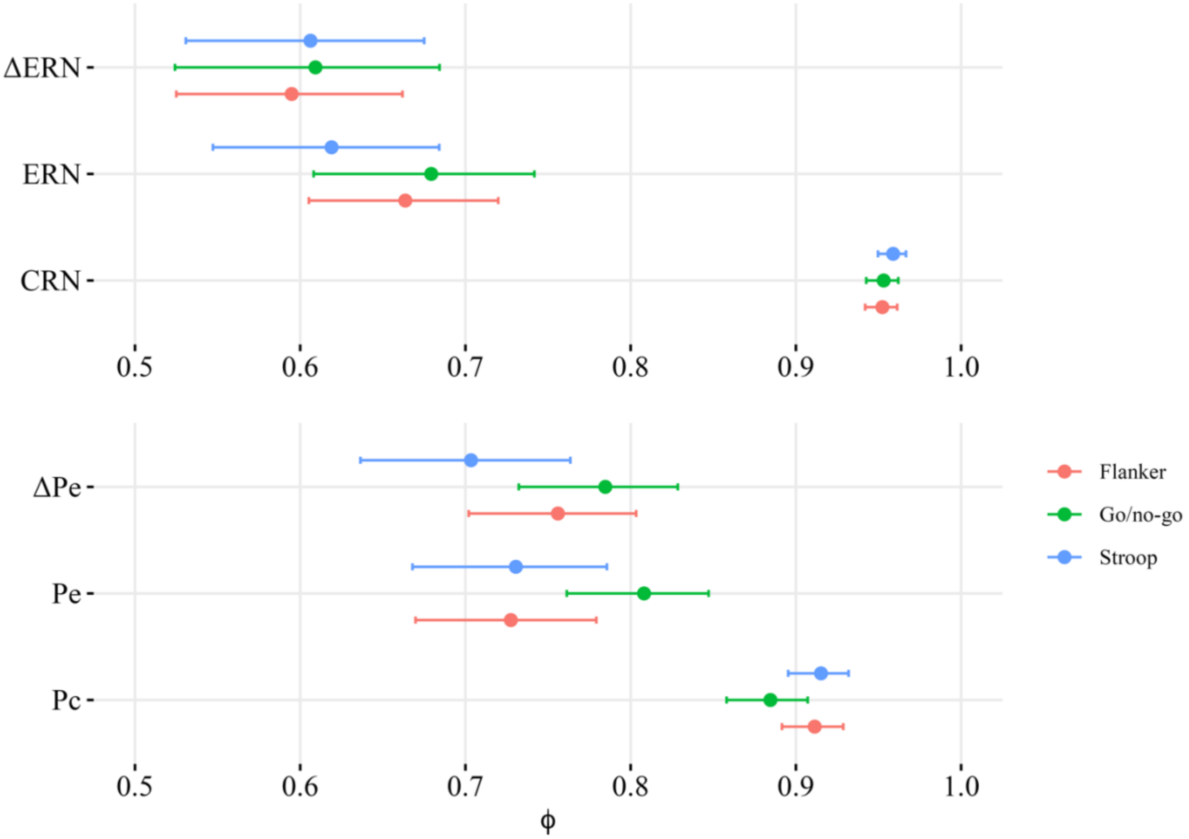
Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 4


ΔERN
ERN
CRN
ΔPe
Pe
Pc
ϕ
Flanker
Go/no-go
Stroop